Application of Data Mining and Machine Learning Techniques to Predict Loan Approval and Payment Time

Ehsan Allah Khoshkhoy Nilash[1] , Mansour Esmaeilpour[2•] , Behrooz Bayat[3], Alireza Isfandyari Moghaddam[3] and Erfan Hassannayebi[4]

***Abstract*-- One of the most important issues regarding banks is knowing the customers, their behaviors, and the decisions these institutions make regarding customers' preferences. Their main task is to provide banking facilities. Bank facilities carry the risk of default in repayment. Failure to evaluate and review factors related to repayment can cause significant damage to banks. On the other hand, investment in the private sector and various industries is also increasingly important. This action can lead to economic growth, increased employment, and national income. This research aims to identify the effective features related to the fixed capital facility data of one of the active banks in Iran, in line with the classification of customers into two categories good customers and overdue customers to predict the duration of the facility payment. The five-step method is based on data mining techniques. The most important steps of this method are data preparation, analysis with rough set methods, and common classification techniques such as artificial neural networks, tree types, Bayes types, and support vector machines. One of the most important results of this research was the identification of the features that affect the repayment and duration of fixed capital facilities. Additionally, among other results of the present research, the ANN method demonstrated superior performance in evaluating credit risk with an accuracy value of 70.27%, and the J48 technique showed superior performance in predicting the duration of payment of facilities with an accuracy of 72.54%.**

***Index Terms*-- Fixed capital facility, Facility repayment, Credit risk, Data mining, Classification**

I. INTRODUCTION

Providing services to customers by banks dates back to 200 years ago[1]. Banks are important and influential institutions in any country. One of the most important issues regarding banks is knowing the customers, their behaviors, and the decisions made by these institutions toward the customers. Banking has undergone changes in many ways throughout the world over the years[2]. Today, most banks offer a wider range of products and services than ever before. Their main duties in the field of operating the society's surplus funds, deposits, and investments by providing banking facilities, i.e. providing loans to people, have remained unchanged. Accurate consumer assessment is very important in the banking sector and other organizations. Loans and bank facilities have the risk of default in repayment. These loans are associated with a lot of capital, and failure to evaluate and review the factors in their repayment can cause a significant loss to the financial institution. Therefore, accurate risk assessment is very important for banks and other organizations. It is not only the minimization of credit risk that is important but the problems and issues of valid customer rejection are equally important [3]. Also, credit risk assessment for loans is an important measure in banking systems to ensure loan payments by lenders and to rank banks as institutions with good performance in implementing regulations [4]. Risk analysis in bank loans requires a precise understanding of the meaning of risk. Additionally, the number of transactions in the banking sector is growing rapidly, and a large amount of data is available, representing diverse customer behaviors. This can increase the risks associated with lending. To gain a better understanding of customers, banks should identify effective factors and characteristics for loan repayment. These institutions take measures based on customer information and behavior to provide services, particularly banking facilities. This process is known as the validation process [5].

Statistical studies have shown that, even though financial institutions receive numerous guarantees from their customers in exchange for receiving facilities, in many cases, repayment of claims faces a significant delay. Therefore, customer validation, which means identifying the characteristics and factors affecting repayment of facilities, is still one of the most important issues facing bank managers. Some researchers have investigated several statistical classification models for consumer validation. They concluded that on this basis there is no best method and the choice of the best method depends on parameters such as data structure, other variables, and contextual characteristics. In addition, they stated that when the data is not structured, it is better to use flexible intelligent methods based on data mining such as neural networks [6]. On the other hand, as the repayment of bank facilities is important, investment in the private sector and various industries is also increasingly important. This action can lead to economic growth, an increase in employment, and national income[7]. Therefore, providing bank facilities, especially fixed capital, at the right time to production, tourism, and mining units, as well as timely repayment of facilities, is of great importance and

1. Department of Management, Hamedan Branch, Islamic Azad University, Hamedan, Iran.
2. Department of Computer Engineering, Hamedan Branch, Islamic Azad University, Hamedan, Iran.
• Corresponding author Email: esmaeilpour@iauh.ac.ir
3. Department of Knowledge and Information Science, Hamedan Branch, Islamic Azad University, Hamedan, Iran.
4. Department of Industrial Engineering, Sharif University of Technology, Tehran, Iran.

value. Therefore, to identify their customers and optimize the processes of providing facilities, banks need to process and discover hidden knowledge from the data in the information systems related to these facilities based on efficient, intelligent, and up-to-date information technology methods [8].

Data related to information systems regarding customers and facilities are one of the most valuable sources for any bank or financial institution to discover knowledge in line with banking activities, such as customer validation or payment time prediction. One effective and useful method in this regard is data mining. Data mining provides the possibility of extracting knowledge from historical data and predicting the results of future conditions. This method helps optimize business decisions, increase the value of each customer, improve communication, and enhance customer satisfaction [9]. In this regard, considering the many issues and problems of banks and credit institutions in obtaining their claims from customers in connection with banking facilities, especially fixed capital, and the high importance of timely payment of such loans to various industries and active businesses, the purpose of this research is to investigate the factors affecting the repayment on time and without delay. Additionally, the research aims to predict the payment time of such facilities using data mining methods.

In this research, we aim to investigate the factors that affect repayment in both the good and overdue customer classes, as well as predict the duration of fixed capital facility payment. For this case study, we will be using data from the facility information system of an active bank in Iran. The customer classes are categorized based on the duration of delay in payment of installments. Specifically, customers who make timely repayments are referred to as good customers, while those who experience delays in repayment are referred to as overdue customers. The salient point of the present research is the authenticity of the available data related to bank customers and the use of different data mining classification methods and techniques, namely Rough Set, types of trees, types of Bayes, artificial neural networks, support vector machines supported by Weka, RapidMiner, and Rosetta.

In this regard, the questions of the present research are:

1. What features affect the timely repayment of fixed capital facilities based on data mining methods?
2. What factors are effective in predicting the payment time of fixed capital facilities using data mining techniques?
3. Which data mining technique performs best for categorizing customers into good and overdue classes?
4. Which data mining technique demonstrates superior performance in predicting the payment time of fixed capital facilities?

II. RESEARCH LITERATURE

In this section, the concepts and literature related to the research and in the form of research background, the studies related to the present research by other researchers are reviewed.

