Volume 4, No. 9 September 2026
p-ISSN 3032-3037 | e-ISSN 3031-5786

ADVANCES IN SOCIAL HUMANITIES RESEARCH

Business Analytics for Credit Risk Management in Rural Banks Using the CRISP-DM Methodology (Case Study: PT BPR Jabar Perseroda)

Ayip Muhdiatullah[1], Samidi[2]*, Cecep Taofiqurrochman[3]
Universitas Ekuitas Indonesia, Indonesia[1,3]
Universitas Budi Luhur, Indonesia[2]
Email: ayipmuhdiatullah@ekuitas.ac.id[1], samidi@budiluhur.ac.id[2]*,
cecep.taufiqurrohman@ekuitas.ac.id[3]

Keywords:
Business Analytics; Credit Risk Management; CRISP-DM; Machine Learning; Random Forest; Rural Banks (BPR); 5C Credit Principles.

Abstract
Credit risk remains a critical concern for rural banks because lending quality directly affects financial stability and institutional sustainability. Although banking databases contain valuable debtor information, credit assessments often continue to rely on administrative verification and subjective judgment. This study aimed to identify credit risk characteristics, compare the performance of Random Forest, Gradient Boosted Tree, and Random Tree algorithms, and formulate data-driven risk management strategies for PT BPR Jabar Perseroda. A quantitative descriptive research design was employed using the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, integrating descriptive, predictive, and prescriptive analytics. The sample consisted of 1,015 debtor records selected through stratified sampling from a population of 10,150 records. Model performance was evaluated using accuracy, precision, and recall metrics, while correlation analysis was conducted to identify variables associated with credit risk status. Payment delinquency history showed the strongest relationship with risk status ($r = 0.653$), whereas collateral demonstrated the weakest relationship ($r = -0.047$). The Random Forest algorithm achieved the best predictive performance, with 99.00% accuracy, 100.00% precision, and 97.83% recall. The findings indicated that repayment behavior provided more meaningful risk information than static administrative attributes. Therefore, integrating Random Forest-based predictions with the 5C credit assessment principles could strengthen objective credit evaluation, decision-support systems, early warning mechanisms, and proactive credit risk management.

INTRODUCTION

Bank Perekonomian Rakyat (BPR) is one of the financial intermediary institutions that plays an important role in supporting regional economic growth through credit distribution to the community, particularly the micro, small, and medium enterprises (MSMEs) sector, as well as consumer lending. As an institution whose primary revenue source is derived from credit activities, the quality of the credit portfolio is a key determinant of bank health and sustainability (Altman & Saunders, 1998). Therefore, the ability of BPRs to identify potential debtors with a high risk of default is an essential component of implementing prudent banking principles and effective credit risk management (Bougie & Sekaran, 2020).

On the other hand, the increasing volume of transaction data stored in Core Banking systems provides opportunities for the banking industry to adopt data-driven decision-making approaches to support credit analysis processes (Han et al., 2012). This approach can assist credit analysts in identifying risk patterns more objectively rather than relying solely on experience-based assessments or subjective judgments (Creswell & Creswell, 2018). However, in practice, many credit analysis processes in BPRs still emphasize administrative verification and have not fully utilized historical information available in operational databases as a basis for decision-making (Breiman, 2001a, 2001b).

This phenomenon was also reflected in the research data used in this study. The dataset from PT BPR Jabar Perseroda demonstrated that debtor risk characteristics were not equally influenced by all variables involved in the credit analysis process (Delen & Ram, 2018). The results of the correlation matrix analysis revealed differences in the strength of relationships between each variable and debtor risk status (Chang et al., 2024).

The phenomenon illustrated above indicated that credit risk characteristics in the research object were not solely determined by administrative factors, such as collateral availability or credit limits, but were more strongly associated with debtor repayment behavior. From a business perspective, this condition suggested that credit analysis processes focusing only on administrative aspects might not adequately represent debtor risk levels (Hair et al., 2019). Therefore, an analytical approach was required to identify the factors most strongly associated with credit risk, enabling a more objective, measurable, and data-driven decision-making process (Tan et al., n.d.).

Furthermore, the utilization of historical data stored in Core Banking systems provided opportunities for BPRs not only to understand existing risk conditions through descriptive analytics but also to develop credit risk prediction models through predictive analytics and translate analytical results into business strategy recommendations through prescriptive analytics. Thus, credit risk management could shift from a reactive approach toward a more proactive approach supported by analytical techniques and data-driven decision-making.

In addition to the empirical phenomena observed in the BPR industry, previous studies have also indicated a theoretical gap in the development of credit risk analysis models. Recent developments in credit risk research have primarily focused on improving predictive model performance through the application of various machine learning algorithms (Mitchell, 1997), including Random Forest, Gradient Boosting (Friedman, 2001a, 2001b), XGBoost, LightGBM, and ensemble learning methods. The effectiveness of these studies has generally been evaluated based on classification model performance indicators, such as accuracy, precision, recall, and Area Under the Curve (AUC).

(Muñoz-Cancino et al., 2023) showed that the use of credit history and behavioral variables improved models' ability to predict creditworthiness. The study concluded that payment history was the most dominant factor in determining credit risk. (Liu et al., 2022) developed a Tree-Enhanced Gradient Boosting approach combined with TreeSHAP to improve the interpretability of credit scoring models (Sharda et al., 2018). The study enhanced model explainability but remained focused on interpreting prediction outcomes. (Dernsjö, 2023) also demonstrated that Gradient Boosting Decision Tree algorithms provided strong performance in behavioral credit scoring, particularly in improving the classification of debtor risk based on repayment behavior.

On the other hand, (Ayari et al., 2025), through a systematic literature review, concluded that Random Forest, Gradient Boosting, XGBoost, and various ensemble learning methods were among the most widely used algorithms in modern credit scoring research due to their ability to achieve high predictive accuracy. However, the study also indicated that most existing research remained focused on developing predictive analytics models, while studies connecting prediction outcomes with business decision-making through prescriptive analytics remained relatively limited (Géron, 2022).

