REVISTA MEXICANA DE ECONOMÍA Y FINANZAS Nueva Época REMEF (THE MEXICAN JOURNAL OF ECONOMICS AND FINANCE)

Revista Mexicana de Economía y Finanzas, Nueva Época

Volumen 20 Número 1, Enero – Marzo 2025, pp. 1-18, e1205

DOI: https://doi.org/10.21919/remef.v20i1.1205

(Received: December 12, 2023, Accepted: March 18, 2024, Published: December 5, 2024)

imef fundación de investigación

A Bayesian Network Model to Evaluate the Credit Risk of Mexican Microfinance Institutions in 2023

Alondra M. Gress-Guerrero - Universidad Autónoma del Estado de Hidalgo, México
Jedidia Hernández-Vargas - Universidad Autónoma del Estado de Hidalgo, México
José F. Martínez-Sánchez[1] - Universidad Autónoma del Estado de Hidalgo, México
Francisco J. Martínez-Farías - Hospital Juárez de México, México

Abstract

Assessing credit risk is crucial for financial institutions to make timely and accurate decisions. This research proposes a model for microfinance institutions to estimate credit risk in simplified single-client scenarios. The model is based on a Bayesian Network algorithm and uses a historical database to demonstrate the forecast risk potential. The database includes variables such as *age, income, credit history, home ownership*, and the final state of the loan, which can be *pay*, *default*, or *breach*. By establishing a relationship between variables through an inferential network and joint probability tables, the research explores three scenarios to obtain risk probability distributions based on different age and income ranges. The inference probability is obtained via a Bayesian network, where the interrelation between variables is structured in a specific topology. We use the causality assumption to estimate the probability of default or credit risk, which is closer to the reality of credit institutions. Therefore, it is a powerful tool for risk analysis agencies to make informed decisions in credit evaluation.

JEL Classification: G21, G28, G32, G33, K20, C610.

Keywords: Bayesian network, Junction tree, Credit risk, Nodes probability, Risk assessing.

Un Modelo de Red Bayesiana para Evaluar el Riesgo Crediticio de las Instituciones de Microfinanzas Mexicanas en 2023

Resumen

La evaluación del riesgo de crédito es esencial para las instituciones financieras, ya que les permite tomar la decisión correcta en el momento adecuado. Esta investigación presenta un modelo basado en un algoritmo de Red Bayesiana para estimar el riesgo de crédito en escenarios simplificados de perfil de cliente único para instituciones de microfinanzas. Demostramos el potencial de previsión del riesgo aplicando los datos probabilísticos de una base de datos histórica estándar. Nuestra aplicación se basa en una base de datos bancaria con resultados históricos de préstamos. Las variables de la base de datos son la edad, los ingresos, el historial crediticio, la propiedad de la vivienda y el estado final del préstamo, en el que existen tres posibilidades (pago, impago, incumplimiento). Establecemos una relación entre las variables mediante una red inferencial y tablas de probabilidades conjuntas. Además, exploramos tres escenarios, considerando diferentes rangos de edad e ingresos, para obtener distribuciones de probabilidad de riesgo. La probabilidad inferencial se obtiene mediante una red bayesiana donde la interrelación entre variables se estructura en una topología específica. Este trabajo, a diferencia de otros, utiliza el supuesto de causalidad para estimar la probabilidad de impago o riesgo de crédito. Esto se aproxima más a la realidad de las entidades de crédito y, por tanto, es una potente herramienta para la toma de decisiones en la evaluación crediticia por parte de las agencias de análisis de riesgos.

Clasificación JEL: G21, G28, G32, G33, K20, C610.

Palabras clave: Red bayesiana, Árbol de unión, Riesgo de crédito, Probabilidad de nodos, Evaluación de riesgos.

[1] Corresponding author: jmartinez@uaeh.edu.mx , Carretera Apan-Calpulalpan s/n, Colonia, 43920 Chimalpa Tlalayote, Hgo. Teléfono: 771 717 2000 ext. 5800

* No source of funding for research development

1. Introduction

Effective management and mitigation of credit risk is essential in the financial industry. Credit risk is a measure that depends on various variables, including historical information and conditional relations between the structure of stochastic variables (Vargas-Sánchez and Mostajo-Castelú, 2014; Assef et al., 2019). Credit risk methodologies are critical in identifying and addressing potential risks, enabling lenders and investors to improve their risk-adjusted returns and maintain the safety and soundness of their businesses. Risk assets have numerous applications, such as medicine, industrial security, and natural disasters. Credit risk assets are among the most critical issues in the finance industry since a healthy economy depends strongly on them. Although there are standard methodologies such as the BIS 1998 or CreditMetrics proposed by JP Morgan (Risk Metrics Group Inc, 2007), each financial institution is responsible for adequately assessing credit risk. The standard models are changing as the precision of the information stored in significant variables and the relationships between them increase.

