JOIV : Int. J. Inform. Visualization, 8(1) - March 2024 220-230

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Modeling and Application of Credit Scoring Based on A Multi-Objective Approach to Debtor Data in PT. Bank Riau Kepri

Sugianto [a], Yohana Dewi Lulu Widyasari [a,*], Kartina Diah Kesuma Wardhani [a]

[a] *Applied Master of Computer Engineering, Politeknik Caltex Riau, Rumbai, Pekanbaru, 28265, Indonesia*

Correspondent author: [*]yohana@pcr.ac.id

***Abstract*— The development of information technology in Indonesia, marked by the start of Industry 4.0, is very rapid. With the development of technology, many companies use technology to develop their business, one of which is banking, which analyses the process of prospective customers. New employees find it challenging to interpret and tend to agree more easily with prospective customers because they only see the fulfillment of general requirements. This research aims to find an overview of the primary and additional factors to analyze prospective credit customers using The Cross-Industry Standard Process for Data Mining (CRISP-DM). Develop a model in this study using data variables of prospective customers in health insurance as a moderating variable. This model tested the Decision Tree algorithm with an accuracy value of 92.49%, the Random Forest with an accuracy value of 81.72%, the Support Vector Machine (SVM) with an accuracy value of 91.25%, and K-Nearest Neighbor (K-NN) with an accuracy value. 90.58%, Gradient Boosting with an accuracy value of 90.69%, and XGBoost with an accuracy value of 93.27%. The algorithm uses a cross-validation technique at the validation stage by changing the K value to 2, 4, 6, 8, and 10. The results show that the XGBoost Algorithm accuracy is 93.27% with a K value of 8. As the highest model accuracy, this model was implemented using the XGBoost Algorithm.**

***Keywords*— Supervised learning; credit scoring; algorithm; XGBoost; application.**

Manuscript received 22 Dec. 2022; revised 6 Aug. 2023; accepted 21 Sep. 2023. Date of publication 31 Mar. 2024.

I. INTRODUCTION

MSMEs continue to grow every year, from the micro-segment in 2018 of 251.34 trillion, increasing to 277.23 trillion in 2019, as well as the small segment of 312.07 trillion to 332.12 trillion in 2019. The same happened in the middle part, 469.24 trillion to 488.79 trillion in 2019 [1]. That means the MSME segment in Indonesia has a lot of opportunities in terms of credit. This condition also exists in Riau Province and Riau Islands. The way banks analyze customers can be said to be quite conservative. Every time they apply for credit to the bank, the process requires the debtor's SLIK (Financial Information Service System) as a determinant of the eligibility of prospective debtors. The customer has a credit score calculated from 1 to 5 from his collectability record. The lower the collectability level, the greater the potential for the application to be accepted. If they enter collectability 3, 4, and 5, they will automatically be rejected because they are blocklisted. The problem is the high cost to serve is not worth the potential low income in the medium term. Limited credit history, low savings rate, non-existent amount of debt, long duration of non-existent loans, and little awareness of insurance or investment. The low level of accuracy of traditional analysis. Credit scoring, which was analyzed using the algorithm, showed outstanding results, as evidenced by several previous studies. Various credit scoring approaches have begun to utilize data analysis techniques. Still, no studies are available that provide a systematic outline of the customer loan approach of prospective debtors with a combination of analysis of Health Insurance participants to predict prospective debtors' compliance level.

Regional Development Banks (BPD) have a strategic role in the national banking system. BPD is one of the mainstays in the national banking system in distributing financing to micro, small, and medium enterprises. However, the high cost to serve is not worth the potential low income in the medium term. In addition, there is no payment history, including a limited credit history, the amount of debt owed does not exist, the duration of the loan does not exist, and there needs to be more awareness regarding insurance or investment. This results in a low level of accuracy of traditional analysis. As a result, it is difficult to predict the level of credit risk that will be going through in the future. Ultimately, the risk of default on MSME debtors tends to be higher, the NPL is higher, and

bank income is decreasing. Analyzing prospective customers currently being carried out includes conservative aspects, but decisions are still subjective. It indicates when the team member's position changes in this section. New employees find it challenging to analyze and tend to agree more easily with prospective customers because they only see the fulfillment of general requirements.

II. Material and Methods

The Bank is defined as a business entity that collects (funding) funds from the public through savings. It distributes them (lending) to the people through credit or other conditions to improve the community's quality of life by referring to Law Number 10 of 1998 concerning Banking [2]. The term credit comes from the Greek "Credere," which means trust; therefore, the basis of credit is trust. A person or all entities that provide credit (creditors) believe that future credit recipients (debtors) can fulfill everything that has been promised in the form of goods, money, or services [3]. The data used in the analysis are 39 (Table I) related to existing credit processes.

TABLE I
SELECTED FEATURE ON DATASET

No.	Column	Non-Null Count	type
0	age	1116 non-null	float64
1	Tax number	1116 non-null	float64
2	Job Code	1116 non-null	int64
3	Business fields code	1116 non-null	int64
4	financing credit nature code	1116 non-null	int64
5	Debtor category code	1116 non-null	int64
6	economic sector code	1116 non-null	int64
7	project value	1116 non-null	int64
8	interest rate percentage	1116 non-null	int64
9	type of interest rate	1116 non-null	int64
10	government credit programs	1116 non-null	int64
11	platform	1116 non-null	int64
12	fine	1116 non-null	int64
13	debit balance	1116 non-null	int64
14	crash cause code	1116 non-null	float64
15	principal arrears	1116 non-null	int64
16	interest arrears	1116 non-null	int64
17	arrears frequency	1116 non-null	int64
18	condition code	1116 non-null	int64
19	collectability code 3	1116 non-null	float64
20	collectability code 12	1116 non-null	int64
21	number of days in arrears 21	1116 non-null	float64
22	collectability code 24	1116 non-null	float64
23	number of days in arrears 24	1116 non-null	int64
24	collateral type code	1116 non-null	float64
25	binding type code	1116 non-null	float64
26	Collateral value according to personal tax	1116 non-null	int64
27	membership status	1116 non-null	int64
28	class rights	1116 non-null	int64
29	participant segment	1116 non-null	float64
30	bill 2	1116 non-null	int64
31	status 3	1116 non-null	int64
32	status 5	1116 non-null	int64
33	payment 7	1116 non-null	int64
34	bill 8	1116 non-null	int64
35	status 9	1116 non-null	int64
36	bill 11	1116 non-null	int64
37	status 12	1116 non-null	int64
38	arrears	1116 non-null	int64
39	clusters	1116 non-null	int64

Credit scoring is a way to find a different group when someone can't see a location visit that separates the groups but only those related. Fisher introduced the idea of differentiating groups within a population in statistics [4], [5]. Machine learning is the knowledge of computer science that provides learning capabilities for computers to know something without explicit programming [6], [7].