This condition demonstrated a theoretical gap in existing research. Most previous studies have concentrated on selecting optimal algorithms or improving classification model performance, while predictive outputs have rarely been translated into strategic recommendations applicable to banking business processes. In other words, previous studies have been predominantly model-centric, focusing on producing highly accurate prediction models, rather than decision-centric, emphasizing how analytical results can support managerial decision-making (Xia et al., 2017).

This study was developed to address this gap by comprehensively integrating a Business Analytics approach through the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. Unlike previous studies that primarily focused on the predictive analytics stage, this study integrated three analytical stages: descriptive analytics, predictive analytics, and prescriptive analytics within a unified analytical framework. The descriptive analytics stage was used to characterize credit risk patterns, the predictive analytics stage was applied to compare the performance of Random Forest, Gradient Boosted Tree, and Random Tree algorithms, and the prescriptive analytics stage utilized correlation analysis results as a basis for formulating credit risk management recommendations aligned with the 5C principles (Character, Capacity, Capital, Collateral, and Condition of Economy).

Therefore, the primary contribution of this study lies not only in identifying the best-performing prediction model but also in developing a Business Analytics framework capable of connecting data analysis results with business decision-making processes (Greuning & Bratanovic, 2020). This approach enabled machine learning outputs to extend beyond predictive information and be translated into operational strategies that support improvements in credit analysis quality and risk management practices at BPRs (Kohavi, 1996). Accordingly, this study was entitled "Business Analytics for Credit Risk Management in Rural Banks Using the CRISP-DM Methodology (Case Study: PT BPR Jabar Perseroda)."

This study examined credit risk management challenges in BPRs, particularly limitations in utilizing historical data, ensuring analytical objectivity, implementing machine learning techniques, and integrating descriptive, predictive, and prescriptive analytics. The study aimed to identify credit risk characteristics, compare the performance of Random Forest, Gradient Boosted Tree, and Random Tree algorithms based on accuracy, precision, and recall metrics, and formulate risk management strategies using

Business Analytics. Theoretically, this study contributed to the literature on Business Analytics, machine learning, and credit risk management through the application of CRISP-DM. Practically, the findings were expected to support more objective credit decision-making, the development of Decision Support Systems and Early Warning Systems, and the improvement of credit risk management quality in BPRs.

METHOD

Research Design

This study employed a quantitative descriptive method to analyze credit risk in Bank Perekonomian Rakyat (BPR). The analysis applied a Business Analytics approach using the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, consisting of business understanding, data understanding, data preparation, modeling, evaluation, and deployment stages. Descriptive analysis was conducted to identify data characteristics, while quantitative analysis was applied to develop and evaluate credit risk prediction models.

The descriptive analytics stage involved analyzing variables including PlafondJT, Debt-to-Income Ratio (DTI), collateral, late payment history, permanent employment status, and risk status using descriptive statistics (Evans, 1996). The predictive analytics stage involved developing classification models using the Random Forest, Gradient Boosted Tree, and Random Tree algorithms. Model performance was evaluated and compared based on accuracy, precision, and recall metrics derived from the confusion matrix. The best-performing model was subsequently used as the basis for analysis and the formulation of credit risk management recommendations.

Population, Sample, and Unit of Analysis

The research population is all credit debtor data of PT BPR Jabar Perseroda recorded in the Core Banking system with a data position of December 31, 2025, which is as many as 10,150 debtor data. The study sample amounted to 1,015 data or 10% of the selected population using the Stratified Sampling technique by maintaining the proportion of characteristics and Risk Status classes in the population (Cochran, 1977).

The unit of analysis of the research is each credit debtor recorded in the Core Banking system. Each debtor is an observation that contains information about the JT Ceiling, DTI Ratio, Collateral, Late History, Permanent Employment, and Risk Status.

Data Collection Sources and Techniques

The research uses secondary data obtained from the Core Banking system of PT BPR Jabar Perseroda through the process of extracting credit debtor data as of December 31, 2025. The data then goes through a data cleaning process, attribute selection, and data preparation before being used in modeling. In addition, the research uses literature studies from books, scientific journals, regulations, and previous research related to Business Analytics, Credit Risk Management, Machine Learning, and CRISP-DM.

The data collection technique consists of documentation and literature studies. Documentation is carried out by taking debtor data from the Core Banking system, while literature studies are carried out to obtain theoretical foundations and references that support the research.

Data Processing and Analysis Techniques

Data processing and analysis was carried out using RapidMiner Studio with a Business Analytics approach and the CRISP-DM methodology. The process begins with data preparation, including data analysis, data type validation, categorical variable coding, and data division into training sets and testing sets. Furthermore, three classification models were built, namely Random Forest, Gradient Boosted Tree, and Random Tree.

The model evaluation was carried out using a Confusion Matrix with Accuracy, Precision, and Recall indicators. Accuracy is used to measure the overall accuracy of predictions, Precision to measure the accuracy of risky debtors' predictions, and Recall to measure the model's ability to detect truly risky debtors. The best model is determined based on the performance of the three indicators, considering Recall as an important indicator to minimize False Negatives in credit risk management.

The next best model is used in the Prescriptive Analytics stage. Correlation Matrix analysis was carried out to determine the relationship between the research variables and the Risk Status. The results of the analysis are then used to prepare credit risk management recommendations based on the 5C principles, namely Character, Capacity, Capital, Collateral, and Condition of Economy. Thus, the research not only produces predictive models, but also recommendations that can support data-driven credit decision-making.

Research Location and Time

The research was carried out at PT BPR Jabar Perseroda using credit debtor data obtained from the Core Banking system. The data used is cross-sectional data with a position of December 31, 2025. The entire processing and analysis process is carried out while still paying attention to the confidentiality of customer data and the company's data governance provisions.

The implementation of the research took place from January 2026 to July 2026, which included problem identification, proposal preparation, literature study, data collection and extraction, data preparation, modeling using RapidMiner, analysis of results, preparation of conclusions and suggestions, and completion of thesis.