The rise of digital banking and the advent of big data have increased lending rates and financing opportunities (Transparency, 2021). Bayesian methods, such as inference models, have been used to assess credit risk and the probability of default (Huang and Yu, 2010). However, the use of massive data analysis algorithms has also opened up new possibilities for risk management (Wang *et al*., 2020). The administration of financial resources has traditionally been in the hands of banks. However, with financial globalization, it has become increasingly challenging to supervise financial institutions and their security, impacting their risk management abilities. This situation is exemplified in a study conducted by Deloitte & Touche, as discussed by Prieto (2000), Roisenvit and Zárate (2006), and Murillo-López and Venegas-Martínez (2011).

Credit institutions in Latin America were established in the mid-19th century (Marichal, 2008). 1990, Mexico published the Law of Credit Institutions (SHCP, 1990) in the Federal Gazette. This law aims to set the framework for credit institutions' commercial activities, limit and regulate banking and credit services, and ensure financial institutions' healthy and balanced development. In summary, this legal instrument regulates the activities and operations that credit institutions can undertake.

Bayesian networks (BN) are powerful machine learning and artificial intelligence tools (Hernán and James, 2023). These networks model probabilistic relationships between variables and causality connections. A Bayesian network is a directed acyclic graph representing these probabilistic relationships in a graphical form. In the graph, each node represents a different variable, while the edges represent the probabilistic dependencies between them. The conditional probability distribution for each node is defined in terms of the probabilities of its parents in the graph. One of the most critical features of Bayesian networks is that they can be used to perform probabilistic inference. Bayesian inference algorithm can compute the probabilities of other variables in the graph, which makes Bayesian networks valuable for prediction, diagnosis, and decision-making tasks. They have been applied in various fields, including medical diagnosis, risk analysis, image recognition, and natural language processing. Bayesian networks are beneficial in situations where there is uncertainty or incomplete information. They can incorporate prior knowledge and update

probabilities based on new evidence. This makes them a valuable tool in situations requiring accurate predictions.

Bayesian Networks are a new computational tool that uses statistical inference algorithms. They have been applied in medicine, chemistry, and Money Laundering Risk (Sánchez *et al*., 2020). The research uses the Bayesian Network model to measure credit risk in financial institutions using specific examples of stochastic variable distributions based on a historical database. In BNs, each node represents a random variable with an associated probability distribution. The structure of these networks provides information about the conditional dependency and independence relationships between variables, simplifying the joint probability representation. This representation is the product of the conditional probability functions of each of the variables (Sucar, 2021, Leong, 2015).

We aim to simplify decision-making by providing a reference framework. Although this model is straightforward, it can be expanded by introducing additional relevant variables and representing them at different nodes in the network. The investigation of this model is divided into six sections, starting with an introduction. The second section outlines the elements, definitions, and models used to measure credit risk. The third section explains how to use Bayesian Networks for measuring credit risk. The fourth and fifth sections cover the database, the analysis of the model, its structure, and its results. Finally, the sixth section concludes the investigation. Overall, the BN Model is a valuable tool for financial institutions to make informed decisions about granting credit to applicants.

1.1 Credit risk

Credit risk is the intrinsic probability associated with a breach of contract by the borrower (Dimitris, 2000).Credit risk in financial institutions is assumed to be assigned in the markets; therefore, this price is included in the market purchase price for a contracted payment. The part of the price that belongs to the credit risk is the credit spread (Chatterjee, 2016). The credit risk measurement model uses general economic conditions and company status. Those let us generate a result known as a credit spread—a great diversity of models to assess credit risk. The most common are based on historical data and probability decision distribution of the data fields. Credit risk is a critical consideration for lenders, investors, and other stakeholders in the financial industry. Some of the key technical features of credit risk methodologies include:

1. Credit scoring: This involves evaluating the creditworthiness of a borrower based on a variety of factors, such as credit history, income, and debt to income ratio. Credit scoring models assign numerical scores to borrowers, which are used to determine the likelihood of default and to set the terms of the loan or investment.
2. Risk assessment: This involves evaluating the overall credit risk exposure of a lender or investor. Risk assessment frameworks consider factors such as the loan portfolio's size and diversity, individual borrowers' creditworthiness, and overall economic and market conditions.
3. Risk management involves developing strategies and controls to mitigate or avoid credit risk. Risk management strategies may include diversifying the loan portfolio, setting appropriate

loan terms and interest rates, and purchasing credit insurance or other risk management tools.

Some typical applications of credit risk methodologies include:

1. Consumer lending includes credit risk management in mortgages, auto, and personal loans.
2. Commercial lending: This includes credit risk management in business loans and lines of credit.
3. Investment management includes credit risk management for investors holding fixed-income securities, such as bonds.

Overall, credit risk methodologies are critical for managing and mitigating credit risk in the financial industry.By identifying and addressing potential credit risks, lenders and investors can improve their risk-adjusted returns and maintain the safety and soundness of their businesses.