A. Machine Learning

Machine Learning is the knowledge of computer science that provides learning capabilities for computers to know something without explicit programming [8]. Machine Learning is a subset of artificial intelligence. This discipline includes designing and developing algorithms that allow computers to develop behaviors based on historical data [9], such as from databases of sensor data [10]. In machine learning, there are things such as:

1) Supervised Learning. Learning uses learning data input that is labeled. After that, make predictions from its dataset.

2) Unsupervised Learning: Learning uses learning data input that is not labeled. After that, try to group the data based on the characteristics encountered.

3) Reinforcement learning: The learning and test phases are mixed. To actively collect learner information by interacting with the environment to get a reply for every action from the learner [11], [12].

Machine learning is a technique for classifying and predicting new data by applying models to learn patterns from previous data sets [13]. Machine learning is also deployed because it can process data more effectively and efficiently [14]. According to the research, the decision tree model is also used to identify relevant factors for the prediction process [15], [16].

B. Decision Tree

In a study entitled Comparison of Supervised Learning Classification Methods on Data Bank Customers Using Python Decision trees, also known as top-down induction decision trees (TIDIDT), are supervised learning techniques that construct representations of classification rules in a hierarchical sequential structure by recursively partitioning training data sets [17].

C. Random Forest

In a study entitled Chemosphere Hybrid Decision Tree-based machine learning model for short-term water quality prediction, Chemosphere results that Random Forest is an integrated learning method for classification and regression. It is representative learning and is a model based on a bagging algorithm, Random Forest, when constructing each tree, using

a random sample of predictors before each node segmentation, which can reduce the unsegmented ones [18], [19].

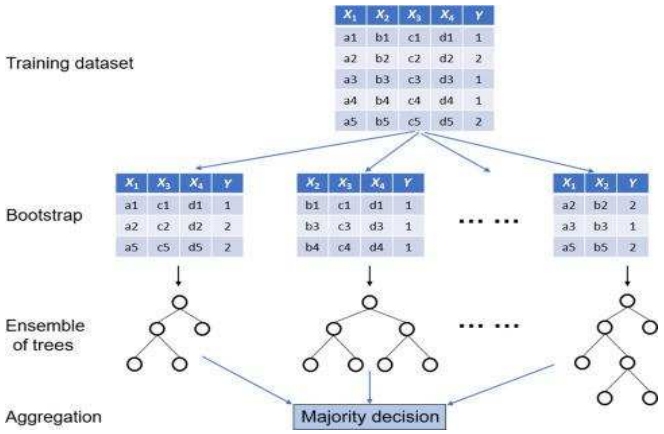

Fig. 1 Random Forest classifier

Fig.1 describes an example of implementing the Random Forest algorithm classification on a dataset that has four features (X1, X2, X3, and X4) and two classes (Y 1 and 2) [19].

D. Gradient Boosting Algorithm

Gradient Boosting Machine is a decision tree algorithm that makes predictions using a different format. Unlike ordinary decision tree algorithms, which parallelize the tree construction process, the Gradient Boosting Machine takes a sequential approach to get predictive results. The Gradient Boosting algorithm develops a regression model sequentially by adjusting the previous classification to the following classification; thus, the new classification will be trained to correct the misjudgment of the last classification, adaptively improving the prediction performance with high efficiency [20], [21].

E. Extreme GBoost (XGBoost) Algorithm

The eXtreme Gradient Boosting (XGBoost) algorithm is an alternative to the Gradient Boosting algorithm. (Gradient Boosting Machine-GBM) It can be optimized to prevent redundant learning and manage blank data. It is fast and has high predictive power compared to the known machine learning methods and algorithms Chen and Guestri first developed. XgBoost algorithm works ten times faster. XGBoost has good predictive power and computational ease compared to popular machine learning methods. Multidimensional data analysis uses these calculations quickly [22].

F. Support Vector Machine (SVM) Algorithm

The Support Vector Machine (SVM) Algorithm is a machine learning system using a theory in the form of a linear function in a feature based on optimal theory [23]. Boser, Guyon, and Vapnik developed SVM, first published in 1992 at the Annual Workshop on Computational Learning Theory. The basic theory of SVM is obtained from mixing pre-existing computational theory [16].

G. K-Nearest Neighbor (K-NN) Algorithm

The K-Nearest Neighbor algorithm is the standard prediction algorithm [23]. The k-NN algorithm is one of the classification algorithms in data mining that calculates the similarity of the object with the nearest (k) neighbor [24]. The steps to calculate the K-NN algorithm are as follows:

- Determine the value of k, which is the number of neighbors that have similarities with the data labeled.
- Calculate the square of the Euclidean distance of each object to the given training data using the equation.
- Then, sort the things into groups with the smallest Euclidean distance.
- Collecting class Y labels (K-Nearest Neighbor Classification).
- The calculated query instance value can be predicted using the K-Nearest Neighbor category, which is the majority [24].

H. Analysis Technique

This study uses the clustering technique to analyze the data to determine which risk level is high or low [25]. Several basic methods, including Fuzzy Centroid (PC) and Fuzzy k-mean Partition (FkP), were compared with the proposed multivariate multinomial distribution technique based on several soft sets [26].

I. Cross Industry Standard Process for Data Mining (CRISP-DM)

Cross-Industry Standard Process for Data Mining or CRISP-DM (Fig. 2) is one of the process model tools (data mining framework). Many organizations and European companies develop these frameworks as a non-proprietary standard methodology for data mining. An analysis of data comes from something other than a momentary process but from a mature approach that uses high standards. CRISP DM [27], or in full called the Industry Standard Process for Data Mining is a standard developed in 1996 in Europe. CRISP DM consists of 6 stages, of which the three earliest stages, based on the author's experience, can be Non-Mutually Exclusive. The six stages of CRISPR-DM are presented in the scheme below: [28], [29].