RESULTS AND DISCUSSION

Predictive Model Implementation

After obtaining the best algorithm, namely Random Forest, the model was implemented through prediction simulations on new debtor data that did not have Risk Status information (Chapman et al., 2000). The implementation was carried out using the Apply Model operator on RapidMiner with the attributes of CeilingJT, DTI Ratio, Guarantee, Late History, and Fixed Work as predictor variables. Models that have been trained using historical data are applied to the simulation dataset to generate a classification of debtors in the Risk or No Risk category (Berger & Udell, 1990). This stage aims to prove that the model not only performs well on test data, but can also be used on new data under operational conditions (Nallakaruppan et al., 2024).

Based on the results of the evaluation, Random Forest was selected as the best model with an Accuracy of 99.00%, Precision of 100.00%, and a Recall of 97.83%, so it has excellent ability to classify credit risks (Rose & Hudgins, 2021). The implementation process is carried out by taking historical datasets, assigning RiskStatus as labels, building a Random Forest model, and then applying it to a simulation dataset whose RiskStatus value is not yet known. The model studies patterns from historical data and generates risk predictions for each debtor based on a combination of available credit characteristics (Provost & Fawcett, 2013).

The results of the implementation show that Random Forest can provide initial information about the risk of potential debtors so that it has the potential to be used as a tool in the credit analysis process. This model does not replace the decisions of credit analysts or Credit Committees, but it can support more objective and data-driven decision-making. Thus, the model has the potential to be integrated into PT BPR Jabar Perseroda's credit information system as a Decision Support System (DSS) and Early Warning System (EWS), so that the application of Business Analytics does not stop at model development and evaluation, but can be continued to the implementation stage according to the CRISP-DM methodology (Indonesia, 2023).

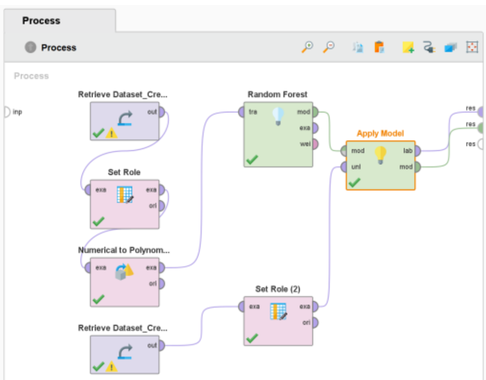


Figure 1. Implementation of the Random Forest Model for Predicting Debtors' Risk Status

Source: RapidMiner Data Processing Results (2026)

Figure 1 shows the flow of the implementation of the Random Forest model in the RapidMiner application. Historical datasets are used as training datasets to build prediction models, while simulation datasets are used as predicted data (unlabeled datasets). Once the Random Forest model is formed, the Apply Model operator applies the model to the simulation dataset resulting in a RiskStatus prediction for each debtor. This process describes the implementation of the model under operational conditions, i.e. helping credit analysts identify the risk level of potential debtors before credit decisions are made.

After the training process, the Random Forest algorithm succeeded in forming a prediction model consisting of 100 decision trees. Each tree is built using a different data sample (bootstrap sampling) and selects a different combination of attributes in each node splitting process. This approach is a key characteristic of the Random Forest algorithm which aims to improve the generalization capabilities of the model while reducing the risk of overfitting that generally occurs with the use of a single decision tree.

In this study, the main parameters used in the formation of the Random Forest model include the number of trees (number of trees) as many as 100, the criterion using the Gain Ratio, and the maximum depth of 10. The use of 100 decision trees is intended to enable the model to be able to study patterns of relationships between variables more stably through the ensemble learning mechanism, while the use of Gain Ratio aims to select the separating attributes that have the best ability to distinguish between Risk and Non-Risk classes. In addition, the maximum depth is limited to 10 levels to maintain a balance between the complexity of the model and the generalizability of new data.

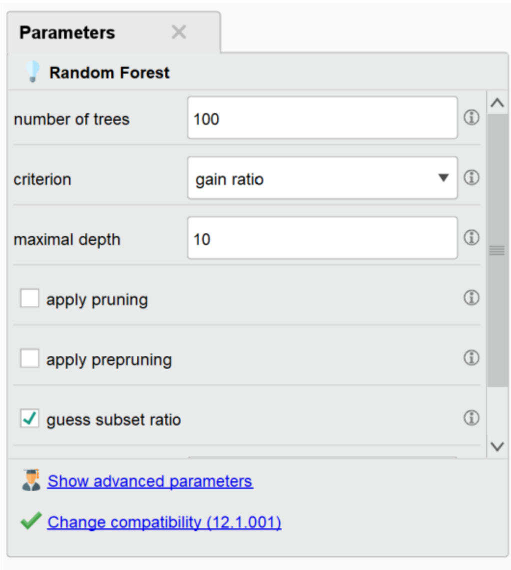


Figure 2. Altair Studio Parameters
Source: RapidMiner Data Processing Results (2026)

Table 1. Random Forest Model Parameters

Parameters	Value	Function
Number of Trees	100	Forming 100 decision trees as an ensemble model so as to improve prediction stability.
Criterion	Gain Ratio	Determine the best attributes for each node based on the *information gain ratio*.
Maximum Depth	10	Limiting the depth of the tree so that the model is less complex and reduces the potential for *overfitting*.
Apply Pruning	Not used	The trees are allowed to grow according to the set parameters as *Random Forest* controls the complexity through an ensemble mechanism.
Guess Subset Ratio	Active	Specifies the number of attributes randomly selected on each node during the tree formation process.

Source: Processing results using RapidMiner (2026).

