International Journal on Emerging Technologies 15(2): 05-12(2024)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Model for the Prediction of Default Risk of Funding Requests Using Data Mining

Sameh Ali[1*], Atef Raslan[2] and Lamiaa Fattouh[3]
[1]Faculty of Graduate Studies for Statistical Research Cairo University Cairo, Egypt.
[2]Department of Information Systems, Higher Institute of Advanced Studies.
[3]Faculty of Graduate Studies for Statistical Research Cairo University Cairo, Egypt.

(Corresponding author: Sameh Ali)*
(Received 02 May 2024, Revised 10 June 2024, Accepted 15 July 2024)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Microfinance institutions currently confront numerous financing challenges, particularly within the non-bank sector where risks abound. Each year, there is a notable incidence of borrowers defaulting on their microfinance obligations, resulting in substantial financial setbacks for these companies. Given the burgeoning volume of electronic data and transactional activity in the banking realm, data mining emerges as a pivotal strategic domain. Leveraging data mining techniques, valuable patterns and insights can be gleaned from vast datasets, thereby furnishing actionable information to mitigate risks associated with nonbank loans. This study employs data mining as a tool to extract pertinent insights from the credit data of microfinance companies, facilitating the construction of a model aimed at assessing borrower eligibility and identifying potential default risks. The study employs the open-source machine learning platform WEKA. This study uses data mining to develop a predictive model for microfinance institutions to enhance decision-making in client financing. By employing cross-validation and percentage splits (80-20, 70-30, 60-40), cross-validation showed slightly higher accuracy. The model performed excellently, especially with preprocessed data, highlighting the importance of data cleaning. The J48 proved to be the most effective algorithm, demonstrating superior accuracy. The study emphasizes the potential of using historical data to assess client credit status during financing approvals, reducing loan defaults, and supporting the growth of non-banking institutions.

Keywords: Data Mining Technique, Classification, Credit Risk, Non-Banking Sector, Microfinance, Fraud detection.

INTRODUCTION

Microfinance institutions and non-governmental organizations are established to provide microcredit to economically disadvantaged individuals who lack stable employment or credit history, making it difficult for them to access traditional banking services. These borrowers often engage in small-scale economic activities but are underserved by conventional banks. Egypt hosts numerous recognized microfinance institutions, with "Tamweely Microfinance" being one of the accredited institutions in the country (Hailemariam *et al*., 2012). The rapid progression of science and information technology has led to the generation and continual maintenance of vast volumes of data. Data mining, utilizing fundamental techniques for extracting information and identifying patterns, has emerged as a crucial tool across various fields of endeavor. Notably, within the non-banking sector, data mining applications play a significant role. One such application is credit scoring, among the earliest tools developed for financial risk management.

Credit scoring provides valuable insights to lenders in the banking industry, aiding in lending decisions. Furthermore, data mining empowers microfinance companies to enhance decision-making throughout the loan approval process. By analyzing factors such as credit history, employment status, and demographic profile, data mining can discern the credit behavior of borrowers. This information enables microfinance companies to assess customers and determine their suitability for a loan, as well as identify potential default risks. Armed with insights into a borrower's likelihood of default, microfinance companies can proactively mitigate risks, thereby enhancing their risk management practices (Subia and Galapon 2020; Zeng *et al*., 2017).

Utilizing data mining techniques, it becomes possible to analyze the behavior and reliability of borrowers associated with microfinance companies. However, ensuring the quality of data is paramount, as the efficacy of decisions hinges upon it. Through meticulous data preprocessing, the raw experimental data can be refined, allowingfor the selection of variables that contain pertinent and essential information exclusively. Real-world data often contains noise, inconsistencies, redundancies, or irrelevant information, necessitating the elimination of outliers, standardization, and data cleaning to attain the desired data quality (Abakar, 2020).

One of the most significant challenges in data mining research pertains to the initial stage of refining gathered data, particularly in large databases. Achieving accurate and useful results mandates that the data be both relevant and original. Handling such data is both time-consuming and crucial. Moreover, clear correlations must exist within the data to yield valuable patterns. Enhancing the efficiency and accuracy of classifications further requires the incorporation of algorithms to develop a meaningful model (Kulkarni and Kulkarni 2016).

In light of these considerations, this paper outlines a systematic approach to creating a sample model for predicting default risk.

This research delineated the following objectives: Investigate the viability of employing data mining techniques to unearth patterns within typical microfinance datasets. Construct a predictive model leveraging these discerned patterns to identify potential microfinance defaulters. Identify key financial parameters and successful project attributes conducive to averting defaults. Facilitate streamlined decision-making by early detection of project success or failure. Present findings and offer recommendations for future research endeavors.

The scope of this research is confined to exploring the potential of data mining techniques in forecasting defaulters to enhance company performance and facilitate informed decision-making. The focus primarily revolves around classification methods, deemed suitable for constructing a predictive model aimed at extracting insights from non-banking investment data. The dataset under scrutiny will be sourced from Tamweely Microfinance Company in Egypt.

BACKGROUND

Effectively managed and modeled data possesses the capacity to provide valuable insights that greatly enhance decision-making processes. Within the financial industry, data warehouses serve as the bedrock, providing organizations with the means to harness data for informed decision-making. Through meticulous analysis of this data, organizations can make well-founded evaluations concerning the feasibility and potential success of projects seeking financial backing. Furthermore, delving into historical data allows for the identification of projects that have demonstrated success in the past, thereby assisting in risk mitigation for both the funding institution and the applicants (Zeng *et al.,* 2017).