1.2 Credit risk measurement methodologies

Credit risk measurement models are utilized to approximate the likelihood of Default (PD), Stress Testing (ST), and Loss given default (LGD) of individual borrowers or groups of loans. Various models are used, such as credit scoring, structural, reduced-form, machine learning, and portfolio models. Credit rating agencies evaluate a borrower's creditworthiness and assign a rating based on various factors, including the borrower's financial strength, operating environment, and overall creditworthiness. These models are typically employed with credit risk measurement frameworks such as the Basel Accords to manage credit risk in financial institutions. Various formalisms are utilized to evaluate loan risk, and below, we have mentioned some of the most used ones.

1.2.1 Probability of Default models

PD models are statistical models used to estimate the likelihood of a borrower defaulting on their debt obligations. These models are widely used in credit risk management by financial institutions, credit rating agencies, and regulatory bodies to measure the creditworthiness of borrowers and assess the overall risk of their credit portfolio. Those models estimate the likelihood of a borrower defaulting on their debt within a given timeframe. PD models typically use historical data to estimate the probability of default and can be calibrated to specific borrower or loan characteristics (Dar et al., 2019, Alkhawaldeh *et al*., 2019, Burova *et al*., 2021).

1.2.2 Loss Given Default models

LGD models estimate the loss a lender or investor would incur if a borrower failed to meet their debt obligations. Typically expressed as a percentage of the outstanding loan balance or security face value, LGD models can be divided into statistical and judgmental categories. Statistical models use historical data and mathematical algorithms to estimate the expected LGD, while judgment models

rely on expert judgment and qualitative information. LGD models consider various factors such as collateral, recovery rates, and credit enhancements to estimate the potential loss to a lender in case of default by a borrower (Kovalova and Cug, 2015; Bellini, 2019).

1.2.3 Stress Testing

Stress testing models evaluate the resilience of financial institutions, portfolios, or individual financial products under adverse economic scenarios. The primary aim of stress testing is to assess the possible impact of severe and unexpected events on the performance and stability of financial systems. Stress testing models can be categorized into two broad groups: macroeconomic and microeconomic. Macroeconomic stress testing models evaluate the impact of systemic risks on the financial system, such as recession, inflation, or financial crises. Microeconomic stress testing models, on the other hand, assess the impact of idiosyncratic risks on individual portfolios or financial products. Stress testing assesses credit risk by simulating adverse scenarios that affect borrowers' ability to repay their debt. These scenarios might include changes in economic conditions, interest rates, or other macroeconomic factors (Varotto, 2012; Buncic and Melecky, 2013).

1.2.4 Big Data Analytics

1. Artificial Intelligence (AI) methods and Big Data analytics are increasingly used in credit risk management to improve the accuracy and efficiency of credit risk models. These methods provide insights into borrower behavior and credit risk that are unavailable through traditional statistical methods. AI methods include machine learning algorithms such as neural networks, decision trees, and random forests. These algorithms learn from historical data to identify patterns and relationships between borrower characteristics and credit risk. They also adapt to new data and changing market conditions, improving the accuracy of credit risk models over time.
2. Big Data analytics uses significant, diverse, unstructured data sources, such as social media, web browsing behavior, and mobile phone usage patterns, to supplement traditional credit data. This approach provides a more comprehensive view of a borrower's creditworthiness and behavior, allowing lenders to make more informed lending decisions. One of the significant advantages of AI methods and Big Data analytics is their ability to identify non-linear relationships and interactions between borrower characteristics and credit risk. For example, machine learning algorithms can identify complex interactions between credit scores, income levels, and loan characteristics that are difficult to detect using traditional statistical methods.
3. AI methods and Big Data analytics can also help lenders proactively identify and manage credit risk. They can identify early warning signals of potential defaults, enabling lenders to take preventative action before losses occur. They can also help lenders identify opportunities for cross-selling and up-selling to existing customers, improving customer retention and profitability. Additionally, these methods are increasingly used to improve credit risk models' accuracy and efficiency. In summary, the advantages of AI and Big Data include their ability to identify complex interactions between borrower characteristics and

credit risk, supplement traditional credit data with diverse unstructured data sources, and identify early warning signals of potential defaults.

1.3 Bayesian Network applied to credit risk

Bayesian networks (BN) are a probabilistic graphical model representing complex relationships between variables in a system. These networks consist of stochastic variables with a known probability distribution and their inter-dependency relationships, which are represented graphically. The BN model is based on inference, which involves estimating the posterior probability of unknown variables based on the known variables. BN models have various applications, including classification, prediction, and diagnosis. They are commonly used in credit risk assessment to model the relationships between borrower characteristics, loan characteristics, and the probability of default. By incorporating subjective and objective information, BN models provide a unified framework for lenders to combine expert knowledge with data-driven modeling, resulting in more accurate estimates of the probability of default.

Credit risk Bayesian network models represent the relationships between borrower characteristics, loan characteristics, and the probability of default using a network of nodes and edges. Each node represents a variable, and each edge represents a conditional dependency between two variables. This model uses Bayes' theorem to update the probability of default as new information becomes available. If a borrower's credit score changes, the model updates the probability of default based on the new information. By examining the conditional dependencies between variables, the model can identify which variables have the most significant impact on the probability of default. Bayesian network models are also advantageous in handling missing data. In credit risk, some borrower or loan characteristics may be missing. Using the available data, Bayesian network models can incorporate the missing values into the model. The model can be applied to credit risk assessment in various contexts, including consumer lending, small business lending, and corporate lending. Compared to traditional methods like logistic regression, Bayesian network models have been shown to improve the accuracy of credit risk assessment.