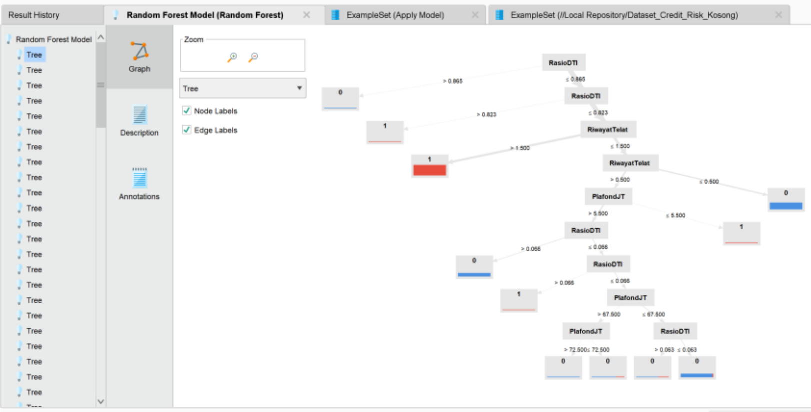


Figure 3. Interpretation of One of the Decision Trees
Source: RapidMiner Researcher Data Processing (2026)

Figure 3 shows one of the decision trees formed from all 100 trees in the Random Forest algorithm. The image is only one representation of the tree structure used in the prediction process, while the final decision of the model is obtained through the majority voting mechanism of all trees built.

It can be seen that the DTI Ratio attribute is the first variable used as the root node, which shows that in that tree the DTI Ratio has the best ability to separate data based on the Gain Ratio value. Furthermore, the classification process is carried out through a series of separations using the Late History and CeilingJT attributes until it reaches the leaf node that represents the StatusRisk class, namely 0 (No Risk) or 1 (Risky).

For example, one of the decision rules formed in the tree can be interpreted as follows: If the DTI Ratio is ≤ 0.823 and the Late History > 1.5, the debtor is predicted to be included in the Risk category (Risk Status = 1). Another example rule shows that: If the Late History ≤ 0.5, the debtor tends to be predicted as No Risk (Risk Status = 0).

In addition, in some branches of the tree, it can be seen that the PlafondJT variable is used as a follow-up attribute after the Late History, which shows that the size of the credit ceiling plays a supporting factor in the classification process if the debtor's payment behavior is known. This indicates that the model's decision is not only influenced by one attribute, but is a combination of several characteristics of the debtor.

However, the interpretation of a single decision tree cannot be used to describe the entire Random Forest model. This is because the Random Forest algorithm produces predictions based on the aggregation of classification results from all decision trees (ensemble prediction). Therefore, the tree structure in Figure 4.13 is only used to provide an overview of the algorithm's working mechanism, while the final decision of the model is the result of consensus (majority voting) of all the trees that are formed.

Overall, the visualization results show that the classification process is carried out through a combination of several attributes, especially DTI Ratio, Late History, and CeilingJT. These findings are in line with the results of previous analyses which showed that these variables contributed to the formation of credit risk prediction models. However, the level of influence of each variable on the Risk Status will be analyzed more comprehensively at the Correlation Matrix stage, which is the basis for the preparation of credit risk management strategy recommendations.

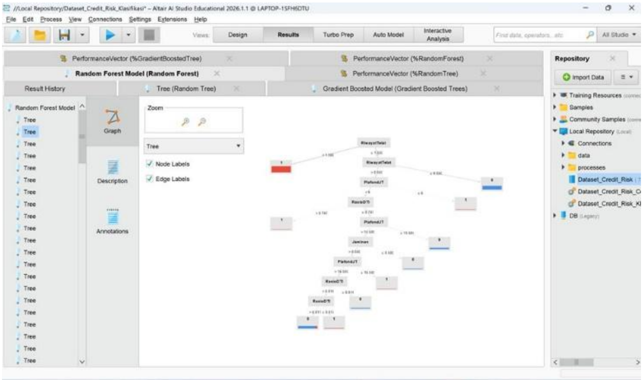


Figure 4. Visualization of One of the Decision Trees in the Random Forest Model
Source: RapidMiner Researcher Data Processing (2026)

Figure 4 shows one of the 100 decision trees built by the Random Forest algorithm. Each node shows the attributes used as the basis for data separation along with the threshold value, while each leaf node shows the final classification result in the form of RiskStatus = 0 (No Risk) or RiskStatus = 1 (Risky). The final decision of the Random Forest model is not determined by a single tree, but is the result of a majority vote of all decision trees built during the training process.

Prediction Simulation Results

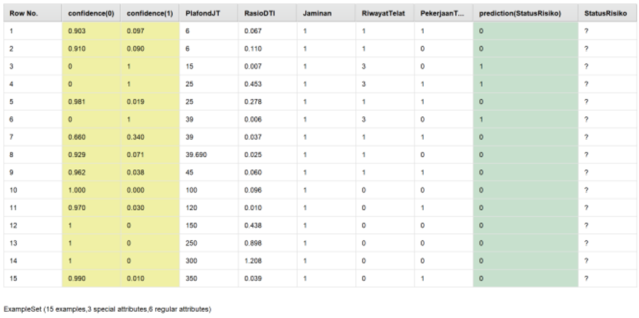
After the Random Forest model has been successfully built and selected as the best model based on the results of the evaluation, the next stage is to apply the model to the simulation dataset to predict the debtor's Risk Status. This process aims to determine the model's ability to classify new data that previously did not have information about the level of credit risk.

Figure 4.19 shows the prediction results generated by the Apply Model operator on the RapidMiner application. The main output of the implementation process is shown in the prediction column (StatusRisk), while the confidence (0) and confidence (1) columns show the confidence score of the model for each prediction result.

In this study, the prediction value (Status Risk) consists of two categories, namely 0 indicating Not Risk, and 1 indicating Risk. This value is the result of the classification of the Random Forest model based on the pattern of relationships between variables studied from historical data of PT BPR Jabar Perseroda.

In addition to producing prediction classes, the model also provides a confidence value that shows the model's level of confidence in the classification results. For example, in the first observation, a confidence value (0) of 0.903 and a confidence (1) of 0.097 was obtained, so that the model predicted the debtor as Not at Risk (Risk Status = 0) with a confidence level of around 90.3%. On the other hand, in the third observation, a confidence value (1) of 1,000 was obtained, so that the model classified the debtor as Risky (Risk Status = 1) with full confidence level.

The prediction results showed that out of 15 simulation data, as many as 12 debtors (80%) were predicted to be in the category of No Risk (Risk Status = 0), while 3 debtors (20%) were predicted as Risk (Risk Status = 1). Debtors who are predicted to be at risk are in the 3rd, 4th, and 6th observations.