It signifies that leveraging data-driven approaches could present fresh opportunities for enhancing business strategies. Data mining, also referred to as Knowledge Discovery, involves extracting significant, non-obvious, implicit, previously unidentified, and potentially valuable information or patterns from extensive databases. The significance of data mining has been evident for over twenty years. Forward-thinking enterprises are prioritizing data in their strategic decision-making processes. Data analysis typically progresses in two stages: Discovery and search. The patterns discovered in the discovery phase can be utilized in the subsequent search phase. It's important to note that data mining isn't mere data reporting. Genuine data mining serves specific purposes and is a statistical process aimed at achieving business objectives, enabling business analysts to unearth valuable patterns within available data. Additionally, data preprocessing plays a crucial role in data mining, as the quality and completeness of data significantly impact the effectiveness of data mining algorithms (Zeng *et al.*, 2017).

Sophisticated data mining technologies, tailored to various data sets and objectives, facilitate the prediction of outcomes for novel scenarios based on patterns identified from familiar instances. Such predictive capabilities offer insights into potential future outcomes of implementing a strategy and enable risk assessment. Big Data has emerged as a valuable resource for microfinance firms, serving as a tool for assessing creditworthiness and detecting fraudulent activities (Nguyen, 2019).

REVIEW OF RELATED LITERATURE

This section outlines the literature review and pertinent research endeavors. Various studies have explored similar themes, and below is a concise overview of several papers we have examined and scrutinized.

The author introduced a novel methodology for evaluating loan risk within the banking sector, leveraging data mining techniques. To predict loan statuses, the model was constructed using data sourced from the banking industry. The dataset comprises 1000 instances, partitioned into a training set (80% of the data) and a testing set (20% of the data). Three algorithms—Bayes Net, J48, and Naïve Bayes—were employed to develop the proposed model, implemented, and evaluated using the WEKA software. The accuracy measures for each algorithm were as follows: Bayes Net (73.8739%), J48 (78.3784%), and Naïve Bayes (73.8739%). The results were extensively discussed, and a comprehensive comparison of the algorithms was conducted. Based on the findings, J48 was selected as the most accurate algorithm (Abakar, 2020).

The primary objective of the study was to explore the application of data mining techniques in examining customer loyalty and forecasting loan default occurrences. This encompassed identifying strategies for customer retention, forecasting liquidity risk, introducing novel services, and enhancing profit margins, rated on a scale from 1 to 5. Through experiments conducted on a dataset comprising 9551 records extracted from a database, it was found that the J48 classifier algorithm yielded favorable outcomes in accurately classifying instances. The study utilized the WEKA data mining tool for analysis (Hailemariam *et al.*, 2012).

The case study introduces the application of various data mining technologies in crafting a mechanism for evaluating loan risk tailored for a subprime lender. Diverse data mining methods were employed to derive the outcomes, with the analysis conducted using the WEKA data mining tool. The dataset comprises 1000 instances, consisting of 700 good cases and 300 bad cases. The experimental process involved training the models on 70% of the dataset and testing them with the remaining 30%. The algorithms employed include J48, EM, Naïve Bayes, K-means. The accuracy rates for each method were determined as follows: J48 achieved

71.44%, EM yielded 28.77%, Naïve Bayes attained 75.09%, and K-means reached 57.47%. It was concluded that a decision tree, represented by the J48 algorithm, is the most suitable data mining technology for developing a loan risk assessment system tailored for subprime lenders (Lee and Wang 2020).

This study constructed a model aimed at evaluating and determining the suitability of a borrower for a loan or assessing the risk of default. The model, implemented using the WEKA software, underscores the crucial role of preprocessing or data cleaning in enhancing accuracy rates. Particularly, the results achieved through the J48 algorithm were noteworthy, showcasing a high correctly classified instances rate of 96.3647%. The dataset utilized comprised 3466 instances with attributes structured in the format (AIRFF) (Nalić and Švraka 2018).

The study utilizes data mining techniques to enhance bank performance and decision-making by predicting defaulters. Experiments were conducted using microfinance data obtained from an agricultural bank in Sudan to forecast microfinance status. The Random Forest, KNN, and Naive Bayes classification algorithms were employed, yielding different accuracy rates: Random Forest achieved 94.6%, KNN reached 87.4%, and Naive Bayes attained 92.3%. Based on these accuracy rates, Random Forest was selected as the optimal algorithm. The Orange application data mining tool was utilized for analysis, with recommendations suggesting the monitoring of funded projects from inception to mitigate default risks (Han *et al.,* 2012).

The study aims to explore factor analysis, data mining modeling, credit scoring, and post-modeling processes. To ensure precision, significant factors determining the creditworthiness of applicants were incorporated into the model, including installment type, age, monthly expenses, job sector, payment method, and income-to-finance ratio. By presenting a systematic.and structured approach to developing a credit scoring model, this study contributes to advancing credit scoring methodologies. Based on the study's findings, banks can utilize this model to construct their credit scoring systems for assessing the creditworthiness of individual loan applicants. Implementing this model can help banks mitigate risks and enhance long-term operational efficiency in the credit system, facilitating informed decision-making processes.

The model empowers loan officers to automate decision-making and accurately predict the creditworthiness of applicants. However, the study acknowledges the exclusion of several variables during the model's development, including net worth, education level, number of dependents, other financial commitments, financing duration, and gender. Incorporating these variables in future iterations could potentially enhance the model's accuracy and predictive capabilities. Additionally, expanding the dataset to include information from other banks can further enhance the model's accuracy.

Lastly, the study emphasizes the importance of ongoing model maintenance to ensure its continued reliability in the dynamic business environment. Regular updates and adjustments are crucial to adapt the model to evolving market conditions and changing customer profiles (Sum *et al.,* 2022).

Table 1: Summarize the literature review.