A. Rough set theory

Today, we are faced with uncertainties and ambiguities when making choices. Many concepts and theories of uncertainty, such as fuzzy sets, have been introduced in the past. In recent years, mathematical tools based on these concepts have been developed rapidly. One important theory in this regard is the rough set theory. The concepts of rough set theory are based on the assumption that each member of a set is associated with certain information features. This theory is particularly important in various branches of artificial intelligence, including classification, learning algorithms, and pattern recognition. It has a strong potential for research in these fields [10, 11]. The Rosetta software supports different algorithms for discretization, data reduction, and rule generation related to this theory. These algorithms include entropy, Johnson, genetics, and Holte's.

B. Data mining

The evolution and growth of information storage and collection technologies have led to a large amount of data in various fields such as financial, health, educational businesses, and other fields. The set of activities involved and related to the analysis of these large databases to obtain useful knowledge to help decision-making is possible in various ways, the most important of which are data mining, knowledge discovery, pattern recognition, and machine learning. In particular, data mining has an increasing role in theoretical and applied studies. Data mining refers to the process of examining and analyzing large datasets to find regular patterns extract relevant knowledge and obtain meaningful repetitive rules [12].

Data mining is an iterative process in which different models and techniques play a key role. This process includes steps such as specifying objectives, data collection, data integration, feature selection, model development and validation, and finally analysis and interpretation [13].

Data mining has a variety of techniques, including classification, clustering, association, and prediction [14]. Classification is the most common data mining method that uses a set of historical data samples pre-categorized labeled or classed to create a model. This model can be used to predict the class or category of other samples in the future. This technique usually has two main stages: training and testing. In the training stage, historical data is trained based on different classification methods and techniques, and then using the test dataset, the accuracy and precision of the predictive model are evaluated. Classification has a variety of methods. Some common and widely used methods include types of trees, types of Bayes, neural networks, support vector machines, etc. [15]. In the following, each of these methods will be discussed briefly.

B.a. Decision tree

A decision tree is a method that can easily display a large amount of data. This method can classify data based on characteristics. Inferring decision trees means learning them from training samples with specific category labels. The structure of the decision tree is similar to the structure of a flowchart, where each intermediate node represents the test on one or more features, each branch represents the output of the test, and each leaf node contains a category label. The decision tree shows the rules for dividing the data into different categories. Decision trees use the feature that has more

information gain as the basis of action. There is a feature in the root that has the most information gain. Decision tree techniques recursively perform top-down construction and bottom-up deletion or pruning. Algorithms such as ID3, C4.5, C5.0, CART, etc. are among the types of decision trees[15, 16].

B.b. Types of Bayes

Among the most common methods based on probabilities, we can refer to the category of naive Bayes and Bayes networks. The naive Bayes method is based on Bayes' statistical theorem. This type of category assumes that all properties are correlated with the class property. This method can predict the probability of each feature belonging to a class. This method states that the belonging or non-belonging of one feature to the class is not affected by other features. The mentioned method is very suitable when the number of inputs is large [16]. This method requires only a small amount of training data to determine the estimation of the required parameters in the classification process, and it often works much better than expected in complex and real-world situations [17, 18]. Bayes network method based on a non-cyclic directed graph includes nodes that represent random variables and arcs that represent possible relationships between variables. Each node has a local probability distribution that is independent of the state of its parents. In other words, the joint distribution of the variables is shown by the product of the conditional distributions of each variable concerning its parents[19].

B.c. Artificial neural networks

The artificial neural network type is modeled on the human brain, and data processing is entrusted to small and many processors that are connected in a network and behave in parallel to solve the problem in question. With the help of programming knowledge, these networks are designed as a special data structure that can act like neurons in the human brain. Neural networks have various types, the most famous and common type of neural network is simple and multi-layer perceptron. Single-layer perceptron networks are used for simple linear problems, and multi-layer networks are used for more complex problems. Each neural network generally consists of several layers, including input, hidden, and output layers. Adjusting the weights for neurons based on different inputs is used to train the neural network. Learning in neural networks is done by the training dataset, and it is in line with the desired output [8]. Artificial neural networks are capable of solving problems in various fields such as agriculture, medical sciences, education, finance, management, security, engineering, business, and art. They can even solve issues and problems in areas such as transportation, computer security, banking, insurance, real estate management, marketing, and energy that cannot be solved with traditional procedures and ordinary mathematics. Despite the widespread applications of neural networks, this method requires a systematic approach in the initial training and configuration stage. For example, it is important to consider data accuracy, data working tools, data standardization, types of inputs, data segmentation, pre-processing, and validation [20].

B.d. Support vector machines

Support vector machine is one of the supervised learning methods used for classification and regression. The working basis of the SVM classifier is the linear classification of the data, and in the linear division of the concrete data on the linear cloud, the data are classified according to the class to have a higher confidence margin. Samples close to the line are called support vectors. Therefore, the line that has the largest distance or margin with the support vectors is more optimal. For non-linear data, the data is linearized using the kernel concept. Among the most common kernels are Polynomial, Radial Basis, and Sigmoidal [21].

C. Research background

Zandi et al. used a dynamic multilayer network based on a graph neural network and recurrent neural network from different sources to assess credit risk. They utilized a US mortgage dataset that included various information related to the borrower's geographic location and loan provider choices. The results demonstrated that the proposed model outperformed traditional methods [22]. Chen, Jin, and Lu used machine learning based on the daily statement data of micro, small, and medium enterprises in China to evaluate the credit risk and arrears of such enterprises to receive loans. The algorithm used in this research was the backpropagation neural network classifier based on a genetic algorithm. The results showed the high accuracy of the method[23]. Montevechi, Andre Aoun, et al reviewed the research related to credit risk in a review study. The results showed that the modeling process for decision tree methods, support vector machines, and neural networks is lengthy in practical applications and commercial use for credit risk analysis. On the other hand, data preprocessing, such as cleaning, transformation, and data reduction for feature selection, is of increasing importance. In addition to the above, this systematic study showed that decision tree and Random Forests models have improved performance, and the application of machine learning in credit risk analysis is undeniable[24]. Xiaoming Zhang and Lean Yu proposed a feature-based modeling framework for credit datasets based on types of traditional single classifiers, intelligent single classifiers, hybrid, and ensemble multiple classifiers to solve credit risk problems[25]. Addy, Wilhelmina Afua, et al. reviewed the studies conducted in the banking sector related to predictors in credit risk management. This study showed that model selection is of increasing importance in such research. The study also points out the obstacles facing such analyses, such as data quality, ethical considerations, model interpretability, and implementation costs. In addition to the aforementioned issues, this comprehensive study confirms the effectiveness of such predictors[26]. Chandrasiri and Premaratne used data mining classification techniques in a study for credit risk management based on a data set of a financial institution in Sri Lanka. Among the algorithms used in this study are Naïve Bayes, Rule Induction, Decision Tree,