Row No.	confidence(0)	confidence(1)	PlafondJT	RasioDTI	Jaminan	RiwayatTelat	PekerjaanT...	prediction(StatusRisiko)	StatusRisiko
1	0.903	0.097	6	0.067	1	1	1	0	?
2	0.910	0.090	6	0.110	1	1	0	0	?
3	0	1	15	0.007	1	3	0	1	?
4	0	1	25	0.453	1	3	1	1	?
5	0.981	0.019	25	0.278	1	1	1	0	?
6	0	1	39	0.006	1	3	0	1	?
7	0.660	0.340	39	0.037	1	1	1	0	?
8	0.929	0.071	39.690	0.025	1	1	0	0	?
9	0.962	0.038	45	0.060	1	1	1	0	?
10	1.000	0.000	100	0.096	1	0	0	0	?
11	0.970	0.030	120	0.010	1	0	1	0	?
12	1	0	150	0.438	1	0	0	0	?
13	1	0	250	0.898	1	0	0	0	?
14	1	0	300	1.208	1	0	0	0	?
15	0.990	0.010	350	0.039	1	0	1	0	?

ExampleSet (15 examples,3 special attributes,6 regular attributes)

Figure 5. Model Prediction Results on Simulation Data
Source: RapidMiner Researcher Data Processing (2026)

Table 2. Summary of Random Forest Prediction Results

Categories Predictions	Quantity	Percentage
No Risk (0)	12	80,00%
Risky (1)	3	20,00%
Total	15	100,00%

Source: Processing results using RapidMiner (2026).

Based on the results of these predictions, it can be seen that the Random Forest model is able to distinguish debtors based on the characteristics of the risk they have. Interestingly, all debtors who are predicted to be at Risk have a Late History value of 3, while most debtors who are predicted to be Not at Risk have a Late History value of 0 or 1. These findings are consistent with the results of the Correlation Matrix analysis which shows that Late History is the variable with the strongest relationship to StatusRisk.

In addition, the model does not consider only one variable in the classification process. For example, the fourth observation is predicted to be Risky despite having a Permanent and Guaranteed Job, because the combination of a high Late History and a relatively large DTI Ratio causes the probability of risk to increase. On the other hand, the tenth to fifteenth observations are still predicted to be Riskless even though they have a relatively high Ceiling value, because they are supported by Late History of 0, thus showing good payment behavior. These findings indicate that the Random Forest model classifies based on a combination of multiple attributes, rather than just one variable separately.

Overall, the implementation results show that the Random Forest model is able to be used to predict the Risk Status on debtor data that does not yet have a label. This ability shows that the built model not only has high performance at the evaluation stage, but can also be applied in operational conditions as a decision-making tool (Decision Support System) or as part of the Early Warning System (EWS) in the credit analysis process of

PT BPR Jabar Perseroda. Thus, the results of the resulting prediction can be supporting information for credit analysts in conducting early identification of potential debtors who have the potential to have higher credit risks.

The following is a table of business interpretations so that the discussion does not only stop at the results of technical predictions but can provide business recommendations.

Table 3. Business Interpretation of Predicted Results

Model Prediction Results	Business Interpretation	Follow-up
RiskStatus = 0 (No Risk)	The debtor has characteristics that match the historical pattern of the debtor performing well.	The credit analysis process can be continued according to normal procedures while still paying attention to the applicable credit policies.
StatusRisk = 1 (Risky)	Debtors have characteristics that resemble the debtor's historical pattern with potential delays or defaults.	A more in-depth credit analysis, additional verification, or risk mitigation is required before a credit decision is made.

Source: Researcher Data Processing (2026)

The table above makes it clear that the output model is not intended to replace the credit analyst's decision, but rather to be a decision-making tool. A presentation like this will further emphasize the practical contribution of your research as an implementation of Business Analytics that supports business processes at PT BPR Jabar Perseroda.

Discussion of the Results of Predictive Model Evaluation Based on Previous Research

The results of the predictive model evaluation showed that the Random Forest algorithm is the best model in predicting the debtor's Risk Status with an accuracy rate of 99.00%, accuracy of 100.00%, and recall of 97.83%. The second position is occupied by the Gradient Boosted Tree with an accuracy of 98.00%, while the Random Tree only gets an accuracy of 68.00%. These findings show that ensemble learning-based algorithms have better classification capabilities than single decision tree-based algorithms. This ability is due to the mechanism of forming multiple decision trees that is able to reduce model variance and improve generalization capabilities to new data.

The findings of this study are in line with the research (Liu et al, 2022) that developed the Tree-Enhanced Gradient Boosting model for credit scoring. The study shows that gradient boosting-based algorithms have very high performance in predicting credit risk because they are able to capture nonlinear relationships between variables while maintaining model interpretability through the TreeSHAP approach. Research (Liu et al, 2022) also confirms that ensemble tree-based models produce a good balance between accuracy, computational efficiency, and interpretability making them feasible for use in modern credit scoring systems. The results of this study reinforce these findings because two ensemble learning-based algorithms, namely Random Forest and Gradient Boosted Tree, both show a much higher level of accuracy than Random Tree. In contrast, the study (Liu et al, 2022) focused on developing the Gradient Boosting architecture as the main model, while this study compared the three algorithms directly and showed that Random Forest performed best on the characteristics of the BPR credit dataset.

The results of this study also have similarities with a study (Dernsjö, 2023) that evaluated the use of the Gradient Boosting Decision Tree (GBDT) in behavioral credit scoring. The study shows that boosting-based algorithms are able to provide better prediction performance than traditional approaches such as Logistic Regression (Montgomery et al., 2021), especially in utilizing payment behavior information to predict the probability of default. The similarity of the research lies in the use of tree-based machine learning algorithms as a credit risk prediction model. However, there is a fundamental difference, namely that Dernsjö's research only evaluated the performance of various variants of Gradient Boosting, while this study conducted a comparative evaluation between Random Forest, Gradient Boosted Tree, and Random Tree so that it was able to show empirically which algorithm is most suitable to be applied in the environment of the People's Economic Bank.