Authors “year”	**Title of paper**	**Methodology**	**Result**
Randula Koralage, “2019”	Data Mining Techniques for Credit Card Fraud Detection	Bayes Net, J48, and Naïve Bayes the model was implemented and evaluated using the WEKA software	Bayes Net 73.87%, J48 78.38%, and Naïve Bayes 73.87%.
Jasmina Nalić and Amar Švraka, “2018”	Using data mining approaches to build credit scoring model: Case study-implementation of credit scoring model in microfinance institution.	J48 classifier algorithm using the WEKA software	J48 classifier algorithm performed relatively well in accurately classifying instances
Jia Wu, Sunil Vadera, Karl Dayson, Diane Burridge and Ian Clough “2010”	A comparison of data mining methods in microfinance	GLM algorithm using the Oracle Data Mining (ODM) software	GLM algorithm, demonstrated excellent results, with a predictive confidence of 97.437% and an overall average accuracy exceeding 98%
Jun-Ya Zeng, Jian-Bang Lin and Tian Wang “2017”	A new competing risks model for predicting prepayment and default using data mining	J48, EM, Naïve Bayes, K-means, Using WEKA software	J48 71.44%, EM 28.77%, Naïve Bayes 75.09%, and K-means 57.47%.
Dr. Md. Rashid Farooqi and Naiyar Iqbal “2017”	Effectiveness of Data mining in Banking Industry: An empirical study	J48 algorithm Using WEKA software	Displaying a correctly classified instances rate of 96.36% on a dataset with 3,466 instances
Rabihah Md, Waidah Ismail, Zul Hilmi Abdullah, and Nurul Fathihin Mohd Noor Shah “2022”	A New Efficient Credit Scoring Model For Personal Loan Using Data Mining Technique For Sustainability Management	Random Forest, KNN, and Naive Bayes classification algorithms Using The Orange data mining tool	Random Forest 94.6%, KNN 87.4%, and Naive Bayes 92.3%.

After reviewing previous research and studies, we've identified an intriguing area for further investigation: exploring how non-bank financial institutions can effectively integrate traditional and digital methods seamlessly. While existing studies primarily focused on challenges within the banking sector, such as payment defaults, installment failures, and risk assessment for decision-making, there's a noticeable gap regarding similar issues within non-banking sectors.

During our search, we found a lack of research or scientific papers addressing these points within non-banking sectors, including companies and non-governmental organizations catering to customers who may lack the financial capabilities or face procedural complexities typical in traditional banking settings. In contrast, these entities often offer faster financing solutions, sometimes disbursing funds within 24 hours.

Given Egypt's current focus on financial inclusiveness in the digital age and the proliferation of microfinance companies and non-governmental organizations in the country, there's an opportunity to explore how data mining algorithms can address specific challenges faced by these entities. One such challenge is accurately estimating the success of projects submitted for financing and predicting their likelihood of success or failure to mitigate non-payment or installment defaults.

DATA MINING PROCESS

The aim is to construct a practical predictive model through data mining techniques and represent the outcomes in a visually comprehensible manner. Fig. 1 illustrates the sequential steps of the data mining process utilized in this particular research investigation.

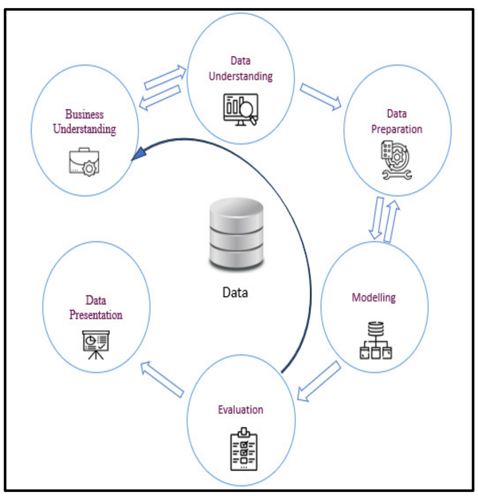


Fig. 1. The Cross-industry standard process for data mining.

BUSINESS UNDERSTANDING

In the initial phase of the CRISP-DM process, the primary focus is on comprehending the business objectives and constraints, maintaining equilibrium amidst various priorities within the organization, Tamweely. This stage is dedicated to pinpointing the Business Success Criteria: Here, clear and quantifiable standards are set to evaluate the success of the model pivotal factors that impact the data analysis objectives. A detailed plan is devised to accomplish both the data mining and business goals, encompassing the preliminary selection of tools and methodologies (Figueiredo *et al.,* 2023). from a business angle. Nevertheless, there are situations where subjective criteria are vital, like delivering valuable insights into customer relationships. Data Mining Success Criteria: This outlines the milestones for a successful outcome from a data mining viewpoint. For instance, attaining a specific level of predictive accuracy in the model.

DATA UNDERSTANDING

Data understanding and preparation are crucial factors influencing the results of data mining efforts. The effectiveness of the constructed model significantly depends on the depth and accuracy of data acquisition, examination, and preprocessing (Jackson, 2002). Thus, the following sections explore data understanding and the essential preprocessing tasks performed in this context.

Since the data typically originates from routine transactions collected for administrative purposes, it's crucial to evaluate the existing data landscape to identify relevant aspects and understand its nature.

In this context, the dataset was obtained from Tamweely Microfinance Institution, categorized as social data, and gathered from various branches spanning from 2018 to 2023. Initially, the primary objective was to consolidate the data into a unified repository, resulting in the accumulation of 534,639 records for preprocessing tasks. Although the dataset comprised 19 attributes in total, certain attributes contained numerous missing values, noises, and inconsistencies requiring resolution during the data cleaning phase of data preparation.

DATA PREPARATION

During this phase, data preparation involves various processes applied to the extracted data to enhance its suitability for the experiment and improve the overall data mining task. At this stage, the most crucial preprocessing tasks were carried out. These include data selection, data cleaning, and data aggregation/summarization, as outlined below.