Random Forest, and KNN. It should be noted that the KNN algorithm performed better[27]. Jumaa, Saqib, and Attar used data mining techniques, especially deep learning, in a study to increase the accuracy of default prediction and credit risk management. The analyzed dataset was a survey of 1000 participants. The Keras library was used to build the model. The accuracy of the proposed model in the evaluation, based on 250 records, was obtained above 95%[28]. Krishnaraj, Rita, and Jitendra Jaiswa used data mining to identify effective features based on telemarketing data on term deposits associated with a bank in Portugal. In this research, decision trees and random forest algorithms were used for analysis. The results showed that the two characteristics of age and inventory had the greatest effect on customer sharing, and the random forest algorithm performed better[29]. Mayank Anand, Arun Velu, and Pawan Whig used multiple logistic regression, decision trees, random forests, Gaussian Naive Bayes, and support vector machines to predict bank loan arrears based on an internet dataset. The results showed the superiority of the decision tree and random forest[30]. Justin Munoz et al. used machine learning techniques based on two categories: the first type identifies customers to receive a loan, while the second type confirms the eligibility of the customer to receive a loan. In other words, it covers factors such as customer credit risk. Among the methods used, deep learning performed better and achieved ideal performance for 20 to 25 customers. Additionally, for 50 customers, it demonstrated a performance with an accuracy of over 85% [31]. Ashenafi Wubshet Desta and J. Sebastian Nixon used naive Bayes and decision tree techniques to identify delinquent customers. In this research, 20,461 random sample data related to a bank's database in a three-year period were used for prediction. The present research proposed the decision tree method for applying data mining in similar applications [32]. Wang et al. proposed an approach called RSFS for rough set-based feature selection and scatter search for validation. In RSFS, conditional entropy is used to search for optimal solutions. They used two datasets from the UCI database and analyzed them using neural networks, decision trees, and logistic regression techniques. The results showed that RSFS has better performance in improving classification accuracy compared to common classification methods [33]. In a study, Crone et al. used sampling to predict consumer repayment behavior. In this research, logistic regression techniques, discriminant analysis, decision trees, and neural networks were used in two datasets with 20 samples of large size and 29 samples of balanced distribution. The results showed that larger samples had better accuracy [34]. Koutanaei, Sajedi, and Khanbabaei developed a data mining model that combines feature selection and learning-based classification algorithms. The results showed that the PCA[1] algorithm is the best feature selection algorithm, and the artificial neural network adaptive enhancement method has a higher classification accuracy[35]. Gulsoy et al. proposed a method to assess the objective risk of lending to SME business customers. They also used classification techniques to validate current customers in the bank. In the present research, six different classification algorithms and Waka software were used. The results showed the high accuracy of these algorithms and the efficiency of the method [36]. Jisha, Kumar, and Vimal used data mining in the banking sector for fraud prevention and detection, customer retention, marketing, and risk management. The results showed the efficiency of techniques and methods based on data mining [37]. Hamid et al. presented a model using data from the banking sector to predict the status of loans. Decision tree, Bayes network, and navie Bayes techniques were used in this model. The results showed that the decision tree had the best accuracy [38]. In research, Homan et al. investigated the application of various common data mining methods and hybrid methods for customer validation. This research is a basis for the researcher to achieve the best methods in this regard [39]. Mojtabi Salehi and Alireza Kurd Katoli used a combination method of the colonial competition optimization algorithm and neural network to evaluate credit risk. In addition to categorizing high-risk and low-risk customers, this method eliminated ineffective features. For the analysis, the real data set of the UCI database as well as the real data of an Iranian private bank were used. The results showed a reduction in the error of the neural network method [8]. Meysham Jafari Eskandari and Milad Rouhi presented a research model that uses data mining methods to predict the percentage of receivables collection for contracts with a high risk of receivables before granting facilities. The results indicate that the proposed model in this research is more accurate[40].

Although each of the reviewed studies has reached its goals, some of them have only focused on the customer's credit risk and the proper use of bank resources, neglecting the examination of the customer's side, specifically the timely payment of facilities to industries. In other words, no research has been conducted on predicting the duration of facility payments, which can be considered a knowledge gap. Another issue is the lack of comprehensive use of classification methods together based on the use of different data mining tools. Additionally, none of the studies reviewed in the research literature included an analysis that combines a process-oriented perspective with data mining. Therefore, this study addresses these knowledge gaps, which is considered a unique contribution of the present study.

III. RESEARCH METHOD

The method used in this research is based on data mining concepts and techniques. In this research, the data of the related facility provision system of an active bank in Iran is used. This five-step method includes data extraction and preparation, classification with the method based on rough set theory,

1. principal component analysis

classification with common data mining methods, analysis of the effects of data characteristics, and comparison and analysis of the results. The mentioned method is depicted in Fig 1.