Furthermore, the results of this study are also consistent with the research The Ensemble Supervised Machine learning for Credit Risk Classification (2024) which concludes that algorithms based on Random Forest, Gradient Boosting, and XGBoost are the most effective approaches in credit risk modeling because they are able to identify nonlinear relationships between variables and produce a more stable level of classification than conventional algorithms. The study emphasizes that the ensemble learning method is very suitable to be applied to credit data because it has the ability to reduce overfitting and increase prediction accuracy. The results of this study support this conclusion, where two ensemble algorithms, namely Random Forest and Gradient Boosted Tree, show much higher performance than Random Tree as a representation of algorithms based on one decision tree.

However, there are interesting differences between this study and some previous studies. Most credit scoring research only stops at the stage of selecting the best algorithm based on performance indicators such as accuracy, precision, recall, AUC, or F1-score. Instead, this study not only evaluates the performance of predictive models, but also continues the results of the evaluation into the Prescriptive Analytics stage. The best model obtained, namely Random Forest, is not only used as a classification model, but also used as a basis for evaluating the relationship between variables using the Correlation Matrix and subsequently translated into business strategy recommendations based on the 5C principles (Character, Capacity, Capital, Collateral, and Condition of Economy). Thus, the results of the research do not stop at the technical aspects of machine learning, but produce policy recommendations that can be implemented in the BPR credit business process.

From a methodological perspective, this study also has different characteristics compared to credit scoring research in general. Most of the machine learning research uses a purely algorithmic experimental approach, while this study integrates the CRISP-DM methodology with a Business Analytics framework that includes Descriptive Analytics, Predictive Analytics, and Prescriptive Analytics. This approach results in a more comprehensive analysis flow, starting from identifying data characteristics, selecting the best predictive model, evaluating the relationship between variables, to preparing a business implementation strategy that is tailored to the credit risk management process at BPR.

Based on a comparison with previous research, it can be concluded that this research has a scientific contribution (novelty) in three main aspects. First, this study compares three decision tree-based algorithms (Random Forest, Gradient Boosted Tree, and Random Tree) in the context of BPR credit risk and shows that Random Forest is the best-performing model. Second, this study not only evaluates the model using classification performance indicators, but also integrates the evaluation results with the Correlation Matrix to identify the main factors that shape credit risk. Third, the results of the evaluation are translated into a prescriptive strategy based on the 5C principle along with its implementation roadmap in BPR business processes. Thus, this study not only produces an accurate credit risk prediction model, but also provides an implementation framework that can support data-driven decision-making in credit risk management.

Table 4. Comparative Synthesis of Research

Aspects	Liu et al. (2022)	Dernsjö (2023)	Supervised ML Set (2024)	This research
Comparing multiple algorithms	Partial	No	Partial	Yes
Random Forest tested	Not dominant	No	Yes	Yes
Gradient Boosted Tree tested	Yes	Yes	Yes	Yes
Random Tree tested	No	No	No	Yes
Evaluation of Accuracy, Precision, Recall	Yes	Yes	Yes	Yes
Using CRISP-DM	No	No	Partial	Yes
Prescriptive Analytics	No	No	No	Yes
5C-based strategy	No	No	No	Yes
Business implementation roadmap	No	No	No	Yes (Novelty)

Source: Results of Researcher Data Processing (2026)

Based on this synthesis, the novelty of this research does not lie in the use of the Random Forest algorithm alone, because the algorithm has been widely used in credit scoring research. The main contribution of this research is the integration between the CRISP-DM methodology and the Business Analytics framework. The research does not stop at identifying the best predictive model, but continues the evaluation results into the Correlation Matrix, developing a prescriptive strategy based on the 5C principles, and developing an implementation roadmap that establishes work programs, person in charge (PIC), performance indicators (KPIs), and expected outcomes for BPRs. This end-to-end approach makes a stronger practical contribution than credit scoring research that focuses on algorithmic evaluation alone, as the results of the analysis can be directly translated into operational policies that support data-driven credit risk management.

Discussion of Correlation Matrix Results Based on Previous Research

The results of the Correlation Matrix analysis showed that the Late History variable had the strongest relationship with Risk Status with a correlation coefficient of

0.653, followed by Ceiling JT (-0.354), DTI Ratio (-0.162), Permanent Employment (-0.120), and Guarantee (-0.047). These findings show that historical payment behavior is the most dominant factor in explaining credit risk compared to administrative characteristics and the existence of collateral. From a Business Analytics perspective, these results show that debtor behavior indicators have a higher ability to explain credit risk than static variables, so they can be the main basis in the preparation of credit risk management strategies in BPRs.

These findings are in line with research (Muñoz-Cancino et al., 2023) which shows that credit history and repayment behavior are the factors that contribute the most to improving the performance of creditworthiness assessments. The study found that the information value of credit history increases significantly in the early months of the credit relationship and then stabilizes once adequate payment history is available. In other words, debtor payment behavior is a stronger predictor than demographic or administrative characteristics. The results of this study strengthen these findings because the Late History variable is also the variable with the strongest relationship to Risk Status, thus showing that payment history is the main indicator in identifying potential non-performing loans.

Another finding that has similarities is a study (Alamsyah et al., 2025) that developed a credit risk assessment model using alternative data. The research shows that modern approaches to credit scoring are starting to reduce dependence on traditional variables such as the existence of collateral and emphasize more on behavioral variables that are able to describe the characteristics of debtors more accurately. The study showed that behavioral data sources contribute more to the quality of predictions than relying solely on administrative information. These findings are consistent with the results of this study, where the Guarantee variable only has a correlation of -0.047, while Late History is the most dominant variable. Thus, both studies show a paradigm shift from collateral-based lending to behavioral credit assessment in modern credit risk management.

The results of this study are also in line with the findings in the systematic literature review by (Ayari et al., 2025) which concluded that the development of machine learning in credit scoring shows a tendency to increase the role of payment behavior variables, credit history, and transaction characteristics compared to administrative variables alone. The study confirms that modern credit scoring models not only pursue the level of prediction accuracy, but also seek to identify factors that actually add value to credit decision-making. The findings of this study support this conclusion because the results of the Correlation Matrix are not only used to explain the statistical relationship between variables (James et al., 2021), but also become the basis for the preparation of prescriptive strategies in the credit risk management process in BPR.