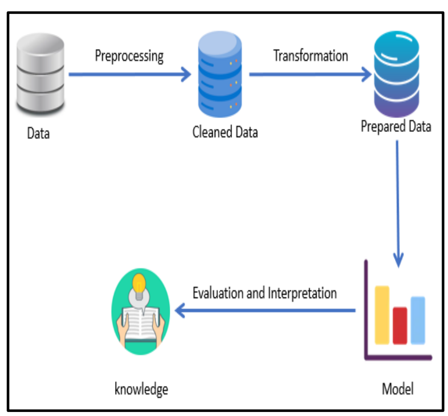


Fig. 2. Data Preparation.

DATA SELECTION

The main aim of data selection is to identify the suitable data type, source, and instrument(s). Initially, data reduction involves eliminating unnecessary or less important attributes from the original dataset. This process is based on the objective of the study at hand.
Microfinance data is obtained from Tamweely Microfinance from 2018 to 2023. In this dataset, there are 534639 instances, and 19 Attribute the Table 1 gives information about the data set.

DATA CLEANING

After attribute selection, the next step is data cleaning, which is applied to the dataset with the selected attributes. Data collected for the mining process often contains missing values, noise, or inconsistencies, which can result in the generation of unreliable information during the mining process. A high-quality data mining process typically produces efficient results, requiring preprocessing of the collected data to enhance its quality and, consequently, the mining outcomes (Aljawarneh *et al.,* 2019).
In this study, various standard data preprocessing tasks are conducted on the dataset, including data integration, data cleaning, data reduction, and data transformation. The initial step of data preprocessing involves Data Filtering, where relevant attributes necessary for prediction are selected from the company dataset. Since the dataset is unorganized, with features nested within each other, efforts a Are made to rearrange similar fields together to ensure accuracy. For example, all features related to monetary details are grouped, and likewise for premium-related features.
The subsequent task is handling missing data. The dataset contains missing and imputed data.which are addressed in this step. For instance, missing data in attributes such as "Total Amount," "Main Activity code," "Activity Type Code," and "Education Code" are handled by replacing the missing value with the mean of all samples belonging to the same class as the given tuples. As shown in the following figure (3,4,5,6).
Filling in missing values, and removing inconsistencies and noises were major data-cleaning activities undertaken at this stage of data preparation. Some fields had missing values, with 12,000 missed values from the "Total Amount" field and 7,600 missed values from the "Education Code" field. These values were considered most probable because they had the highest mode in the original dataset.

Table 2: Information about the data set (Class Attribute).

Flag	Description
Good	Disbursement of funding
Bad	Customer Reject
V-Bad	Reject final

Table 3: Information about the data set (Conditional Attribute).

Attribute	Description
BRCODE	Branch Code
DEBT_TYPE	Product Type
PRIMIUM_VALUE	The value of Installment
DEBT_PRD	Funding duration in months
APPROV_VALUE	Funding value
RATE	Annual interest
APP_FEE	Application submission fees
TOTAL_REQ_ AMTOUNT	Total Funding
REQ_NO_MONTHS	The payment period is in months
TOTAL_AMOUNT	Total Funding with interest
INDUSTRY_CODE	Industry code
MAIN_ACTIVITY	Main activity code
ACTIVITY_TYPE	Sub-activity code
GOV_ID	Governorate code
JOBCODE	Job Code
EDUCATION_CODE	Education Code
SEX	Gender
SCORE	Credit inquiry
OPEN_CREDITS	Number of open Funding

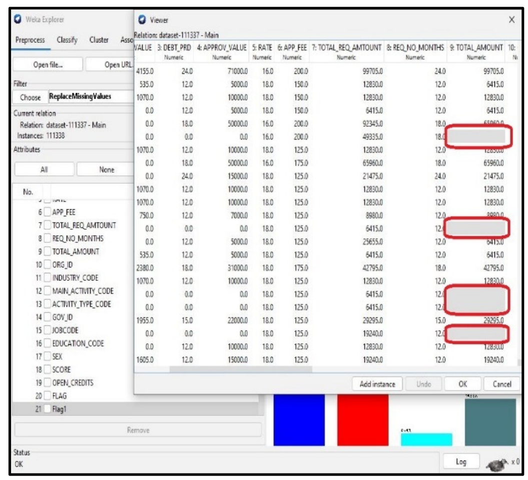


Fig. 3. Missing value [Total Amount].

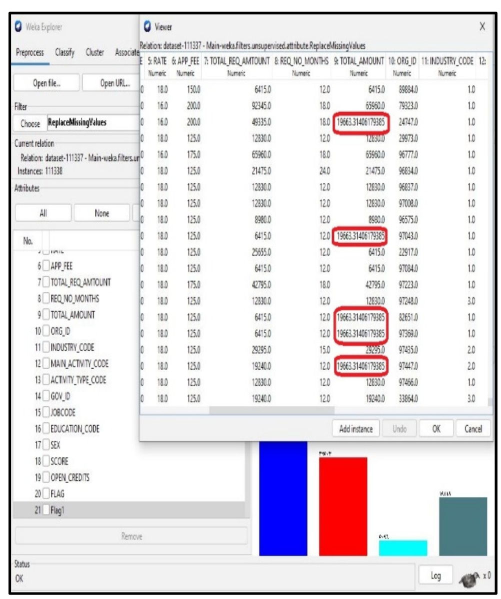


Fig. 4. Replacing missing value [Total Amount].