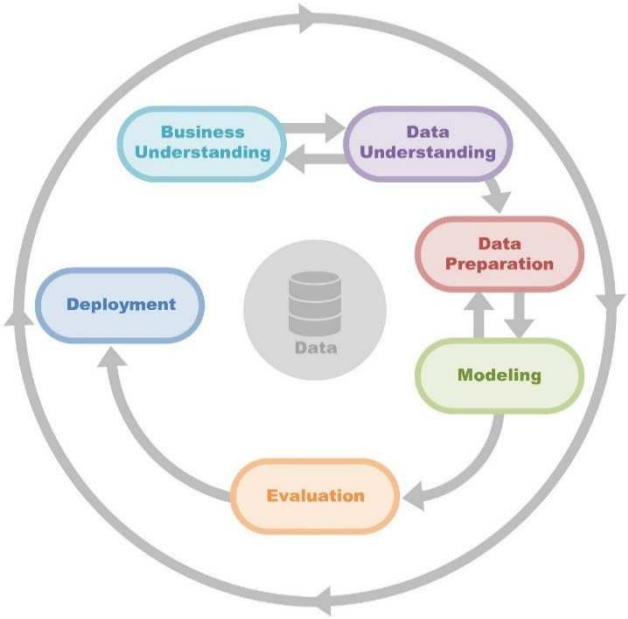

Fig. 2 CRISP-DM model

J. Multi-Objective Design

Multi-Objective Optimization uses the weighted sum method. The way this method works is by combining all objective functions into one scalar objective function and giving weight to each objective function [30]. The weighted-sum method works by assigning weight to each objective function. As an illustration, the optimization model of the multi-objective assignment problem using this method is as follows:

$$plus\ ,w\ sub\ 2\ ,sum\ from\ I\ equals\ 1\ to\ m\ of\ the\ Z = w_1 \sum_{i=1}^{m}\sum_{j=1}^{n} c_{1ij}.x_{ij} + w_2 \sum_{i=1}^{m}\sum_{j=1}^{n} c_{2ij}.x_{ij} + \ldots + w_k \sum_{i=1}^{m}\sum_{j=1}^{n} c_{kij}.x_{ij} \quad (1)$$

This section will solve the multi-objective assignment model using the constraint method. This method of solving multi-objective assignment problems should minimize one of all tasks from the objective function. While other goals serve as a limiter that is less than or equal to the given target value [31].

$$\begin{pmatrix} \varepsilon_1 \\ \ldots \\ \varepsilon_{q-1} \\ \varepsilon_{q+1} \\ \varepsilon_k \end{pmatrix} \quad (2)$$

$$Z_q = \sum_{i=1}^{m} \sum_{j=1}^{n} c_{qij}.x_{ij} \quad (3)$$

III. Result and Discussion

System design addresses everything that is done to achieve a goal (Fig.3).

A. Current Credit Application Business Process

Prospective debtors/debtors apply for Credit through the Debtor Management Unit (Subbranch / Branch / Head Office).

- The Relationship Officer/Manager (RO/RM)/Credit Analyst (CA) is considered for further processing. The next process is collecting the necessary data (related to prospective debtors/debtors or third parties) and verifying data.

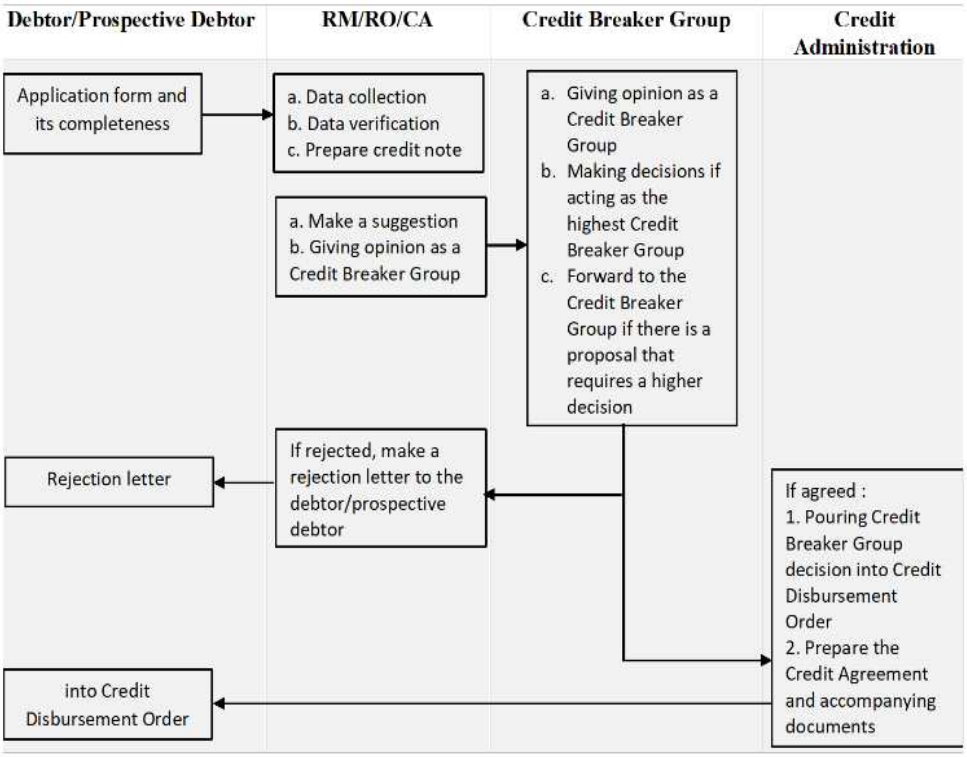

Fig. 3 Credit processing flow

- Based on the data that has been collected and verified, RO/RM/CA conducts qualitative and quantitative analysis as outlined in the Credit Analysis Toolkit.
- Propose Credit to the competent Credit Breaker Group using the Credit Proposal Form (CPF).
- Each member of the Credit Breaker Group provides an opinion in the available Credit Analysis Notification column.
- The highest Credit Breaker Group member decides in the Credit Analysis Notice.
- The opinion and/or decision of each member of the Corruption Eradication Commission must be written clearly and unequivocally regarding the decision to agree or disagree with the granting of Credit and not to repeat the opinion of the previous Credit Breaker Group member so as not to cause a difference of perception for officers in carrying out the said decision.
- Credit Breaker Group decision. It will be submitted to the debtor / prospective debtor in a rejection letter if rejected. However, if approved, poured, and placed into Credit Disbursement Order and Credit Agreement by the Credit Administration Unit.

B. Future Business Design

From Fig. 4 above, the future business process begins when a prospective debtor submits a credit application. Prospective debtors apply for Credit by filling out the application form; the output is the application document. Next, the Account Officer (AO) checks the application, and if it is acceptable, the AO proceeds to carry out a creditworthiness analysis; if not, the AO issues a credit rejection letter.