However, this study also found some differences compared to previous studies. One of the interesting findings is the negative relationship between the JT Ceiling and the Risk Status ($r = -0.354$). Most international studies consider the size of credit exposure as a factor that has the potential to increase risk if not balanced with adequate risk management. However, in this study, debtors with a larger credit ceiling actually showed a tendency to have lower risk. This condition can be explained by the operational characteristics of BPR, where loans with a large nominal value are generally given to debtors who have gone through a stricter analysis process, have better payment capacity, and are supported by good collateral quality and payment history. Thus, the size of the credit ceiling in the context of BPR reflects more of the results of credit selection than is the cause of increased risk.

Another difference is found in the Debt to Income Ratio (DTI) variable which shows a negative relationship of -0.162 to Risk Status. Theoretically, most studies state that the higher the DTI ratio, the higher the probability of default. Therefore, the results of this study show different characteristics compared to common findings in the literature. These differences may be influenced by the characteristics of the dataset, the relatively homogeneous distribution of DTI values, and the credit selection process that has been carried out before the credit is granted. Thus, these results need to be understood as a specific characteristic of the research dataset and not as a rejection of the theory regarding the relationship between DTI and credit risk.

In addition, the Permanent Employment and Assurance variables show a relatively weak relationship with RiskStatus. These findings indicate that the existence of permanent employment and collateral is not the main differentiating factor between risky and non-risky debtors in the research dataset. This condition is likely due to the characteristics of the BPR portfolio, where most debtors have met the minimum administrative requirements, including the provision of collateral. Therefore, the variation of the two variables becomes relatively small, so their contribution in explaining credit risk is also limited.

Compared to previous research, the novelty of this study does not lie in the discovery that Late History is the dominant factor, because these results have been widely supported in the international literature. The main contribution of this research lies precisely in the integration of Correlation Matrix results into the CRISP-DM-based Business Analytics framework. Unlike most studies that use variable relationship analysis only to explain the performance of the prediction model, this study utilizes the results of the Correlation Matrix as the basis for the preparation of a prescriptive strategy mapped into the 5Cs (Character, Capacity, Capital, Collateral, and Condition of Economy) principles. With this approach, the results of statistical analysis do not stop at academic information, but are translated into operational policy recommendations that can be implemented in the credit granting process at BPR.

Research Findings

Based on a comparison with previous research, it can be concluded that the results of this study are in accordance with the international literature, especially related to the dominance of payment history as the main factor in determining credit risk. The finding that Late History has the strongest relationship with Risk Status supports the results of the study (Muñoz-Cancino et al., 2023) and (Alamsyah et al., 2025), both of which emphasize the importance of behavioral variables over administrative variables in the creditworthiness assessment process.

However, this study also has different characteristics compared to previous research. If the study (Muñoz-Cancino et al., 2023) is oriented towards improving model performance through the use of behavioral information and social networks, while (Alamsyah et al., 2025) focuses on the use of alternative data, this study uses operational data commonly owned by BPRs such as credit ceilings, DTI ratios, guarantees, permanent employment, and payment late history. This shows that improving the quality of credit risk management does not necessarily require complex data sources, but can be achieved through the utilization of systematically analyzed internal data of banks.

In addition, a systematic study conducted by (Ayari et al., 2025) concluded that modern credit scoring studies tend to focus on selecting the best machine learning algorithms and improving prediction performance. This research goes further by integrating the results of the Correlation Matrix into Prescriptive Analytics, so that statistical findings do not stop at academic information, but are translated into credit risk management strategies based on the 5C principle and its implementation roadmap.

Research Novelty

Based on this synthesis, the novelty of this research does not lie in the use of Correlation Matrix or Random Forest algorithms, because both approaches have been widely used in credit scoring research. The main contribution of this research is: Integrating the CRISP-DM methodology with the Business Analytics (Descriptive–Predictive–Prescriptive) framework in the context of BPR credit risk management. Translating the results of the Correlation Matrix into a business strategy based on the 5C principle, so that the results of statistical analysis have direct implications for credit policy. Develop an implementation roadmap that connects analytics results with business processes through the establishment of implementation programs, person in charge (PIC), performance indicators (KPIs), and expected outcomes. This contribution provides added value compared to previous research because it produces a data-driven credit risk management framework that not only explains credit risk factors, but also provides implementation guidelines that can be applied by the People's Economic Bank (BPR).

Research Implications

This research provides practical implications for BPR credit risk management through the application of Business Analytics and the CRISP-DM methodology. The results showed that Late History was the most powerful factor related to Risk Status, while Random Forest was the best model with 99% accuracy. These findings can be applied in strengthening credit appraisal, credit scoring, and an Early Warning System based on payment behavior, while still considering DTI, guarantee, and permanent employment variables. The Random Forest model can also be integrated as a Decision Support System to assist credit analysts without replacing the decisions of officers and Credit Committees. In addition, Correlation Matrix results can be used in the Business Rules Engine and Business Analytics dashboards to support risk monitoring and data-driven decision-making.

For further research, the model can be developed by adding variables such as age, income, business sector, collateral value, restructuring history, and macroeconomic factors. Testing can also be extended using algorithms such as XGBoost, LightGBM, CatBoost, SVM, and Deep Learning (Goodfellow et al., 2016), as well as evaluation indicators such as AUC and F1-Score. The use of XAI methods such as SHAP or LIME can also improve model transparency (Montevechi, 2024). In addition, validation on commercial banks (Sinkey, 2002), digital banks, financing institutions, and fintech lending is needed to test the model's generalization capabilities and develop the integration of Descriptive, Predictive, Prescriptive, and Explainable Analytics in credit risk management.