1.3.1 Bayesian Inference

Bayesian networks (BN) use probabilistic meaning to propagate the effects of evidence through the network and determine the posterior probability of variables. This means that BN estimates the probability distribution of unknown variables based on prior information by discarding some information. It adjusts the probability measures using available information to obtain a new analysis that allows us to predict behavior or uncertainty (Sucar, 2021). To apply probabilistic propagation in Bayesian networks, it is essential to consider which nodes represent random variables and which line connections (arcs) represent dependency relationships. In BN, preliminary information is structured in an inferential mode through unidirectional and non-cyclically connected nodes, which form an inferential network. The structured inferential relation between the BN has an advantage over pure statistical models such as decision trees (Enright et al., 2013; Martínez-Sanchez et al., 2016).

Bayesian inference is the theoretical foundation of BN. It is a statistical approach used to make inferences about the parameters of a model based on observed data. The central concept of Bayesian inference is to apply Bayes' theorem, which states that the posterior probability of a hypothesis (i.e., the probability of the hypothesis given the data) is proportional to the likelihood of the data given the hypothesis multiplied by the prior probability of the hypothesis.

Here are some fundamental concepts and steps involved in Bayesian inference:

1. *Prior probability*: Before observing any data, we assign a prior probability distribution to the model's parameters. This distribution represents our prior beliefs or knowledge about the parameters.
2. *Likelihood*: Given the model's parameters, the likelihood function describes how likely the observed data are.
3. *Posterior probability*: Using Bayes' theorem, we can compute the posterior probability distribution of the parameters given the observed data. This distribution incorporates our prior beliefs and the information contained in the data.
4. *Posterior inference*: Once we have the posterior distribution, we can compute various quantities of interest, such as the posterior mean or median, credible intervals, or posterior probabilities of hypotheses.
5. *Model selection*: Bayesian inference allows us to compare different models by computing the posterior probability of each model given the data.
6. *Updating beliefs*: Bayesian inference is a dynamic process. As new data becomes available, we can update our prior beliefs and compute the posterior probability of the parameters given the new data.

Overall, Bayesian inference provides a flexible and robust framework for modeling complex phenomena and making predictions based on data (Gelman et al., 2004, Turner and Van Zandt, 2018). Here we present an introduction to basic mathematical principles.

- x denotes unobservable vector quantities of population parameters of interest.
- y denotes the observed data.
- $\tilde{y}$ represents the unknown but potentially observable quantity. then the sampling distribution $p(y|x)$ could be expressed as

$$P(x, y) = p(x)p(y|x),$$

using the property of conditional probability of Bayes rule,

$$p(y|x) = \frac{p(x,y)}{p(y)} = \frac{p(x)p(y|x|)}{p(y)},$$

where $p(y)$ is the sum over the possible values of x

$$p(y) = \int p(x)p(y|x)dx$$

the distribution probability of $\tilde{y}$ is given by

$$p(\tilde{y}, y) = \int p(\tilde{y}, x\,|y)dx = \int p(\tilde{y}\,|x, y)p(x|y)dx = \int p(\tilde{y}|x)p(x, y)dx$$

This method obtains the posterior predictive probability as an average of conditional prediction over the posterior distributions of x (Gelman et al., 2004).
We use junction tree algorithm for Bayesian inference. For more detail see (Martínez-Sanchez *et al*., 2016).

2. Problem Statement

Our approach involves evaluating their credit risk based on specific fields in financial institutions. We are leveraging the bank database with historical data to calibrate the likelihood of customers paying back loans or defaulting. We analyze the database structure and apply our proposal of the Bayesian network model to infer probabilistic scenarios.

2.1 Probability distribution nodes

In Bayesian networks, each node represents probability distribution data. We use a database bank from Martínez-Farías (2023) to construct these distributions. This database allows us to build probability distribution nodes whose interactions conform to the Bayesian network. The data contained in the database pertains to impersonal information about 32,580 clients.

1. *Age*, concerning the client's age, all values are in the interval [20,70].
2. *Income* that represents the annual income in dollars, and the values are in [0, 300×10^3].
3. *Homeownership* is a categorical variable, represented by the values {*rent, mortgage, owner*}; we assign numerical values {1, 2, 3} respectively.
4. *Employ length*, Represent the years in their last employment. The interval is in [0,30].
5. *Loan grade* is referred to as intrinsic credit risk measured and is a categorical variable {*D, C, B, A*}, where *D* represents the maximum and *A* the minimum Risk; we convert it to numerical values {1, 2, 3, 4}.
6. *Loan amount* concerns the total amount requested in dollars; the interval is [0,100×10^3].
7. *Loan rate* represents the annual rate of the mount and is a perceptual quantity in the interval [0,100].
8. *Credit hist* is measured in years and represents the number of years using credit instruments.
9. *Loan percent* represents the percentage that the loan represents concerning the income.