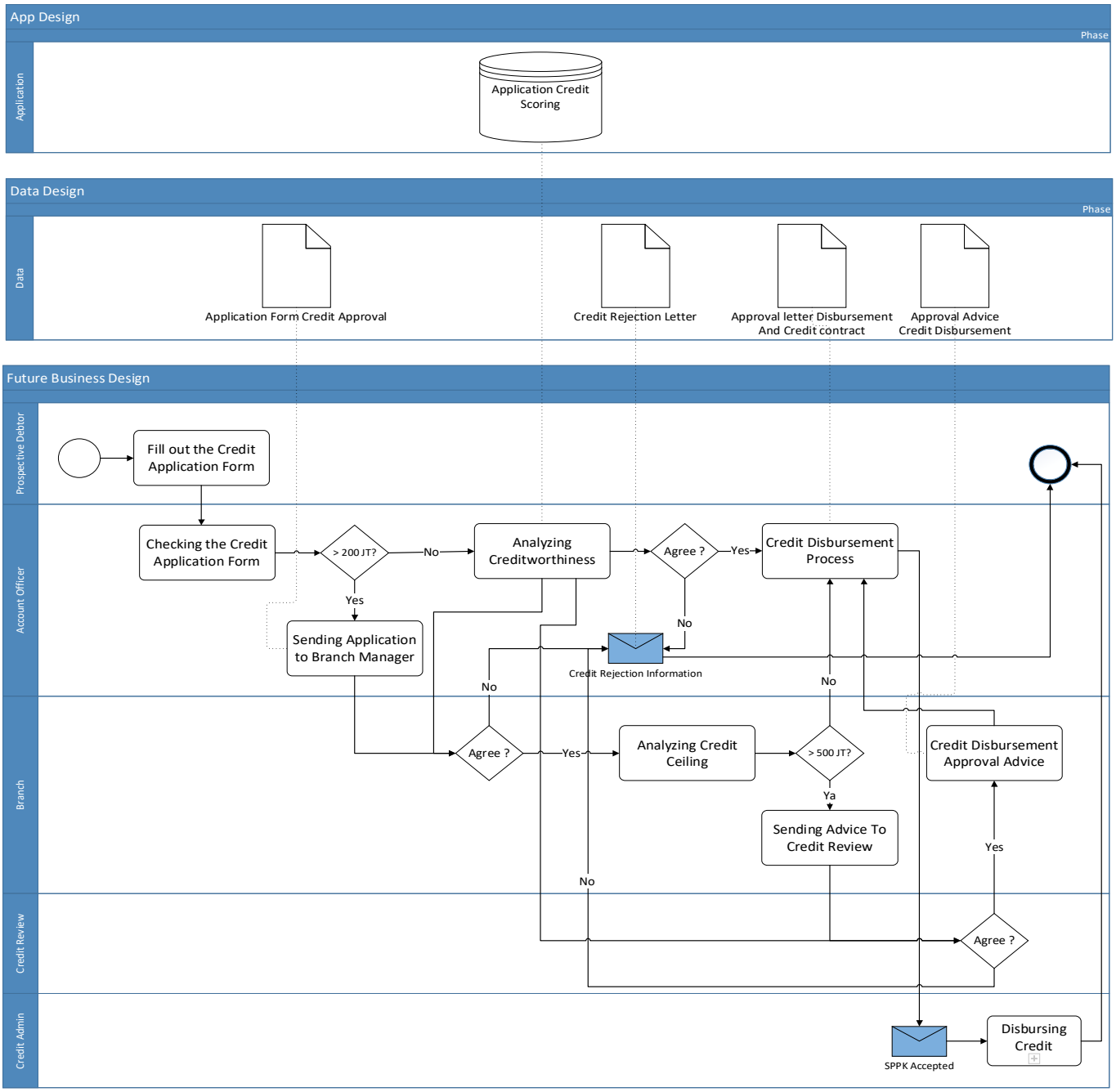

Fig. 4 Future business design

After the AO conducts a creditworthiness analysis, the AO recommends whether the application is to the head of the office. The leadership decides whether this application is accepted; if accepted, then the application is forwarded to the credit administration section for credit disbursement. The output of this process is two documents, namely a letter of approval for credit disbursement and credit contracts. Figure 4 above shows several stages of the credit disbursement process. First, for the status of the sub-office in implementing credit disbursement with a ceiling above 200 million, the head of the office must obtain approval from the head of the consolidation branch.

The output of this stage is a letter of approval from the branch leadership for disbursement approval. Meanwhile, the credit limit for branch managers is up to 500 million. The decision remains with the Branch Manager. If the credit limit exceeds 500 million, then Credit Review provides advice regarding credit disbursement approval. The Figure 4 above shows several stages of the credit disbursement process. First, for the status of the Sub-Branch office in implementing credit

disbursement with a ceiling above or equal to 500 million, the head of the office must obtain approval from the head of the consolidation branch. The output of this stage is a letter of support from the branch leadership for disbursement approval. Meanwhile, the credit limit for branch managers is up to 500 million; the decision remains with the Branch Manager. For above 500 million, Credit Review provides advice regarding credit disbursement approval.

C. System Architecture Design

Credit Review can access the Web Portal anywhere and anytime with the help of the framework. This website will implant a machine-learning model of the Random Forest algorithm, which is helpful for automatically classifying the collectability level and calculating scores based on data obtained through calls from the database. The system workflow begins with a prospective debtor who accesses the web portal, and then the data is sent via a web server connected to the database. The web portal will display the data, and the data will automatically perform automatic classification to get the scoring level.

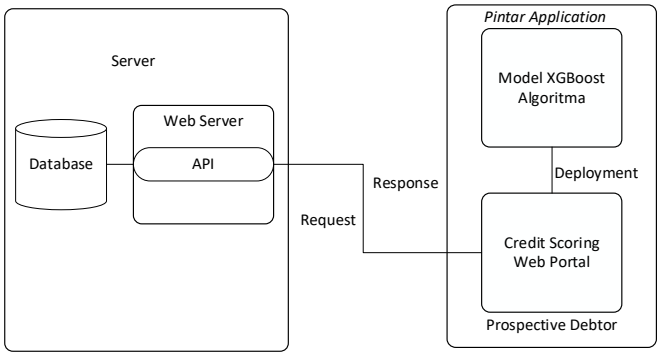

Fig. 5 Architecture design

D. Credit Score Design

The first step in this part is collecting data on the evaluated individual or entity. This data should correspond to the parameters listed in the table. Each parameter is assigned a point value based on the individual's or entity's characteristics. For example, if an individual is 30 to 50 years old, they would receive 5 points for the "Applicant's Age" criterion. The points for each parameter are added together to calculate a total score for the individual or entity. The total score is then used to make decisions. For example, a higher total score in a credit scoring system might indicate a lower risk, which could lead to the individual or entity being more likely to be approved for credit.