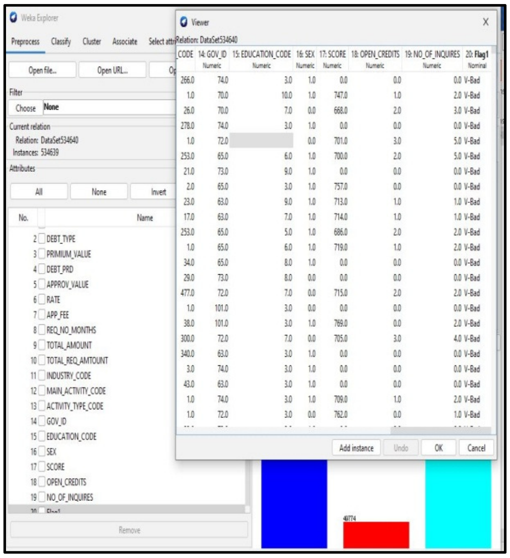


Fig. 5. Missing value value [Education code].

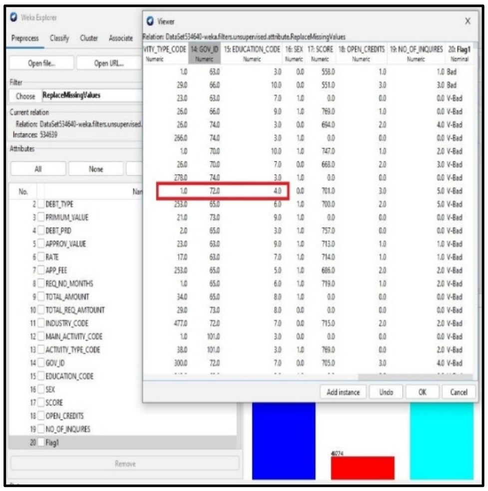


Fig. 6. Replacing missing value [Education code].

MODELLING

This study employed the J48 algorithm, a classification model utilized in data extraction within the Weka toolkit.

CLASSIFICATION

Two methodologies for data analysis are employed to create models for identifying significant categories and predicting future data patterns. These methodologies are referred to as Classification and Prediction.Classification models are crafted to predict categorical class labels, while prediction models are devised to predict continuous valued functions. For instance, a classification Model could be constructed to categorize bank loan applications as either safe or risky. Prediction involves the model's capability to accurately forecast the classification of incoming data. It assesses whether the model can appropriately classify the new data (Han *et al.,* 2012).

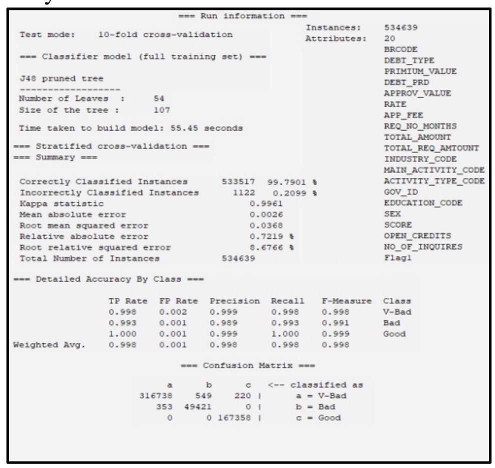
EXPERIMENTS RUN

The experiment has been carried out. This section delineates the various activities conducted regarding the implementation and evaluation of model-building experiments. We opted for the J48 classification algorithm to build the model to achieve higher precision. A tree classifier is particularly effective in determining whether an individual is a suitable applicant for a loan or if there's a high risk of default.

TYPES OF EXPERIMENT

A. Cross-Validation Method

During this experiment, the J48 Tree algorithm was implemented on the dataset utilizing all features and instances. The experiment employed 10-fold Cross-Validation, as depicted in Figure 7. The achieved accuracy was 99.7901%.



```
=== Run information ===
Test mode:    10-fold cross-validation          Instances:    534639
                                                Attributes:   20
                                                              BRCODE
=== Classifier model (full training set) ===                  DEBT_TYPE
                                                              PRIMIUM_VALUE
J48 pruned tree                                               DEBT_PRD
------------------                                            APPROV_VALUE
Number of Leaves  :     54                                    RATE
Size of the tree :      107                                   APP_FEE
                                                              REQ_NO_MONTHS
Time taken to build model: 55.45 seconds                      TOTAL_AMOUNT
                                                              TOTAL_REQ_AMTOUNT
=== Stratified cross-validation ===                           INDUSTRY_CODE
=== Summary ===                                               MAIN_ACTIVITY_CODE
                                                              ACTIVITY_TYPE_CODE
Correctly Classified Instances      533517   99.7901 %        GOV_ID
Incorrectly Classified Instances      1122    0.2099 %        EDUCATION_CODE
Kappa statistic                          0.9961               SEX
Mean absolute error                      0.0026               SCORE
Root mean squared error                  0.0368               OPEN_CREDITS
Relative absolute error                  0.7219 %             NO_OF_INQUIRES
Root relative squared error              8.6766 %             Flag1
Total Number of Instances           534639

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  Class
               0.998    0.002    0.999      0.998    0.998      V-Bad
               0.993    0.001    0.989      0.993    0.991      Bad
               1.000    0.001    0.999      1.000    0.999      Good
Weighted Avg.  0.998    0.001    0.998      0.998    0.998

=== Confusion Matrix ===

      a      b      c   <-- classified as
 316738    549    220 |     a = V-Bad
    353  49421      0 |     b = Bad
      0      0 167358 |     c = Good
```

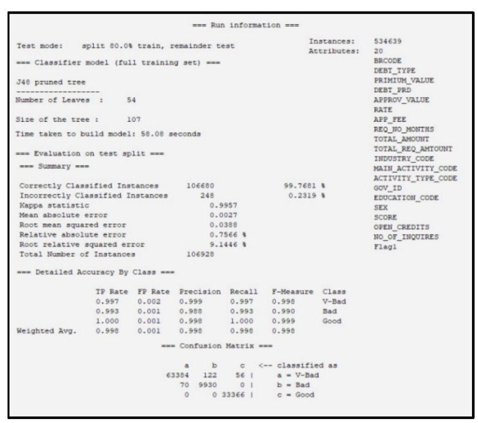
Fig. 7. J48 Tree (10-fold Cross-Validation.