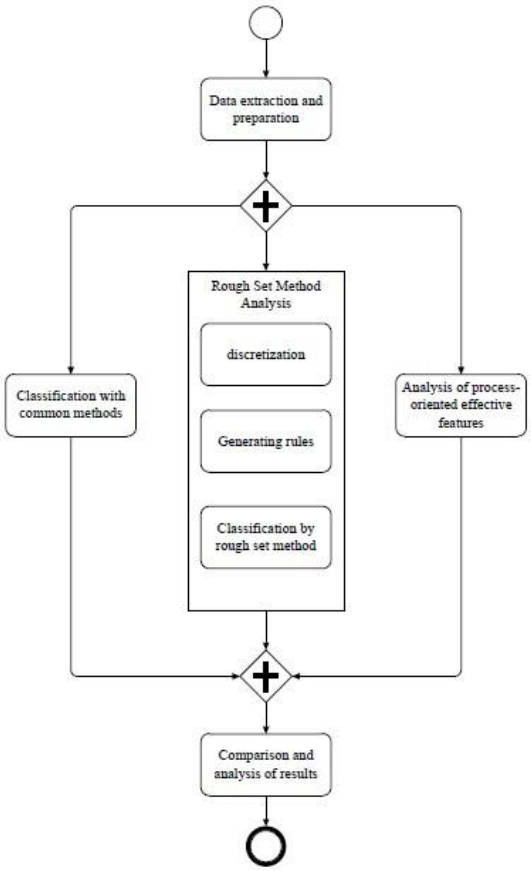

Fig. 1 Proposed method

A. Data extraction and preparation

The dataset used in the present research is extracted from the database of the facility provision system of an active bank in Iran. This dataset includes two main tables with data-oriented and process-oriented views extracted from the database of the facility system. Facilities with natural customers and loans other than fixed capital are excluded from the research. Additionally, to extract the process-oriented data, event logs resulting from the execution of the fixed capital process are necessary; therefore, loans that have not reached the payment stage are also excluded from the study.

The data-driven type includes only features related to the payment of loans to legal entities. This data encompasses information regarding the legal entity receiving the loan, the financial characteristics of the loan, and the associated label. According to the information system data, individuals without arrears are labeled as good customers, while those with bank arrears exceeding one overdue installment are classified as overdue customers. The information features in this view include customer size, type of institution, currency, type of industry, facility source, type of borrowing, type of contract, type of basic contract, credit line, payment method, number of installments, type of fixed capital application, loan amount in millions of Rials, and loan repayment status as a class.

The second table is the process-oriented one, including event logs related to the fixed capital facility process. It includes the loan acceptance number, activity or task name, start timestamp, end timestamp, and data features such as industry type, customer size, facility source, contract type, basic contract type, credit line, profit rate, loan amount, number of installments, and payment period. The Fluxicon Disco software is used to calculate the loan payment period. Loans of more than one year are labeled as class A, while those less than one year are labeled as class B. At the end of the work, the selected tables are cleaned from noise, missing data, etc., and saved in CSV, Excel, and other required formats. This table is used to predict the duration of facilities and analyze the effect of data features. To predict the duration of the loan, the fields of loan acceptance number, activity, start timestamp, and end timestamp are filtered.

B. Classification using methods based on rough set theory

In this step, based on the rough set theory method under the support of Rosetta, the work of customer credit risk assessment is carried out. This involves categorizing customers into good and overdue classes and predicting the loan payment period. The data prepared in the previous step is used for this purpose. The prepared tables, which are data-oriented and process-oriented separately, are first discretized using the entropy algorithm. Then, based on Johnson's, genetics, and Holte's algorithms, the features that affect the class are specified, the data is reduced, and the rules are generated. The best algorithm with the least number of rules and the highest LHS[1] is selected. Finally, the trained data are evaluated separately using the naive Bayes classification technique to check the customer's credit risk and predict the duration of the facilities.

C. *Classification with common data mining methods*

In this step, the data prepared in the previous step is used to classify customers into good and overdue classes and predict the loan payment period. Common classification methods such as tree types, Bayes types, artificial neural networks, and other techniques supported by Weka and RapidMiner are employed. Customer credit assessment is performed based on a prepared data-oriented table, while payment time prediction is based on a prepared process-oriented table, using various common classification methods. The accuracy and correctness of these methods are compared separately in a tabular format. Additionally, facilities are classified based on the mentioned techniques, with payment term tags of more than one year and less than one year.

D. Analysis of the effect of information features

Based on the process-oriented table, using the Celonis tool, the effect of information features on the entire process, i.e., its execution time, is analyzed. The information fields required by the process mining, in addition to the mentioned data features, are the prerequisites of this stage. Comparison of methods

At this stage, the results of the classification based on

[1] left and side

customer credit assessment and forecasting the duration of facility payment based on the rough set method and common classification techniques are compared with each other based on the percentage of correct classification and the accuracy value of each method for the desired class.

IV. Results and findings

In this section, the findings and results related to the research are presented according to the steps mentioned in the research method.

A. Data extraction and preparation

The number of extracted records for the data-driven table included 77,836 records, after filtering non-legal customers, non-fixed capital loans, and noise data such as the negative loan amount, 3,594 records remained. Out of all the remaining records, 2,433 had a deferred customer class and the other records had a good customer class. For the process-oriented table, out of 82,387 records containing event logs related to loan stages, extracted from the facility system after cleaning, 57,208 records remained. Also, with the filter of samples that have not reached the payment stage, the payment duration was calculated by entering the data into the DISCO tool. Furthermore, by filtering samples with a duration of less than 60 days and removing records related to the stages of the facility process, a total of 422 samples of payment facilities with the mentioned information features remained for data mining. The tables were saved in the required formats of Weka, Rapid Miner, Rosetta, and Celonis tools, such as CSV and ARFF. Additionally, to analyze the effect of the information features of all 57,208 records, including event logs resulting from the execution of process steps, they were stored in a separate table.

According to the number of dataset records, for the first type, 66% of the data was used for training and the rest for testing, and for the second type, 80% of the data was used for training and the rest for testing. In Table No. I, the features used in the research are listed according to customer credit assessment and loan payment duration prediction.

TABLE I
Features Used in the Study

Credit Risk		Facility Throughput Time	
Features	Type	Features	Type
Customer Size	P	Base Contract Type	P
Institute Type	P	Contract Type	P
Currency	P	Credit Line	P
Facility Source Type	P	Customer Size Type	P
Moratorium Type	P	Economic Section	P
Base Contract Type	P	Amount Of Loan	P
Contract Type	P	Profit Rate	P
Credit Line	P	Number Of Installments	P
Payment Method	P	Facility Source	P
Industry Type	P	Case Duration Label	C
Facility Application	P	C: Class Feature P: Prediction Feature	
Loan Amount	P		
Number Of Installments	P		
Customer Label	C		

Fields case ID, Activity, End Time Stamp, and Start Time Stamp are also used for effect analysis in addition to process time prediction features.

Below are explanations of some of the information features, such as Customer size, Contract Type, and Credit Line.