TABLE II
CREDIT SCORE DESIGN

Data Source	Scorecard Criteria	Parameter	Points
Internal	Applicant's Age	30 yo up to 50 yo	5
		50 yo up to 58 yo	3
		21 yo up to 30 yo	2
		or > 58 yo up to 63 yo	1
	Credit Installment Track Record	Never Delinquent	5
		Never got credit	3
		Currently has a loan with current collectability but has been in arrears with a maximum collectability of 3 (substandard) which occurred > 1 year	2
	The number of dependents	Up to 2 people	5
		2 up to 4 people	3
		5 up to 6 people	2
		6 people	1
	Business Key Person	The business is managed independently and is not dependent	5
		Self-managed but dependent on the expertise of other parties (wife/husband, siblings)	3
		Self-managed but dependent on employee expertise	1
	How long the business has been running	4 years	5
		2 up to 4 years	3
		1 up to 2 years	2
		6 mo. up to 1 years	1
	Sales system	All goods/services sold/produced must be sold	5
		There is no certainty that all goods/services sold/produced will be sold	3
	Business Location	The business location is near the market/crowd/economic center with a maximum distance of 1 km	5
		The business location is near the market/crowd/economic center with a distance of > 1 km to 2 km	3
		The business location is near the market/crowd/economic center with a distance of > 2 km	2
	Types of products	Apart from plants and animals	5
		Plant and animal commodities	3
		Saturated product	0
	Ownership of business premises	One's own	5
		Family owned	3
		Rent/contract/retribution payments	2
		Apart from the criteria above	1
		Illegal/can be evicted	0
	Merchandise inventory (availability of raw materials)	Can be obtained easily at the same place or from other places in < 1 day	5
		Can be obtained easily at the same place or from other places within > 1 day to 3 days	3
		Can be obtained easily at the same place or from other places within > 3 days	2
	Debtor's Domicile	Own house	5
		Family-owned house	3
		Rent/contract and have parents/brothers/sisters domiciled in their own house in the same city/district and willing to	2

Data Source	Scorecard Criteria	Parameter	Points
		communicate with the debtor/provide information	
		Other	0
	Length of domicile	Have been domiciled for > 5 years as proven by KTP/KK	5
		Have been domiciled for > 3 to 5 years, proven by KTP/KK	3
		Have been domiciled for <3 years, proven by KTP/KK	0
	Membership status	Active	5
		Inactive at the end of the month	2
		Inactive From Center	1
		Inactive Due to Premium	0
	Treatment Class Rights	Class 1	5
		Grade 2	3
		Grade 3	1
External	Participant Segment	Independent Worker	5
		Private employees	1
		BUMN employee	1
		BUMD employees	1
		PBI Participants (Contribution Assistance Recipients)	1
		civil servants	1
		Retired	1
		Village Head/Apparatus	1
		Participants in the Agreement with an employment contract	1
	Billing History	Paid off	5
		Reception	2
		Not yet paid	0

E. Data Understanding

This research uses a Multi-Objective Approach in determining the dataset because the dataset used comes from more than one and has many attributes. The data used in this research is debtor data consisting of 1772 data, with 237 features. Debtor data is data on customers who have loans from banks. And the other dataset is PT.XYZ insurance participant data. Furthermore, using the Multi-Objective Approach, bank debtor data were determined, which at the same time were also participants in PT. XYZ insurance. To obtain a dataset totaling 1116 rows and 126 features.

F. Feature Selection

Feature Selection in this research was done using Featurewiz, which finds essential variables from a dataset concerning the target variable [32-35]. Visualization of Featurewiz in Figure 6.

Fig. 6 Feature selection using Featurewiz

G. Algorithm Testing

1) Decision Tree Algorithm: With an accuracy of 90.178%, the degree of proximity between the measured quantity and the actual value is denoted.

```
Decision Tree Model Accuracy :  90.17857142857143 %

Confusion matrix :
[[  3   0   0  10   4]
 [  0 161   0   0   0]
 [  0   0  26   0   0]
 [  2   0   0   9   3]
 [  0   0   0   3   3]]

Classification report:
              precision    recall  f1-score   support

           0       0.60      0.18      0.27        17
           1       1.00      1.00      1.00       161
           2       1.00      1.00      1.00        26
           3       0.41      0.64      0.50        14
           4       0.30      0.50      0.37         6

    accuracy                           0.90       224
   macro avg       0.66      0.66      0.63       224
weighted avg       0.91      0.90      0.90       224
```

Fig. 7 Test result decision tree algorithm

2) Random Forest Algorithm: With an accuracy of 90.93%, the level of measurement for the quantity is near the actual value.

```
Random Forest Model Accuracy :  90.93 %

Confusion matrix :
[[  3   5  14   0]
 [  0 261   0   0]
 [  1   6  24   2]
 [  1   3   0  33]]

Classification report:
              precision    recall  f1-score   support

           0       0.60      0.14      0.22        22
           1       0.95      1.00      0.97       261
           2       0.63      0.73      0.68        33
           4       0.94      0.89      0.92        37

accuracy                           0.91       353
macro avg          0.78      0.69      0.70       353
weighted avg       0.90      0.91      0.89       353
```

Fig. 8 Test result random forest algorithm

3) Support Vector Machine (SVM) Algorithm: The accuracy of the quantity measurement is 89.29%, indicating a high level of agreement with the actual value.

```
Support Vector Classifier Accuracy :  89.29 %

Confusion matrix :
[[  0   0   0  17   0]
 [  0 161   0   0   0]
 [  0   0  25   1   0]
 [  0   0   0  14   0]
 [  0   0   0   6   0]]

Classification report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        17
           1       1.00      1.00      1.00       161
           2       1.00      0.96      0.98        26
           3       0.37      1.00      0.54        14
           4       0.00      0.00      0.00         6

accuracy                           0.89       224
macro avg          0.47      0.59      0.50       224
weighted avg       0.86      0.89      0.87       224
```

Fig. 9 Test result support vector machine algorithm

4) k-Nearest Neighbor (k-NN) Algorithm: The accuracy of the quantity measurement is 87.95%, indicating a high level of agreement with the real value.

```
KNN Model Accuracy :  87.95 %

Confusion matrix :
[[  1   0   0  12   4]
 [  0 161   0   0   0]
 [  1   0  25   0   0]
 [  3   0   0   9   2]
 [  0   0   0   5   1]]

Classification report:
              precision    recall  f1-score   support

           0       0.20      0.06      0.09        17
           1       1.00      1.00      1.00       161
           2       1.00      0.96      0.98        26
           3       0.35      0.64      0.45        14
           4       0.14      0.17      0.15         6

accuracy                           0.88       224
macro avg          0.54      0.57      0.54       224
weighted avg       0.88      0.88      0.87       224
```

Fig. 10 Test result KNN algorithm

5) Gradient Boosting Algorithm: The accuracy of the measurement is 89.73%, indicating a high level of precision and proximity to the true value.

```
Gradient Boosting Model Accuracy :  89.73%

Confusion matrix :
[[  2   0   0  11   4]
 [  0 161   0   0   0]
 [  0   0  26   0   0]
 [  2   0   0  11   1]
 [  0   0   0   5   1]]

Classification report:
              precision    recall  f1-score   support

           0       0.50      0.12      0.19        17
           1       1.00      1.00      1.00       161
           2       1.00      1.00      1.00        26
           3       0.41      0.79      0.54        14
           4       0.17      0.17      0.17         6

accuracy                           0.90       224
macro avg          0.61      0.61      0.58       224
weighted avg       0.90      0.90      0.89       224
```

Fig. 11 Test result gradient boosting algorithm.