Our research concerns client profile scenarios to measure the credit risk probability in a historical bank database. For example, Figure 1 shows the pay, default, and breach probability as a function of age. We proposed a Bayesian Network built on the inferential relationship between the fields of age, Income, Loan Grade, Joint A-I (age-Income), and Homeowner. The results are divided

into three profiles, identifying each with the pessimistic, neutral, and optimistic cases. A credit risk profile can be assigned to a new applicant, knowing the abovementioned fields. The credit risk asset models based on Bayesian Networks are a powerful computational tool based on statistical inference techniques; these algorithms and the structural models between variables are still at the research frontiers. However, in a later work, comparing the results of other models based on similar information will be necessary. Another exciting aspect is that the presented probability tables can be enriched with information from the credit market, enhancing our work's applicability.

Fig. 2 shows the pay distribution frequency for all database fields. Is it possible to observe the sensibility of each variable concerning *pay* probability (P_{pay}). Graph (a) Shows the probability of growth concerning age, until a maximum of around 60, to decay, possibly related to the mortality rate. (b) The probability increases quickly concerning the *income*, stabilizing for major *income* mounts. (c) Categorical variable *home ownership* with states {*rent* (A), *mortgage* (B), *own* (c)} is converted into numerical values respectively {0, 1, 2}, is appreciated that homeowners pay with height probability. (d) A person's time working in the last job is directly proportional to the probability of paying their loan. (e) The intrinsic risk measured by the categorical variable *loan grade* is converted in numerical values {0} for high risk, {1} for medium height risk, {2} medium-low risk, and {3} for low risk. It is appreciated that *pay* probability is directly proportional to the low risk. (f) The *loan amount* seems unrelated to the *pay* probability, except for significant loans, where the probability shows a more significant variation. (g) The *loan rate* does not have a notorious tendency concerning the *pay* probability, although it is shown that the pay probability variability increments with the *loan rate*. (h) *Credit hist pay* probability is when a person has used credit instruments, showing an increment with the time length. (i) *pay* probability of the *loan percentage* decreases when the loan represents a high percentage of the *income*. After analyzing the available data, the most relevant variables in terms of probability variations were selected to build the nodes of the Bayesian network. This approach ensures that the network is accurate and can provide valuable insights into the relationships between the variables.

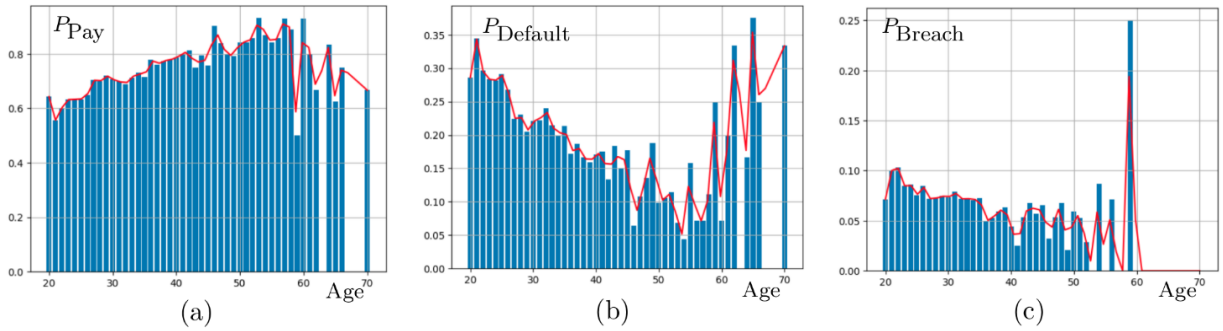

Figure 1. Evolution of normalized frequency distribution [PPay + PDefault + PBreach = 1] for age in the cases.

(a) Pay. (b) Default. (c) Breach. Continuous lines represent the linear interpolation of probability.

Figure 2 shows a continuous line representing the linear interpolation of the data. It is important to note that we do not have all the data probability for every field, but that does not necessarily mean the probability is zero. We have proposed a simple approximation to help solve the problem. We also tested using cubic splines but found that this technique created artificial peaks in

the data distribution, and this approach needed to be revised. The probability nodes calculation was performed using these splines as a near-real probability distribution.