1: Customer Size

This informational feature refers to the size of the applicant for the facility. In the bank under study, these customers usually include small and medium-sized industrial units as well as large industries. It is worth noting that this ranking is usually done based on their production capacity.

2: Contract Type

This informational feature refers to the type of contract for receiving the facility. For example, the famous types of contracts include interest-free, Salf, Civil Partnership, etc. It is worth noting that the type of contract has many variations based on the type of basic contract. The types of basic contracts used in the dataset under study are: Ja'ala, Salf, Civil Partnership, interest-free, Mudarabah, and Installment Sales.

3: Credit Line

This informational feature refers to the line of credit provided to access financial resources for the facilities. For example, from Note 29 of the 2010 budget - Managed funds.

A part of the coded data used for the analysis of risk assessment and prediction of loan payment time is given in Fig. 2.

Credit Risk Classification

Customer SizeType	Institute Type	Currency ID	Facility Source	Moratorium Type	Base Contract Type	Contract Type	CreditLine ID	Payment Method	Industry Type	Facility AppType	Loan Amount	Number Of Installments	Class
1	7	IRR	740	180	12	102	5	0	27	1	1264	56	overdue_loan
1	7	IRR	730	190	12	102	1	0	27	1	605	54	overdue_loan
1	1	IRR	730	180	12	102	45	0	24	1	8251	43	overdue_loan
1	1	IRR	730	180	12	102	61	0	14	2	2899	43	overdue_loan
2	1	IRR	730	180	20	200	1	1	24	2	14700	54	Good
2	1	IRR	730	180	12	100	1	1	24	2	4000	48	Good
1	1	IRR	730	190	12	102	38	0	15	1	10000	60	Good
1	1	IRR	730	190	12	102	61	0	15	1	32300	36	Good
1	1	IRR	740	180	12	102	3	0	24	1	10000	25	overdue_loan

Facility Throughput Time Classification

Base Contract Type	Contract Type	CreditLine ID	Customer Size	Economic Section	Amount Of Loan	Profit Rate	Facility Source	Number Of Installments	Duration Class
B	A	E	A	A	25000	16	A	60	B
B	A	P	C	C	2000	18	A	60	A
B	A	E	A	A	10000	16	A	60	A
C	B	P	C	A	1500	18	A	60	A
B	A	P	A	A	37000	18	A	60	A
B	A	S	B	A	44500	18	A	60	A
C	B	V	C	A	1000000	12	A	60	A
B	A	X	A	A	5450	12	C	60	A
B	A	AG	A	A	10000	17	A	60	A

Fig. 2. A part of the data used.

Rough set-based entropy and genetic algorithms, as well as Weka-based decision trees, were used to reduce features.

B. Classification based on rough set theory

At this stage, the finalized data was entered into the Rosetta tools separately in the direction of the determined goal. After discretization using the entropy algorithm, based on the genetic, Johnson, and Holte's algorithms, the process of reduction and generation of rules is adopted. In the following, the results are given separately for credit risk assessment and predicting the duration of facility payment.

B.a Customer credit risk assessment

According to the rough set methods, among the 13 features used, 11 of them were found to be effective in the analysis as follows, which are:

Customer Size, Institute Type, Moratorium Type, Base Contract Type, Contract Type, Credit Line, Payment Method, Industry Type, Facility Application, Loan Amount, and

Number of Installments.

Further, according to the generation of rules based on Johnson and genetic methods, the highest LHS value was 0.01. However, based on Holte's algorithm, 184 rules were produced, of which 3 rules had an LHS greater than 0.9. Finally, the performance of the method was evaluated using the naive Bayes technique based on the good customer class. The accuracy of the method was found to be 66%, with the accuracy of the good and overdue customer classes being 54% and 73% respectively.

B.b Predicting the time of payment for the facility.

According to the rough set methods, among the 9 features used, 7 of them were found to be effective in the analysis as follows, which are:

Contract Type, Credit Line, Customer Size, Economic Section, Amount of Loan, Profit Rate, Number of Installments

Further, according to the generation of rules based on Johnson and genetics methods, the highest LHS value was equal to 0.02. However, based on Holte's algorithm, 95 rules were generated, of which 3 generated rules had LHS greater than 0.9. Finally, using the naive Bayes technique based on the good customer class, the performance of the method was evaluated. The accuracy of the method was 66%, with the accuracy of the classes less than one year and more than one year being 76% and 59%, respectively.

In Fig. 3, the results of classification evaluation are depicted separately for credit risk and predicting the duration of facilities based on rough set methods.

Since the rough set method uses the Naive Bayes algorithm, it works based on the probability value of each class according to the Bayesian probability formula. This means that if the probability calculated for each sample in the class is larger, that entry is assigned to that class. In some cases, the probability calculated for a particular sample may be equal for both classes. In such cases, the classifier predicts the label "Undefined".

Facility Throughput Time Classification

		Predicted			
		B	A	Undefined	
Actual	B	16	12	2	0.533333
	A	11	40	3	0.740741
	Undefined	0	0	0	Undefined
		0.592593	0.769231	0.0	0.666667
ROC	Class	A			
	Area	0.747549			
	Std. error	0.054685			
	Thr. (0, 1)	0.628			
	Thr. acc.	0.608			

Credit Risk Classification

		Predicted			
		overdue _loa	Good	Undefined	
Actual	overdue _loa	670	123	24	0.820073
	Good	246	146	13	0.360494
	Undefined	0	0	0	Undefined
		0.731441	0.542751	0.0	0.667758
ROC	Class	Good			
	Area	0.696034			
	Std. error	0.01685			
	Thr. (0, 1)	0.364			
	Thr. acc.	0.576			

Fig. 3. Evaluation results based on rough set methods

C. *Classification based on common techniques*

At this stage, the finalized data was entered into Weka and Rapid Miner software, according to the set goal, by separating the risk assessment tables and predicting the duration of the loan. Classification was done based on the mentioned classes using techniques such as tree types, Bayes types, support vector machines, and artificial neural networks. The accuracy values of different algorithms, according to Weka and Rapid Miner tools, are listed in Table No. II.

For different algorithms, default values of software parameters were used. For example, radial and polynomial kernels were used for the SVM method. The artificial neural network model used in the research was two-layer.