6) XGBoost Algorithm: The accuracy is 89.73%, which means that the quantity measurement is quite close to the actual value.

```
XGBoost Model Accuracy :  89.73 %

Confusion matrix :
[[  2   0   0  11   4]
 [  0 161   0   0   0]
 [  0   0  26   0   0]
 [  2   0   0  11   1]
 [  0   0   0   5   1]]

Classification report:
              precision    recall  f1-score   support

           0       0.50      0.12      0.19        17
           1       1.00      1.00      1.00       161
           2       1.00      1.00      1.00        26
           3       0.41      0.79      0.54        14
           4       0.17      0.17      0.17         6

accuracy                           0.90       224
macro avg          0.61      0.61      0.58       224
weighted avg       0.90      0.90      0.89       224
```

Fig. 12 Test result XGBoost algorithm

H. Modelling

Data modeling produces a descriptive diagram of the relationship between various types of information to be stored in the database. One of the data modeling goals is to create the most efficient method of storing information while providing complete access and reporting. The modeling process for predictive models uses training data. Training data is a collection of data where the label results are known. The training data acts as a measure of whether the model needs further adjustment or not. Furthermore, the model is developed for the following algorithm using the XGBoost algorithm.

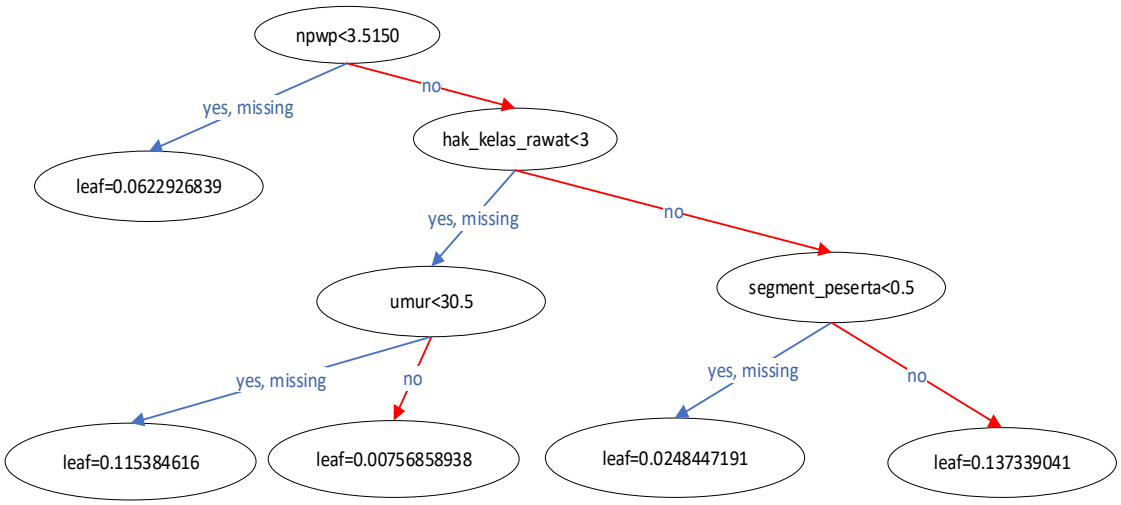

Fig. 13 Modelling XGBoost

I. Evaluation

At this stage, evaluating the quality and effectiveness of one or more models submitted in the modeling phase will be done before placing them for use. The following process is data evaluation. In the previous process, a model was formed to perform classification with the XGBoost algorithm. So, to make sure the model has worked well, a model evaluation can be carried out.

The analysis of the results of the accuracy algorithm aims to obtain a tree model with the best accuracy results based on initial data, namely data from credit debtors who apply for credit and data from insurance participants of PT.XYZ.
The following table shows the results of the accuracy of the algorithm tested, namely the cognate tree algorithm:

TABLE III
ANALYSIS OF TEST RESULTS

No	Algorithm Name	Accuracy
1	Decision Tree Algorithm	90.178%
2	Random Forest Algorithm	83.035%
3	Support Vector Machine (SVM) Algorithm	89.285%
4	k-Nearest Neighbor (k-NN) Algorithm	87.946%
5	Gradient Boosting Algorithm	89.732%
6	XGBoost Algorithm	89.732%

Table 2 shows the highest accuracy in the Decision Tree Algorithm with an accuracy value of 90.178%. The next stage is testing by validating using K-Fold on the algorithm above. For this stage, 80% was training data, and 20% was testing data, resulting in Table 4.

TABLE IV
VALIDATING USING K-FOLD

No	Algorithm name	Accuracy					
		Kfold (2)	**Kfold (4)**	**Kfold (6)**	**Kfold (8)**	**Kfold (10)**	**Average**
1	Algorithm Decision Tree	90.69%	90.58%	91.48%	91.48%	92.49%	91.34%
2	Algorithm Random Forest	77.35%	79.93%	81.28%	81.72%	81.50%	80,35%
3	Algorithm Support Vector Machine (SVM)	91.25%	91.03%	91.14%	91.03%	91.14%	91.11%
4	Algorithm k-Nearest Neighbor (k-NN)	90.58%	90.02%	90.02%	89.45%	89.57%	89.92%
5	Algorithm Gradient Boosting	90.69%	92.26%	92.82%	92.37%	92.60%	92.14%
6	Algorithm XGBoost	89.91%	92.37%	92.94%	93.27%	93.16%	92.33%

From Table 3, this study compares the test results of each algorithm carried out in the table above. The first column is the decision tree algorithm with validation tests carried out by trying to determine the K-Fold value of 2, 4, 6, 8, and 10. This study shows that the highest accuracy value is on K-Fold 10, with an accuracy value of 92.49%. In the second column, the same validation was carried out with a different algorithm, namely the Random Forest algorithm, with the highest value on K-Fold 8 with an accuracy of 81.72%. In the third column, validation is carried out with the Support Vector Machine (SVM) Algorithm. The highest value is on K-Fold 2, with an accuracy of 91.25%.

In the fourth column, validation is carried out with the k-nearest Neighbor (k-NN) algorithm. The highest value on the K-Fold is equal to 2, with an accuracy of 90.58%. In the fifth column, validation is conducted using the Gradient Boosting Algorithm with the result that the highest value on the K-Fold is equal to 2, and the accuracy value is 90.69%. Finally, validation was carried out with the XGBoost Algorithm, with the highest result on K-Fold equal to 8 and an accuracy value of 93.27%. These results concluded that the best accuracy results on K-Fold are similar to 8 using the XGBoost Algorithm, where the experiment is k times for one model with the same parameters. K-Fold Cross Validation aims to obtain maximum accuracy results.