B. Percentage Split Method

In this experiment, the J48 Tree algorithm was utilized on the dataset incorporating all features and instances. The experiment was repeated several times, altering the sizes of the training and test sets (80% training, 20% test - 70% training, 30% test - and 60% training, 40% test). The most favorable outcome was observed when the data was partitioned into 60% training and 40% test sets. The achieved accuracy was 99.79%.

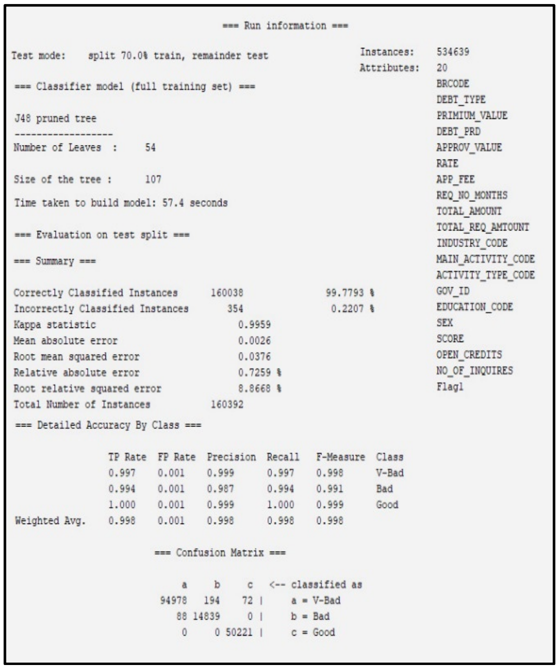
Table 4: Information about the Accuracy for J48 Tree.

	Training	Test	Accuracy
J48Tree algorithm	80%	20%	99.7681%
	70%	30%	99.7793%
	60%	40%	99.79%



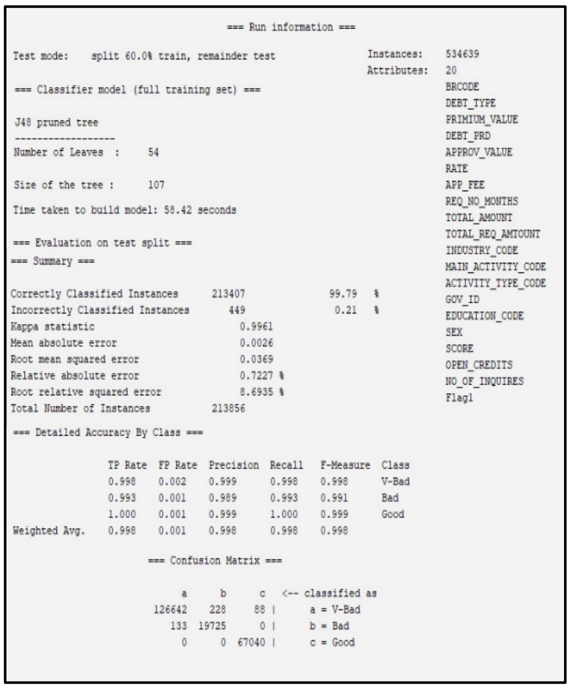
```
=== Run information ===
Test mode:    split 80.0% train, remainder test     Instances:    534639
                                                    Attributes:   20
=== Classifier model (full training set) ===                      BRCODE
                                                                  DEBT_TYPE
J48 pruned tree                                                   PRIMIUM_VALUE
------------------                                                DEBT_PRD
Number of Leaves  :     54                                        APPROV_VALUE
                                                                  RATE
Size of the tree :      107                                       APP_FEE
                                                                  REQ_NO_MONTHS
Time taken to build model: 58.08 seconds                          TOTAL_AMOUNT
                                                                  TOTAL_REQ_AMTOUNT
=== Evaluation on test split ===                                  INDUSTRY_CODE
 === Summary ===                                                  MAIN_ACTIVITY_CODE
                                                                  ACTIVITY_TYPE_CODE
 Correctly Classified Instances      106680          99.7681 %    GOV_ID
 Incorrectly Classified Instances       248           0.2319 %    EDUCATION_CODE
 Kappa statistic                          0.9957                  SEX
 Mean absolute error                      0.0027                  SCORE
 Root mean squared error                  0.0388                  OPEN_CREDITS
 Relative absolute error                  0.7566 %                NO_OF_INQUIRES
 Root relative squared error              9.1446 %                Flag1
 Total Number of Instances           106928

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  Class
               0.997    0.002    0.999      0.997    0.998      V-Bad
               0.993    0.001    0.988      0.993    0.990      Bad
               1.000    0.001    0.998      1.000    0.999      Good
Weighted Avg.  0.998    0.001    0.998      0.998    0.998

=== Confusion Matrix ===

     a     b     c   <-- classified as
 63384   122    56 |     a = V-Bad
    70  9930     0 |     b = Bad
     0     0 33366 |     c = Good
```