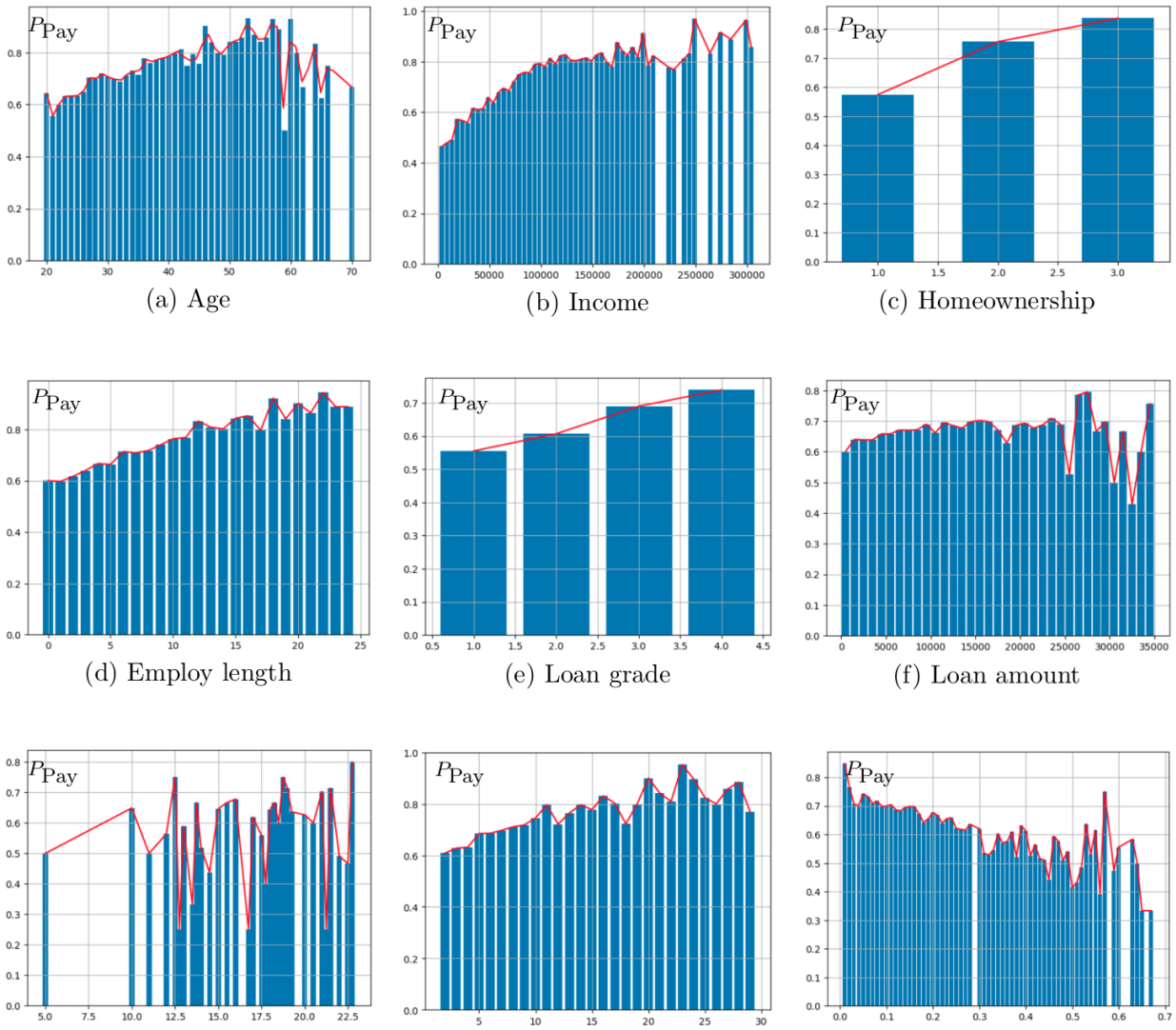

Figure 2. Pay distribution frequencies.

The data is being modeled using linear interpolation, with the continuous line representing the estimated values not included in the database. The bars in the graph represent the actual database values. The probabilistic distribution described by the interpolation was used in the calculations. (a) Age; represents the pay probability for ages in the range [20,70]. (b) Pay probability as a function of income. (c) Pay probability for home; rent (1), mortgage (2), and (3) owner. (d) Pay probability for employment length measured in years. (e) Pay probability of Loan grade, i.e., the qualification of the previous loans. (f) Pay probability as a function of the Loan amount. (g) Pay probability as a function of the loan rate. (h) Pay probability as a function of credit history measured in years. (i) Pay probability as a function of loan percent concerning income.

2.2 Bayesian Network Implementation algorithm to risk assessment

We calculate the likelihood of a credit applicant paying back their lender based on various factors that define them. To determine the probability distribution of the risk of default, we use a Bayesian network that provides us with the posterior probability. We use the "pgmpy" Bayesian network algorithm developed by Ankan et al. (Ankan and Panda, 2015) to simulate the different scenarios for granting a credit loan based on specific profile information. Using historical data, we can obtain information to describe the probability tables of the nodes. Bayesian inference helps us understand the risk's probability distribution, allowing us to decide whether to approve or reject the credit loan.

In Figure 3, we can see a graphical representation of the BN Model. It consists of two significant variables or nodes with statistical relevance, according to Cole (1998). These variables are (1) Age, which is related to the age of the credit applicant, and (2) Income, which parameterizes the source of income of the loan applicant. The information these nodes provide will help us decide on the granting of credit.

The Bayesian network model we have developed considers *age*, while *income* is the independent variable that defines the junction probability table (Joint A-I (c)). The *income* node also influences the loan intrinsic risk type. The *Homeowner* node is the probability distribution of *pay*, default, and breach for the case of *homeowners*, *mortgage*, and *rent*, as shown in Fig. 4. We propose this because these variables have a significant variation in *pay* probability, and we believe that this relationship reflects their interaction. There are other ways to structure the BN, but we have started with linear interpolated historical data to demonstrate an easy way to define BN structures.