J. Deployment

The following process is deployment. Where to use this machine learning model in the production stage, changes are needed so that it can make predictions with other tools such as websites and mobile. In the XGBoost model itself, it converted into json form.

```
[434]  1 XGB_model.save_model("model_XGboost.json")
```

Fig. 14 Deployment

An analysis of data comes from something other than a momentary process but from a mature approach using high standards. The first stage of CRISP-DM is Business Understanding, where the Business problem is precisely defined. Data Understanding is a process where we reconcile what data we have and what data we should need. Data preparation is a data treatment process that leads to a useful quality model. This stage is the most draining of resources for the analysis team. A model is a quality description or knowledge built by a system or process from acceptable calculations and predictions. The adjective proper here refers to at least several things, namely technically correct and economically correct. Evaluation is the validation stage of the model that is formed based on the relevant parameters. Finally, the Deployment Phase, where analysts and engineers pack and deliver the data analysis process. Some considerations are visualization, ease of use of the model, maintenance of the model in the future, and the legal umbrella that accompanies the use of the model. Why it is legally permissible is a factor considered because modeling will lead to decision-making full of regulatory signs.

Figure 15 is the interface of the Smart Syariah Bank XYZ web application, which consists of information related to products, credit schemes, contacts who can be contacted, and office addresses. The prediction model from the analysis stage is embedded in this application.

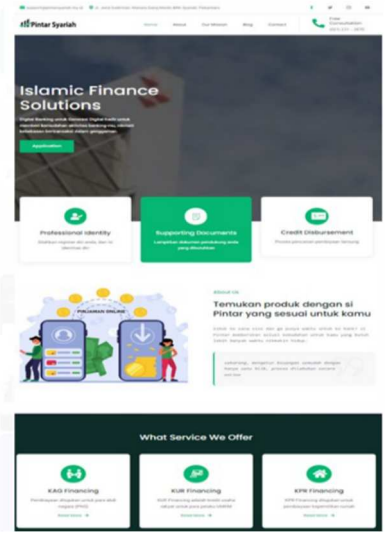

Fig. 15 Pintar Syariah Application

Users who are prospective customers of PT. Bank Riau Kepri can access the application button to submit a loan application to PT. Riau Islands Bank. The following is the dashboard page for logging in to the PT. Bank Riau Kepri in Fig. 16.

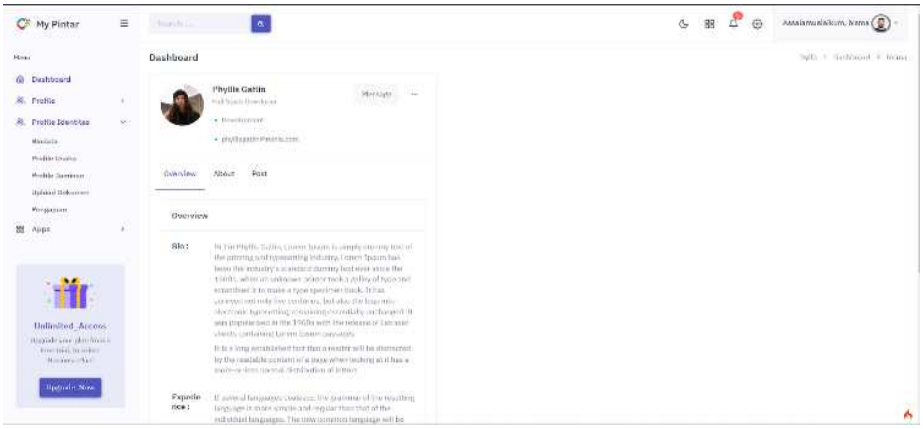

Fig. 16 User of Pintar Syariah Application

IV. CONCLUSION

After testing the model, this study finds several conclusions. There are 39 best features out of 126 features in the dataset using the Feature Selection technique implemented in this research. The most accurate model using the 39 best features is the Decision Tree Algorithm with 90.178%. The K-Fold validation shows that the XGBoost algorithm has the best accuracy, with 93.27% on a K-Fold equal to 8. Development of a web-based application in this research by implementing the XGBoost Algorithm capable of predicting the Risk Grading Matrix using the XGBoost algorithm with prediction results type A+, A, B, C+, C-.

ACKNOWLEDGMENT

We thank PT. Bank Riau Kepri for permits to use some of the data and DIKSI funded for this project.

REFERENCES

[1] Badan Pusat Statistik, "Proporsi Kredit UMKM Terhadap Total Kredit (Triliun Rupiah), 2017-2019," pp. 2–3, 2020.

[2] O. J. KEUANGAN and REPUBLIK INDONESIA, "Peraturan Otoritas jasa keuangan republik indonesia No. 42 /POJK.03/2019," vol. 42 /POJK.0, 2019.

[3] V. K. J. Pongilatan et al., "Evaluation Of The Suitability Of The Allowance For Impairment Losses On Credit With Sfas 55 At Sulutgo Bank Branch Ratahan Oleh: Jurusan Akuntansi , Fakultas Ekonomi dan Bisnis E-mail : Keywords : SFAS 55 , recognition and measurement , allowance for impa," vol. 9, no. 55, pp. 625–632, 2021.

[4] L. C. Thomas, "A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers," International Journal of Forecasting, vol. 16, no. 2, pp. 149–172, Apr. 2000, doi:10.1016/s0169-2070(00)00034-0.

[5] Y. Religia, A. Nugroho, and W. Hadikristanto, "Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 5, no. 1, pp. 187–192, 2021

[6] E. Dumitrescu, S. Hue, C. Hurlin, and S. Tokpavi, "Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects," *Eur. J. Oper. Res.*, vol. 297, no. 3, pp. 1178–1192, 2022.

[7] Peng Du and Hong Shu, "Exploration of financial market credit scoring and risk management and prediction using deep learning and bionic algorithm, " Journal of Global Information Management (JGIM), 30(9), 1-29, 2022

[8] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," Journal of Applied Science and Technology Trends, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi:10.38094/jastt20165.

[9] H. K. Yaseen and A. M. Obaid, "Big Data: Definition, Architecture & Applications," JOIV : International Journal on Informatics Visualization, vol. 4, no. 1, pp. 45–51, Feb. 2020, doi: 10.30630/joiv.4.1.292.

[10] M. Aljanabi et al., "Large Dataset Classification Using Parallel Processing Concept," pp. 1–4, 2020.

[11] H. Park and J. Jeon, "Optimal Data Transmission and Improve Efficiency through Machine Learning in Wireless Sensor Networks," JOIV : International Journal on Informatics Visualization, vol. 6, no. 2–2, p. 463, Aug. 2022, doi: 10.30630/joiv.6.2-2.1125.