Fig. 8. J48 Tree (80% training, 20% test).



```
=== Run information ===

Test mode:    split 70.0% train, remainder test          Instances:    534639
                                                          Attributes:   20
=== Classifier model (full training set) ===                            BRCODE
                                                                        DEBT_TYPE
J48 pruned tree                                                         PRIMIUM_VALUE
------------------                                                      DEBT_PRD
Number of Leaves  :     54                                              APPROV_VALUE
                                                                        RATE
Size of the tree :      107                                             APP_FEE
                                                                        REQ_NO_MONTHS
Time taken to build model: 57.4 seconds                                 TOTAL_AMOUNT
                                                                        TOTAL_REQ_AMTOUNT
=== Evaluation on test split ===                                        INDUSTRY_CODE
                                                                        MAIN_ACTIVITY_CODE
=== Summary ===                                                         ACTIVITY_TYPE_CODE
                                                                        GOV_ID
Correctly Classified Instances      160038              99.7793 %       EDUCATION_CODE
Incorrectly Classified Instances       354               0.2207 %       SEX
Kappa statistic                          0.9959                         SCORE
Mean absolute error                      0.0026                         OPEN_CREDITS
Root mean squared error                  0.0376                         NO_OF_INQUIRES
Relative absolute error                  0.7259 %                       Flag1
Root relative squared error              8.8668 %
Total Number of Instances           160392

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  Class
                 0.997    0.001    0.999      0.997    0.998      V-Bad
                 0.994    0.001    0.987      0.994    0.991      Bad
                 1.000    0.001    0.999      1.000    0.999      Good
Weighted Avg.    0.998    0.001    0.998      0.998    0.998

                      === Confusion Matrix ===

                          a     b     c   <-- classified as
                      94978   194    72 |     a = V-Bad
                         88 14839     0 |     b = Bad
                          0     0 50221 |     c = Good
```

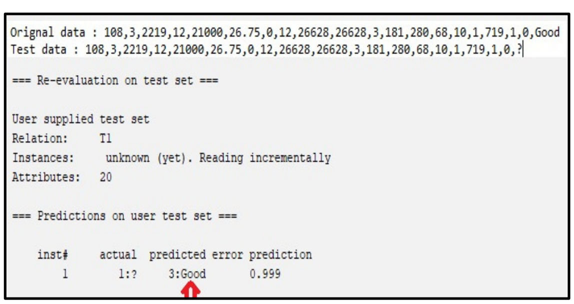
Fig. 9. J48 Tree (70% training, 30% test).



```
=== Run information ===

Test mode:    split 60.0% train, remainder test          Instances:    534639
                                                          Attributes:   20
=== Classifier model (full training set) ===                            BRCODE
                                                                        DEBT_TYPE
J48 pruned tree                                                         PRIMIUM_VALUE
------------------                                                      DEBT_PRD
Number of Leaves  :     54                                              APPROV_VALUE
                                                                        RATE
Size of the tree :      107                                             APP_FEE
                                                                        REQ_NO_MONTHS
Time taken to build model: 58.42 seconds                                TOTAL_AMOUNT
                                                                        TOTAL_REQ_AMTOUNT
=== Evaluation on test split ===                                        INDUSTRY_CODE
=== Summary ===                                                         MAIN_ACTIVITY_CODE
                                                                        ACTIVITY_TYPE_CODE
Correctly Classified Instances      213407              99.79   %       GOV_ID
Incorrectly Classified Instances       449               0.21   %       EDUCATION_CODE
Kappa statistic                          0.9961                         SEX
Mean absolute error                      0.0026                         SCORE
Root mean squared error                  0.0369                         OPEN_CREDITS
Relative absolute error                  0.7227 %                       NO_OF_INQUIRES
Root relative squared error              8.6935 %                       Flag1
Total Number of Instances           213856

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  Class
                 0.998    0.002    0.999      0.998    0.998      V-Bad
                 0.993    0.001    0.989      0.993    0.991      Bad
                 1.000    0.001    0.999      1.000    0.999      Good
Weighted Avg.    0.998    0.001    0.998      0.998    0.998

                    === Confusion Matrix ===

                         a      b      c   <-- classified as
                    126642    228     88 |     a = V-Bad
                       133  19725      0 |     b = Bad
                         0      0  67040 |     c = Good
```

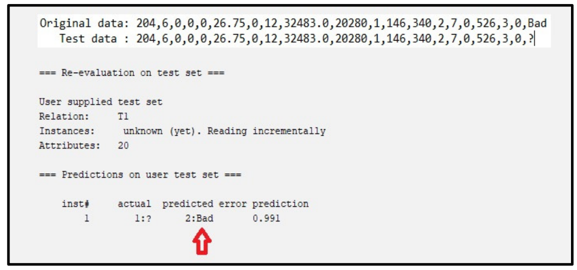
Fig. 10. J48 Tree (60% training, 40% test).

C. The result test of the experiment



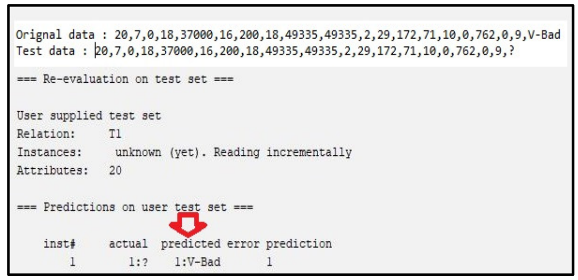
```
Orignal data : 108,3,2219,12,21000,26.75,0,12,26628,26628,3,181,280,68,10,1,719,1,0,Good
Test data : 108,3,2219,12,21000,26.75,0,12,26628,26628,3,181,280,68,10,1,719,1,0,?

=== Re-evaluation on test set ===

User supplied test set
Relation:     T1
Instances:     unknown (yet). Reading incrementally
Attributes:   20

=== Predictions on user test set ===

    inst#     actual  predicted error prediction
        1        1:?     3:Good       0.999
```

Fig. 11. Re-evaluation test (Good).



```
Original data: 204,6,0,0,0,26.75,0,12,32483.0,20280,1,146,340,2,7,0,526,3,0,Bad
   Test data : 204,6,0,0,0,26.75,0,12,32483.0,20280,1,146,340,2,7,0,526,3,0,?

=== Re-evaluation on test set ===

User supplied test set
Relation:     T1
Instances:     unknown (yet). Reading incrementally
Attributes:   20

=== Predictions on user test set ===

    inst#     actual  predicted error prediction
        1        1:?      2:Bad       0.991
```

Fig. 12. Re-evaluation test (Bod).



```
Orignal data : 20,7,0,18,37000,16,200,18,49335,49335,2,29,172,71,10,0,762,0,9,V-Bad
Test data : 20,7,0,18,37000,16,200,18,49335,49335,2,29,172,71,10,0,762,0,9,?

=== Re-evaluation on test set ===

User supplied test set
Relation:     T1
Instances:     unknown (yet). Reading incrementally
Attributes:   20

=== Predictions on user test set ===

    inst#     actual  predicted error prediction
        1        1:?    1:V-Bad       1
```

Fig. 13. Re-evaluation test (V-Bod).