The node *age* is related to the node *Loan grade(g)*, which refers to the credit history of the credit applicant (for example, credit history, bureau points, overdue balance) also the bank's internal evaluation of the borrower's ability to pay. Considering the results of section 2.1 and our credit market experience, we will use the following probability distributions in the variables. Figure 4 shows the conditional probability tables of the variables assigned to the network: a*ge* (a), *Income* (i), *Loan grade* (g), and the *Homeowner* (h).

2.2.1 Nodes

This section describes the node probability distribution. The node *age*(a) has the following probability distribution (see Table 1)

Table 1. Age node probability distributions for indicated intervals.

34-80 (a_0)	**28-33 (a_1)**	**20-27 (a_2)**
0.37	0.33	0.30

The probability distribution table of the *age* (a) node allows us to see that our client is 34 to 80 years is 37%, while the probability that he is 28-33 is 33%, and the probability that he has 20-27 is 30%.

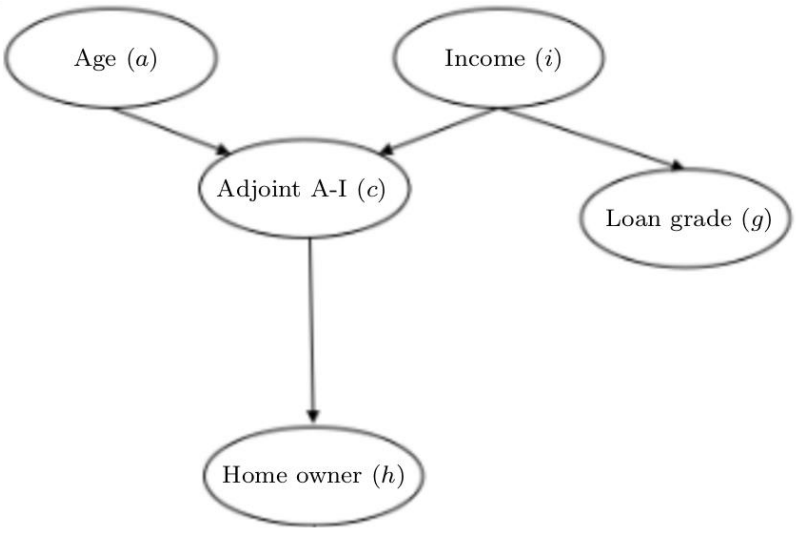

Figure 3. Graphic topology representation of nodes as probabilistic variables model, and the arrows represent their dependency relationships.

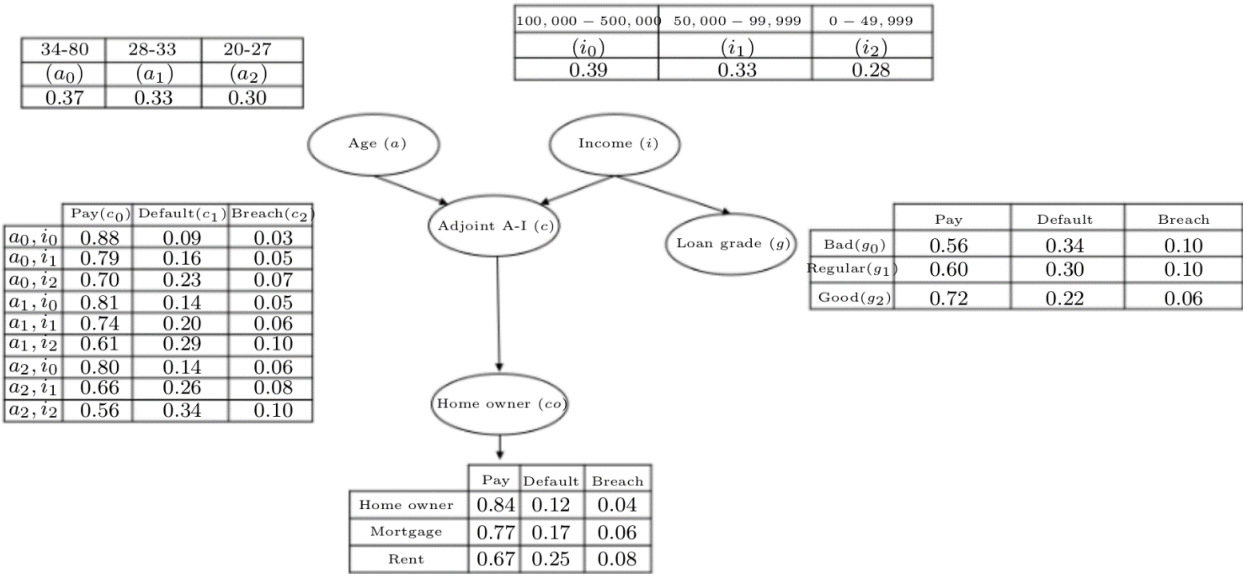

Figure 4. Distribution tables assigned to each node for Bayesian Network

The *Loan Grade* node is divided into three categories based on the borrower's payment history. The first category, considered harmful, has values of 0 and 1. According to the database, 56% of the people with this history paid back their debt, 10% were late, and 34% defaulted. The second category is considered regular, with values of 1 and 2. Our database states that 60% of the borrowers with these values settled their debt, 10% were late on one or more payments, and 30% committed fraud. Finally, the third category is considered good and has values between 2 and 4. The number of customers in this category who paid back their debt was 72%; 6% were late, and 22% failed to yield with the pay.