TABLE II
Comparison of Accuracy of the Used Algorithms.

Tools	Classifier	Accuracy for Duration Time Predictive	Accuracy for Credit Risk Predictive
Weka	J48	**72.61**	68
	Random Forest	72.61	67.26
	Random Tree	61.9	62.43
	REP Tree	69.04	69.23
	Decision Stump	69.04	67.18
	Logistic model tree	64.28	70.37
	Bayes Network	64.28	69.31
	Naive Bayes	50	42.88
	Naïve Bayes Multinomial Text	69.04	67.18
	Naïve Bayes Updateable	50	42.88
	Multilayer Perceptron	67.85	**70.54**
	Voted Perceptron	69.04	67.18
RapidMiner	Decision Tree	52.94	68.8
	Decision Stump	63.53	**70.27**
	Random Forest	64.71	67.57
	Naïve Bayes	56.47	69.70
	Gradient Boosted Tree	64.34	63.23
	Naïve Bayes(kernel)	59.52	52.05
	Deep learning	60	61.83
	Lib SVM	52.94	67.57
	SVM(linear)	60	69.45
	SVM(Evolutionary)	63.53	67.81
	NN	61.18	67.08

D. Analysis of process-oriented effective features

The table prepared in the first stage, including the prerequisite features of the process mining along with the data features, was entered into the Celonis tool. Based on various features such as the customer size, the type of basic contract, and the type of contract, etc., with the target completion time of 271 days, that is, the average of cases throughput time, an analysis was done. The results showed that three features of the basic civil partnership contract, the civil partnership contract type, and the SME customer respectively had the most effect in reducing the time of the process. The results are shown in Fig. 4.

Fig. 4. Analysis of the effects of features

E. Comparison of methods

At this stage, the performance of different methods based on different tools is compared with each other, separately with the objectives of credit risk assessment analysis and facility payment time prediction.

E.a Customer credit risk assessment

According to the obtained results, when comparing the techniques of common classification methods and rough set methods, the multi-layer perceptron artificial neural network algorithm based on Weka tools showed the best performance with an accuracy of 70.54%. The next best algorithm, based on the Rapid Miner tool, was the decision stump technique with an accuracy of 70.27%.

According to the good performance of the J48 algorithm based on Weka, some rules with a high approximate coverage rate from all the tested samples are listed as follows:

Rule1: FacilityAppType = 1: overdue _loan (2481.0/672.0)
Rule2: FacilityAppType = 2
CreditLineID = 61
| BaseContractType = 12: overdue _loan (219.0/65.0)
Rule3: FacilityAppType = 2
CreditLineID = 38: overdue _loan (95.0/31.0)

E.b Predicting the time of payment for the facility

According to the obtained results, when comparing the techniques of common classification methods with Rough Set methods, the J48 decision tree and Random Forest algorithms based on Weka tools performed the best with an accuracy of 72.61%. Additionally, the best method based on Rapid Miner tools was the Random Forest technique with an accuracy of 64.71.

Due to the good performance of the J48 algorithm based on Weka, the following rules with a high approximate coverage rate from all tested samples are listed as follows:

Rule1:
BaseContractType = A
| CreditLineID = P: A (78.0/25.0)
Rule2:
BaseContractType = A
| CreditLineID = Q
| AmountOfLoan_Million > 25000
| | | ProfitRate > 17: B (63.0/22.0)
Rule3:
BaseContractType = B: A (78.0/18.0)

In Table No. III, different methods based on different analyses and various tools are compared with each other.

TABLE III
Comparison of the Best Algorithms Based on the Methods Used

Analysis Type	Technique Domain	Tools	Best algorithm	Accuracy(%)
Credit Risk	Rough Set	Rossetta	Naïve Bayes	66.7
	Common techniques	Weka	Multilayer Perceptron	70.54
	Common techniques	Rapid Miner	Decision Stump	70.27
	All techniques	All Tools	**Multilayer Perceptron**	**70.54**
Throughput Time Prediction	Rough Set	Rossetta	Naïve Bayes	66.6
	Common techniques	Weka	J48	72.61
	Common techniques	Rapid Miner	Random Forest	64.71
	All techniques	All Tools	**J48**	**72.61**

Fig. No.5 shows the performance of the best method based on credit risk assessment and predicting the duration of the facility.

Facility Throughput Time Classification -J48

Correctly Classified Instances	61	72.619	%
Incorrectly Classified Instances	23	27.381	%

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.346	0.103	0.600	0.346	0.439	0.293	0.616	0.433	B
0.897	0.654	0.754	0.897	0.819	0.293	0.616	0.761	A
0.726	0.483	0.706	0.726	0.701	0.293	0.616	0.660	

Credit Risk Classification -Multilayer Perseptron

Correctly Classified Instances	862	70.5401	%
Incorrectly Classified Instances	360	29.4599	%

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.974	0.845	0.702	0.974	0.816	0.241	0.650	0.766	overdue _loan
0.155	0.026	0.747	0.155	0.256	0.241	0.650	0.521	Good
0.705	0.576	0.717	0.705	0.633	0.241	0.650	0.685	

Fig. 5. According to the analysis done, the performance of the best techniques.

V. Discussion and Conclusion

This study showed the efficiency of data mining methods in assessing credit risk and predicting the duration of payment for facilities. The methods utilized data features related to legal clients and fixed capital facilities. In other words, these methods can identify customers based on their classification as good or overdue, and accurately predict the duration of facility payments for classes under one year and over one year.

One of the initiatives of the present research is to pay attention to and examine the customer's credit risk at the same time as ensuring timely payment of resources to legal customers, i.e. industry activists. By examining the studies of predecessors, a knowledge gap was identified. In this regard, this study is based on different techniques and algorithms of machine learning and artificial intelligence. Another initiative of the present method is the data-oriented analysis, along with the process-oriented method, to discover knowledge from the real data of one of the active banks in Iran, which was not observed much in the studies of others.

In line with the answer to the first research question based on the rough set, the features of Customer Size, Institute Type, Moratorium Type, Base Contract Type, Contract Type, Credit Line, Payment Method, Industry Type, Facility Application,

Loan Amount, and Number of Installments, and based on the J48 decision tree, the features of Facility Application Credit Line, Base Contract Type, Moratorium Type, Industry Type, Contract Type, Loan Amount, Payment Method, and Number of Installments for Customer credit assessment were recognized as effective.