EVALUATION AND DATA PRESENTATION

In this study, two approaches were employed to train and assess the model (cross-validation and percentage splits). As mentioned earlier, the main objective of developing the classification model is to discern patterns in each borrower's status, aiding in the prediction of a new borrower's status based on these characteristics. Various classification trees were examined using the J48 classifier algorithm, with the model achieving the highest accuracy among them. This experiment yielded the most favorable outcome compared to all other experiments conducted, primarily due to its superior accuracy level. Two methodologies were employed for training and testing the model: cross-validation and percentage splits (80-20, 70-30, 60-40). The findings indicated that cross-validation resulted in slightly higher accuracy compared to percentage splits, with an accuracy rate of 99.7901%. Overall, the accuracy at the model level was highly satisfactory. Precision peaked, and the confusion matrix displayed commendable outcomes. Furthermore, the classifier demonstrated robust performance when applied to a preprocessed dataset.

CONCLUSIONS

The paper utilizes data mining to develop a predictive model, focusing on the loan histories of existing borrowers. This model aims to aid in comparing potential loan applications by identifying characteristics indicative of a good or bad loan record, drawing from credit background and demographic profiles. Emphasis is placed on the importance of preprocessing or cleaning data to achieve higher accuracy rates. The results obtained using the J48 algorithm are particularly noteworthy, with a correctly classified instances rate of 99.7901%. Additionally, the preprocessing stage can reveal patterns useful for identifying target loan markets, devising income-enhancing strategies, reducing default risk, and improving loan products.

FUTURE SCOPE

Future research could further investigate advanced techniques in data mining and predictive analytics, aiming to refine the methodologies used and enhance their applicability across different sectors. Exploring these avenues will provide a more comprehensive understanding of customer behavior and microfinance dynamics, ultimately contributing to more effective and strategic business practices.

Acknowledgements. First of all, I would like to extend my special gratitude to Dr. Atef Raslan for his unreserved support and assistance throughout my research work. His constructive comments and suggestions, in my study and in writing the thesis, were highly valuable. In general, his helpful personality, cooperativeness, and dedication are remarkably worth mentioning.

My special thanks also go to my supervisor Prof. Dr. Lamiaa Fatouh for his continued efforts during all phases of the research.

REFERENCES

Abakar, M. I. (2020). Microfinance Data Analysis in Banking Sector Using Data Mining Techniques) Case Study: Agricultural Bank of Sudan). Sudan University.

Aljawarneh, S., Yassein, M. B., & Aljundi, M. (2019). An enhanced J48 classification algorithm for the anomaly intrusion detection systems. *Cluster Computing, 22,* 10549-10565.

Figueiredo, R., Magalhães, C., & Huber, C. (2023). How to predict the innovation to SMEs? Applying the data mining process to the spinner innovation model. *Social Sciences*, *12*(2), 75.

Hailemariam, G., Hill, S., & Demissie, S. (2012). Exploring data mining techniques and algorithms for predicting customer loyalty and loan default risk scenarios at wisdom microfinance, Addis Ababa, Ethiopia. In Proceedings of the International Conference on Management of Emergent Digital EcoSystems (pp. 183-184).

Han, J., Kamber, M., & Pei, J. (2012). Data mining concepts and techniques third edition. University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University.

Jackson, J. (2002). Data Mining; A Conceptual Overview. Communications of the Association for Information Systems. *Communications of the Association for Information System, 8.*

Kulkarni, E. G., & Kulkarni, R. B. (2016). Weka powerful tool in data mining. *International Journal of Computer Applications, 975,* 8887.

Lee, Y. C., & Wang, Y. (2020). Credit risk analysis of microcredit using the data of a chinese microfinance company and data mining techniques. *ICFE 2020,* 394.

Nalić, J., & Švraka, A. (2018). Using data mining approaches to build credit scoring model: Case study—implementation of credit scoring model in microfinance institution. In 2018 17th International Symposium Infoteh-Jahorina (INFOTEH) (pp. 1-5). IEEE.

Nguyen, C. (2019). The credit risk evaluation models: An application of data mining techniques, Association for Information Systems AIS Electronic Library (AISeL).

Subia, M. G., & Galapon, A. C. (2020). Sample model for the prediction of default risk of loan applications using data mining. *Journal of Journal of Scientific & Technology Research, 9*(6), 222-227.

Sum, R. M., Ismail, W., Abdullah, Z. H., Fathihin, N., Shah, M. N., & Hendradi, R. (2022). A New Efficient Credit Scoring Model for Personal Loan using Data Mining Technique Toward for Sustainability Management. *Journal of Sustainability Science and Management, 17*(5), 60-76.

Zeng, J. Y., Lin, J. B., & Wang, T. (2017). A new competing risks model for predicting prepayment and default using data mining. In 2017 2nd International Conference on Image, Vision and Computing (ICIVC) (pp. 985-989). IEEE.

How to cite this article: Sameh Ali, Atef Raslan and Lamiaa Fattouh (2024). Model for the Prediction of Default Risk of Funding Requests Using Data Mining. *International Journal on Emerging Technologies, 15*(2): 05–12.