CONCLUSION

This study concluded that integrating descriptive, predictive, and prescriptive analytics within the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework provided an effective and data-driven approach to credit risk management at PT BPR Jabar Perseroda. Based on 1,015 debtor records, late payment history demonstrated the strongest relationship with risk status ($r = 0.653$), indicating that repayment behavior provided more meaningful risk information than static administrative characteristics, such as collateral, employment status, and credit ceiling. Among the three machine learning algorithms evaluated, the Random Forest algorithm achieved the best classification performance, with 99.00% accuracy, 100.00% precision, and 97.83% recall. Therefore, the model could support credit analysts and credit committees by providing an objective preliminary assessment of debtor risk. Furthermore, translating analytical findings into strategies based on the 5C principles demonstrated that machine learning outputs could be operationalized within credit appraisal processes, Decision Support Systems, and Early Warning Systems while maintaining the role of professional judgment in credit decision-making.

Future research should expand the model by incorporating broader debtor characteristics, including age, income, business sector, collateral value, restructuring history, transaction behavior, and macroeconomic conditions. Comparative evaluations using other machine learning algorithms, such as XGBoost, LightGBM, CatBoost, Support Vector Machines, and deep learning methods, supported by additional performance metrics, including F1-score, Area Under the Curve (AUC), and out-of-sample validation, are also recommended. Explainable Artificial Intelligence (XAI) techniques, particularly SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), should be applied to improve model transparency and support regulatory and managerial interpretation. Moreover, external validation involving other rural banks, commercial banks, digital banks, financing companies, and fintech lending institutions is necessary to assess the model’s generalizability. Longitudinal research should also examine whether integrating the model into actual credit decision-making processes reduces nonperforming loans and improves the quality and consistency of lending decisions over time.

REFERENCES

Alamsyah, A., Hafidh, A. A., & Mulya, A. D. (2025). Innovative credit risk assessment: Leveraging social media data for inclusive credit scoring in Indonesia's fintech sector. *Journal of Risk and Financial Management*, *18*(2), 74. https://doi.org/10.3390/jrfm18020074

Altman, E. I., & Saunders, A. (1998). Credit risk measurement: Developments over the last 20 years. *Journal of Banking & Finance*, *21*(11–12), 1721–1742. https://doi.org/10.1016/S0378-4266(97)00036-8

Ayari, H., Guetari, P. R., & Kraïem, P. N. (2025). Machine learning powered financial credit scoring: a systematic literature review. *Artificial Intelligence Review*, *59*(1). https://doi.org/10.1007/s10462-025-11416-2

Berger, A. N., & Udell, G. F. (1990). Collateral, loan quality and bank risk. *Journal of Monetary Economics*, *25*(1), 21–42. https://doi.org/10.1016/0304-3932(90)90042-3

Bougie, R., & Sekaran, U. (2020). *Research methods for business: A skill-building approach*. John Wiley & Sons, Inc.

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Breiman, L. (2001b). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chang, V., Sivakulasingam, S., Wang, H., Wong, S. T., Ganatra, M. A., & Luo, J. (2024). Credit risk prediction using machine learning and deep learning: A study on credit card customers. *Risks*, *12*(11), 174. https://doi.org/10.3390/risks12110174

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*.

Cochran, W. G. (1977). *Sampling techniques*. Wiley.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed., Ed.). Sage Publications.

Delen, D., & Ram, S. (2018). Research challenges and opportunities in business analytics. *Journal of Business Analytics*, *1*(1), 2–12. https://doi.org/10.1080/2573234X.2018.1507324

Dernsjö, A. (2023). *A gradient boosting tree approach for behavioural credit scoring*.

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Brooks/Cole Pub. Co.

Friedman, J. H. (2001a). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman, J. H. (2001b). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, Inc.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Greuning, H. Van, & Bratanovic, S. B. (2020). *Analyzing banking risk (Fourth Edition): A framework for assessing corporate governance and risk management*. World Bank. https://doi.org/10.1596/978-1-4648-1446-4

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate Data Analysis*. Cengage.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques (3rd Edition)*. Elsevier.

Indonesia. (2023). Undang-Undang Republik Indonesia Nomor 4 Tahun 2023 Tentang Pengembangan Dan Penguatan Sektor Keuangan. In *Negara Republik Indonesia* (Vol. 1, Number 163979).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: with applications in R*. Springer.

Kohavi, R. (1996). *Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid*.

Liu, W., Fan, H., & Xia, M. (2022). Credit scoring based on tree-enhanced gradient boosting decision trees. *Expert Systems with Applications, 189*, 116034. https://doi.org/10.1016/j.eswa.2021.116034

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Montevechi, A. A. (2024). Advancing credit risk modelling with machine learning: A comprehensive review of the state-of-the-art. *Engineering Applications of Artificial Intelligence*, *137*, 109082. https://doi.org/10.1016/j.engappai.2024.109082

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons, Inc.

Muñoz-Cancino, R., Bravo, C., Ríos, S. A., & Graña, M. (2023). On the dynamics of credit history and social interaction features, and their impact on creditworthiness assessment performance. *Expert Systems with Applications*, *218*, 119599. https://doi.org/10.1016/j.eswa.2023.119599

Nallakaruppan, M. K., Chaturvedi, H., Grover, V., Balusamy, B., Jaraut, P., Bahadur, J., Meena, V. P., & Hameed, I. A. (2024). Credit risk assessment and financial decision support using explainable artificial intelligence. *Risks*, *12*(10), 164. https://doi.org/10.3390/risks12100164

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly.

Rose, P. S., & Hudgins, S. C. (2021). *Bank management & financial services (10th Edition)*. McGraw-Hill Education.

Sharda, R., Delen, D., Turban, E., Aronson, J. E., Liang, T.-P., & King, D. (2018). *Business intelligence, analytics, and data science: a managerial perspective*. Pearson.

Sinkey, J. F. (2002). *Commercial bank financial management in the financial-services industry*. Prentice Hall.

Tan, P. H., Thike, A. A., & Tan, W. J. (n.d.). *Phyllodes tumours of the breast: a consensus review*.

Xia, Y., Liu, C., Li, Y., & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, *78*, 225–241. https://doi.org/10.1016/j.eswa.2017.02.017