In order to predict the payment time of fixed capital facilities based on rough set methods, the features of Contract Type, Credit Line, Customer Size, Economic Section, Amount of Loan, Profit Rate, and Number of Installments were found to be effective. Also, based on the common data mining algorithms of the best technique, namely J48, the features of Base Contract Type, Credit Line, Amount of Loan, Profit Rate, and Customer Size were recognized as effective.

In the present study, different types of techniques and algorithms of Rough Set, types of trees, artificial neural networks, types of Bayes, and support vector machines were used. Additionally, different tools such as Weka, Rapid Miner, and Rosetta were utilized. For credit risk, Perceptron's multi-layer artificial neural network algorithm performed best with 70.54% accuracy based on Weka. On the other hand, the J48 algorithm performed well with 68% accuracy. In this regard, one of the production rules indicates that if the facility is used to create a new business, it is expected that the customer will have arrears. In another rule, it was mentioned that if the use of facilities is of the completion and development type, the line of credit is of the industry and internal resources type, and the type of the basic contract is the installment sale type, it is expected that the customer will be in arrears. Also, in another rule, it was stated that if the use of development and supplementary facilities and the credit line of those economic and entrepreneurial enterprises is from internal sources, it is expected that the customer will have arrears.

On the other hand, in the analysis of predicting the payment time of the facilities, the j48 and Random Forest algorithms based on the Waka tool had the best performance with 72.61% accuracy. In this regard, one of the production rules indicates that if the basic contract is of the civil partnership type and the credit line is of the internal resources of the branches, the loan payment period is expected to be less than one year. In another rule, it was determined that if the type of contract is an installment sale, then the loan payment period is expected to be less than one year. Also, another rule states that if the basic contract is of the civil partnership type, the credit line is from the internal sources of the headquarters, the loan amount is more than 25 billion Rials, and the interest rate is more than 17%, it will take more than one year to pay off the loan.

The present research showed that tree-based algorithms had high performance in both analyses. One of the other results of this research was the closeness of the performance of the rough set method to the best common methods.

On the other hand, based on the analysis type separation, to evaluate the credit risk of the features Facility Application, Credit Line, Base Contract, Moratorium Type, Industry Type, Contract Type, Loan Amount, Payment Method, and Number of Installments, and also to predict the payment period of the features Credit Line, Amount of Loan, Profit Rate and Customer Size were recognized as effective jointly between the two mentioned methods.

Another initiative of the present research was a process-oriented analysis based on Celonis tools. The results showed that the three features of the basic civil partnership contract, rial civil partnership contract type, and medium and small customer size have the greatest effect in reducing the processing time. In other words, the civil partnership contract type is up to 180 days, the Riyal civil partnership contract type is up to 148 days, and medium and small customers up to 141 days can reduce the period of payment for facilities. This means that the facilities with these features had a longer period for payment.

According to the three types of rough set analysis, common methods and process-oriented analysis for predicting the duration of payment of facilities, the features of the type of basic contract and the size of the customer and the type of contract were recognized as more important and effective.

In addition to the mentioned results, we can point out the weaker performance of the support vector machine method compared to other used techniques.

According to the research findings, it is predicted that overdue customers for fixed capital facilities with creative applications and completion and development are possible. In light of this, it is recommended that the payment of facilities to customers be made in stages and monitored based on the progress of the facility user's project. Additionally, to address the forecasted loan payment period of over one year for small and medium-sized customers or Riyal civil partnership contracts, the bank must eliminate any obstacles in the process by closely monitoring the steps until the final payment is made to the customer.

Given that the J48 and artificial neural network methods performed best, the strengths and weaknesses of these methods will be discussed below.

One of the strengths of the J48 algorithm in the present study was the high speed of learning and tree generation. This method also did not require any parameters. In addition, the generated model or tree, which represents the generated rules, was very understandable and interpretable. It was expected that this method would have higher accuracy.

One of the weaknesses of the artificial neural network method was the long learning time, especially for the 10-fold cross-validation mode. Another weakness of this method was that it was difficult to interpret. One of the strengths of this method in the present study, similar to other studies, was its suitability for real data, i.e. data related to fixed capital facilities.

One of the most important limitations of the present research in terms of payment time prediction is the small number of samples. In addition, in both types of analysis, there was a lack of access to informational features such as customer history in previous loans, credit information, and customer checkbook history.

To the extent possible, these limitations (e.g. sample size) have been addressed, especially for risk assessment. However, to predict the duration of loan payments, event logs are needed, which are related to the information system under study. In other words, in the past, information systems did not record and store events related to the stages of the process of providing

facilities. In this regard, credit information is not a limitation for analyzing the prediction of loan repayment duration. For credit risk analysis, if the customers are individuals, credit information can be very effective. Therefore, researchers need to pay attention to this limitation. However, since the present analysis is based on data from the information system of the bank under study, credit information was not available. On the other hand, given the main function of fixed capital facilities, which is to support the private sector, industries, and businesses active in production, in reality, credit characteristics are not examined much.

For future research, it is suggested to use other common supervised data mining methods for classification, such as KNN. Additionally, it is suggested to group the data using clustering methods at the beginning and then use classification methods.

REFERENCES

[1] 1.Khoshhaikel, et al., Identification of barriers to the development of electronic banking. Business Intelligence Management Studies, 2016. 4(16): p. 123-145..

[2] 2.Gholamian, Mozafari, and Azimeh, Predicting the value of new bank customers based on the R model. F. M using an improved decision tree to reduce the maximum memory required.

[3] 3.Mittal, A., et al. *A study on credit risk assessment in the banking sector using data mining techniques*. In *2018 International Conference on Advanced Computation and Telecommunication (ICACAT)*. 2018. IEEE.

[4] 4.Mandala, I.G.N.N., C.B. Nawangpalupi, and F.R. Praktikto, *Assessing credit risk: An application of data mining in a rural bank.* Procedia Economics and Finance, 2012. **4**: p. 406-412.

[5] 5.Thomas, L.C., *A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers.* International journal of forecasting, 2000. **16**(2): p. 149-172.