[12] D. Lee, J.-Y. Hwang, Y. Lee, and S.-W. Kim, "Informatics and Artificial Intelligence (AI) Education in Korea: Situation Analysis Using the Darmstadt Model," JOIV : International Journal on Informatics Visualization, vol. 6, no. 2, p. 427, Jun. 2022, doi:10.30630/joiv.6.2.1000.

[13] T. Wellem, Y. Nataliani, and A. Iriani, "Academic Document Authentication using Elliptic Curve Digital Signature Algorithm and QR Code," JOIV : International Journal on Informatics Visualization, vol. 6, no. 3, p. 667, Sep. 2022, doi: 10.30630/joiv.6.2.872.

[14] S. Irawan and R. Firsandaya Malik, "Credit Scoring Menggunakan Algoritma Classification And Regression Tree (CART)," vol. 2, no. 1, pp. 82–85, 2017

[15] F. Irawan and F. Samopa, "A Comparative Assessment of the Random Forest and SVM Algorithms Using Combination of Principal Component Analysis and SMOTE For Accounts Receivable Seamless Prediction case study company X in Surabaya," 2018.

[16] F. Sodik, B. Dwi, and I. Kharisudin, "Perbandingan Metode Klasifikasi Supervised Learning pada Data Bank Customers Menggunakan Python," Jurnal Matematika, vol. 3, pp. 689–694, 2020.

[17] H. Lu and X. Ma, "Hybrid decision tree-based machine learning models for short-term water quality prediction," Chemosphere, vol. 249, p. 126169, Jun. 2020, doi: 10.1016/j.chemosphere.2020.126169.

[18] S. Misra and H. Li, "Noninvasive fracture characterization based on the classification of sonic wave travel times," Machine Learning for Subsurface Characterization, pp. 243–287, 2020, doi: 10.1016/b978-0-12-817736-5.00009-0.

[19] P. Palimkar, R. N. Shaw, and A. Ghosh, "Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach," Lecture Notes in Networks and Systems, pp. 219–244, Jul. 2021, doi:10.1007/978-981-16-2164-2_19.

[20] Y. Christian, "Predicting Consumer Interest in All You Can Eat Restaurants with Gradient Boosting Algorithm," Journal of Informatics and Telecommunication Engineering, vol. 6, no. 1, pp. 91–100, Jul. 2022, doi: 10.31289/jite.v6i1.7209.

[21] M. Fatih Yuruk, "Xgboost (Extreme Gradient Boosting) Tabanli Algoritma Ile Gümüş Fiyatlarinin Tahmin Edilmesi Some of the authors of this publication are also working on these related projects: Prediction of Silver Prices With Xgboost (Extreme Gradient Boosting) Based Algorithm View project," 2022. [Online]. Available: https://www.ispecongress.org/sosyal-bilimler

[22] A. Deharja, M. W. Santi, M. Yunus, and E. Rachmawati, "Sistem Prototype Klasifikasi Risiko Kehamilan Dengan Algoritma k-Nearest Neighbor (k-NN)," JTIM : Jurnal Teknologi Informasi dan Multimedia, vol. 4, no. 1, pp. 66–72, May 2022, doi:10.35746/jtim.v4i1.229.

[23] S. Styawati, A. Nurkholis, A. A. Aldino, S. Samsugi, E. Suryati, and R. P. Cahyono, "Sentiment Analysis on Online Transportation Reviews Using Word2Vec Text Embedding Model Feature Extraction and Support Vector Machine (SVM) Algorithm," 2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE), Jan. 2022, doi: 10.1109/ismode53584.2022.9742906.

[24] R. Saedudin, I. T. Riyadi Yanto, A. Budiono, S. Novita Sari, M. Mat Deris, and N. Senan, "Data Clustering for Identification of Building Conditions Using Hybrid Multivariate Multinominal Distribution Soft Set (MMDS) Method," JOIV : International Journal on Informatics Visualization, vol. 6, no. 2, p. 284, Jun. 2022, doi:10.30630/joiv.6.2.986.

[25] - Sarmini, A. Alhabeeb, M. M. Abusharhah, T. Hariguna, and A. R. Hananto, "An Investigation into Indonesian Students' Opinions on Educational Reforms through the Use of Machine Learning and Sentiment Analysis," JOIV : International Journal on Informatics Visualization, vol. 6, no. 3, p. 604, Sep. 2022, doi:10.30630/joiv.6.3.894.

[26] A. N. Iffah'da and A. Desiani, "Implementasi Algoritma K-Nearest Neighbor (K-NN) dan Single Layer Perceptron (SLP) Dalam Prediksi Penyakit Sirosis Biliari Primer," J. Ilm. Inform., vol. 7, no. 1, pp. 65–74, 2022.

[27] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," Procedia Comput. Sci., vol. 181, pp. 526–534, 2021.

[28] E. Kristoffersen, O. O. Aremu, F. Blomsma, P. Mikalef, and J. Li, "Exploring the relationship between data science and circular economy: an enhanced CRISP-DM process model," in Conference on e-Business, e-Services and e-Society, 2019, pp. 177–189.

[29] V. Singh, A. Singh, and K. Joshi, "Fair CRISP-DM: Embedding Fairness in Machine Learning (ML) Development Life Cycle.," in HICSS, 2022, pp. 1–10

[30] A. Pradhan and M. P. Biswal, "Linear fractional programming problems with some multi-choice parameters," International Journal of Operational Research, vol. 34, no. 3, p. 321, 2019, doi:10.1504/ijor.2019.098310.

[31] S. K. Singh and S. P. Yadav, "Scalarizing fuzzy multi-objective linear fractional programming with application," Computational and Applied Mathematics, vol. 41, no. 3, Mar. 2022, doi: 10.1007/s40314-022-01798-2.

[32] X. Dastile, T. Celik, and M. Potsane, "Statistical and machine learning models in credit scoring: A systematic literature survey," Applied Soft Computing, vol. 91, p. 106263, Jun. 2020, doi:10.1016/j.asoc.2020.106263.

[33] N. Kozodoi, S. Lessmann, K. Papakonstantinou, Y. Gatsoulis, and B. Baesens, "A multi-objective approach for profit-driven feature selection in credit scoring," Decision Support Systems, vol. 120, pp. 106–117, May 2019, doi: 10.1016/j.dss.2019.03.011.

[34] E. S. Kamimura, A. R. F. Pinto, and M. S. Nagano, "A recent review on optimisation methods applied to credit scoring models," Journal of Economics, Finance and Administrative Science, vol. 28, no. 56, pp. 352–371, Jun. 2023, doi: 10.1108/jefas-09-2021-0193.

[35] H. He, Z. Wang, H. Jain, C. Jiang, and S. Yang, "A privacy-preserving decentralized credit scoring method based on multi-party information," Decision Support Systems, vol. 166, p. 113910, Mar. 2023, doi:10.1016/j.dss.2022.113910.