[6] 6.Sadatrasoul, S.M., et al., *Credit scoring in banks and financial institutions via data mining techniques: A literature review.* Journal of AI and Data Mining, 2013. **1**(2): p. 119-129.

[7] 7.Elah, T.F. and A.N. Majid, The role of the banking system's payment credits and the government's budget in the formation of gross domestic fixed capital. 2005.

[8] 8.Salehi and K. Katoli, choosing the optimal features in order to determine the credit risk of bank customers. Business Intelligence Management Studies, 2018. 6(22): p. 129-154.

[9] 9.BASHA, S.G., *Importance of Data Mining in Banking Sectors.* 2017.

[10] 10. Marek, W. and Z. Pawlak, *Rough sets and information systems.* Fundamenta Informaticae, 1984. **7**(1): p. 105-115.

[11] 11. Pawlak, Z., *Rough sets.* International journal of computer & information sciences, 1982. **11**: p. 341-356.

[12] 12. Papakyriakou, D. and I.S. Barbounakis, *Data mining methods: A review.* Int. J. Comput. Appl, 2022. **183**(48): p. 5-19.

[13] 13. Jackson, J., *Data mining; a conceptual overview*. Communications of the Association for Information Systems, 2002. **8**(1): p. 19.

[14] 14. Padhy, N., D.P. Mishra, and R. Panigrahi, *The survey of data mining applications and feature scope.* arXiv preprint arXiv:1211.5723, 2012.

[15] 15. Kesavaraj, G. and S. Sukumaran. *A study on classification techniques in data mining*. in the *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)*. 2013. IEEE.

[16] 16. Jadhav, S.D. and H. Channe, *Comparative study of K-NN, naive Bayes and decision tree classification techniques.* International Journal of Science and Research (IJSR), 2016. **5**(1): p. 1842-1845.

[17] 17. Gerhana, Y., et al. *Comparison of naive Bayes classifier and C4. 5 algorithms in predicting student study period*. In *Journal of Physics: Conference Series*. 2019. IOP Publishing.

[18] 18. Vijayarani, S. and S. Dhayanand, *Liver disease prediction using SVM and Naïve Bayes algorithms.* International Journal of Science, Engineering and Technology Research (IJSETR), 2015. **4**(4): p. 816-820.

[19] 19. Muralidharan, V. and V. Sugumaran, *A comparative study of Naïve Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis.* Applied Soft Computing, 2012. **12**(8): p. 2023-2029.

[20] 20. Abiodun, O.I., et al., *State-of-the-art in artificial neural network applications: A survey.* Heliyon, 2018. **4**(11).

[21] 21. Fletcher, T., *Support vector machines explained.* Tutorial paper, 2009. **1118**: p. 1-19.

[22] 22. Zandi, S., et al., *Attention-based Dynamic Multilayer Graph Neural Networks for Loan Default Prediction.* arXiv preprint arXiv:2402.00299, 2024.

[23] 23. Chen, B., W. Jin, and H. Lu, *Using a genetic backpropagation neural network model for credit risk assessment in the micro, small and medium-sized enterprises.* Heliyon, 2024. **10**(14).

[24] 24. Montevechi, A.A., et al., *Advancing credit risk modeling with Machine Learning: A comprehensive review of the state-of-the-art.* Engineering Applications of Artificial Intelligence, 2024. **137**: p. 109082.

[25] 25. Zhang, X. and L. Yu, *Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and learning methods.* Expert Systems with Applications, 2024. **237**: p. 121484.

[26] 26. Addy, W.A., et al., *Predictive analytics in credit risk management for banks: A comprehensive review.* GSC Advanced Research and Reviews, 2024. **18**(2): p. 434-449.

[27] 27. Chandrasiri, T.D. and S.C. Premaratne *Enhancing Credit Risk Analysis of SME Loans by Using Data Mining Techniques.* 2023.

[28] 28. Jumaa, M., M. Saqib, and A. Attar, *Improving credit risk assessment through deep learning-based consumer loan default prediction model.* International Journal of Finance & Banking Studies (2147-4486), 2023. **12**(1): p. 85-92.

[29] 29. Chen, Q., *Interpretable Data Mining Approaches to Predict Term Deposits Subscriptions.* BCP Business & Management, 2023. **44**: p. 345-350.

[30] 30. Anand, M., A. Velu, and P. Whig, *Prediction of loan behaviour with machine learning models for secure banking.* Journal of Computer Science and Engineering (JCSE), 2022. **3**(1): p. 1-13.

[31] 31. Munoz, J., et al., *Deep learning based bi-level approach for proactive loan prospecting.* Expert Systems with Applications, 2021. **185**: p. 115607.

[32] 32. Desta, A.W. and J.S. Nixon, *Data mining application in predicting bank loan defaulters.* International Journal of Innovative Technology and Exploring Engineering, 2020. **9**(4).

[33] 33. Wang, J., et al., *Rough set and scatter search metaheuristic based feature selection for credit scoring.* Expert Systems with Applications, 2012. **39**(6): p. 6123-6128.

[34] 34. Crone, S.F. and S. Finlay, *Instance sampling in credit scoring: An empirical study of sample size and balancing.* International Journal of Forecasting, 2012. **28**(1): p. 224-238.

[35] 35. Koutanaei, F.N., H. Sajedi, and M. Khanbabaei, *A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring.* Journal of Retailing and Consumer Services, 2015. **27**: p. 11-23.

[36] 36. Gulsoy, N. and S. Kulluk, *A data mining application in credit scoring processes of small and medium enterprises commercial corporate customers.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2019. **9**(3): p. e1299.

[37] 37. Jisha, M. and D.V. Kumar, *A CASE STUDY ON DATA MINING APPLICATIONS ON BANKING SECTOR.* 2018.

[38] 38. Hamid, A.J. and T.M. Ahmed, *Developing prediction model of loan risk in banks using data mining.* Machine Learning and Applications: An International Journal, 2016. **3**(1): p. 1-9.

[39] 39. Hooman, A., et al., *Statistical and data mining methods in credit scoring.* The Journal of Developing Areas, 2016. **50**(5): p. 371-381.

[40] 40. Eskandari, J. and Rouhi, credit risk management of bank customers using improved decision vector machine method with genetic algorithm with data mining approach. Asset Management and Financing, 2017. 5(4): p. 17-32.