In Figure 4, the Joint A-I (c) node distribution table displays the probability distribution of our client's payment behavior for a credit application. It covers the likelihood of the client paying on time, being late with the payment, or defaulting on the credit. The table includes scenarios based on the client's age and annual income range. For instance, the first row of the table shows the probabilities for clients between 34 and 80 years old with an annual income of $100,000 to $500,000. In this case, there is an 88% chance that the client will pay on time, a 0.03% chance of being late with the payment, and a 0.09% chance of committing fraud.

Similarly, the second row represents clients between 34 and 80 years old with an annual income of $50,000 to $99,999. In this scenario, there is a 79% probability that the client will pay on time, a 5% probability of delayed payment, and a 16% chance of fraud. The table also includes scenarios for younger clients with lower incomes. For example, if our client is between 20 and 28 years old and has a salary between 0 and 50,000 dollars, there is a probability of 70% that they will make the loan payment, 7% that they will be late in the same payment, and a probability of 23% of committing fraud. On the other hand, if our client is between the ages of 28 and 34 and has a salary between 50,000 and 99,999 dollars, there is an 81% chance that the client will pay on time, a 5% chance of delayed payment, and a 14% chance of non-compliance.

It is crucial to consider these probabilities when assessing the risk of granting a loan to a potential client. The probability of a client making their loan payment varies depending on their age and salary range, so it is essential to analyze each scenario carefully before making any decisions.

Table 2. Distribution probability of loan grade is separated into categories.

100000-500000 (i_0)	50000-99999 (i_1)	0-49999(i_2)
0.39	0.33	0.28

Table 3. Distribution probability of loan grade separated into categories.

	Pay	Default	Breach
Bad 1 (g_0)	0.56	0.34	0.10
Regular 2 (g_1)	0.60	0.30	0.10
Good 3, 4 (g_2)	0.72	0.22	0.06

Table 4. The probability distributions Joint A-I (c_k) node is obtained with the joint probability of age (a_k) and income (i_k).

Scenario	Pay (c_0)	Defaulter (c_1)	Breach (c_2)
a_0, i_0	0.88	0.09	0.03
a_0, i_1	0.79	0.16	0.05
a_0, i_2	0.70	0.23	0.07
a_1, i_0	0.81	0.14	0.05
a_1, i_1	0.74	0.20	0.06
a_1, i_2	0.61	0.29	0.10
a_2, i_0	0.80	0.14	0.06
a_2, i_1	0.66	0.26	0.08
a_2, i_2	0.56	0.34	0.10

Table 5. Probability distributions of Node *Homeowner* (h), the first row corresponds with the probability distribution of graph (c) in Fig. 2. Raws two and three represent the probability distribution of Mortgage and Rent conditions.

	Pay	Default	Breach
Homeowner 3	0.84	0.12	0.04
Mortgage 2	0.77	0.17	0.06
Rent 1	0.67	0.25	0.08

The *Homeowner* node (CO) probability distribution table has the following information regarding the housing status of credit applicants. This section is divided into three categories: Homeowner, Mortgage, and Rent. Homeowner has a value of 3, indicating people who own their homes. According to the database, 84% of homeowners paid off their credit, 4% made late payments, and 12% defaulted on their credit. The mortgage category has a value of 2 and comprises people with a mortgage. The data shows that 77% of people with a mortgage paid their credit, 6% were delinquent, and 17% committed fraud. Lastly, the Rent category includes people who rent their houses with a value of 1. According to our historical information, 67% settled the loan, 8% made late payments, and 25% committed fraud.

3. Scenarios and Results

By utilizing the Bayesian Inference implementation discussed in subsection 2.2, it becomes possible to estimate the posterior probability of a scenario based on the nodes or probability distribution, as depicted in Figure 4. Our team has simplified the scenarios into three categories for each result, but it is important to note that creating the most complex cases is also possible. It is worth mentioning that in real life, the complexity of the scenery depends on each solicitor's profile.

3.1 Scenario 1

That scenario corresponds with someone with age in the interval [31,80] and an Income of [100000,500000]. The rest of the parameters remain fixed and are the same for all scenarios. Then Credit probability distribution P(Age ∈ [31, 80] & Income ∈ [100, 500] × 10^3) has as result the table 6.

Table 6. Results of scenrio 1: The probability distribution for the age in the interval [31,80] and Income [100000,500000] is shown. Both intervals correspond with a high probability of paying.

C	ϕ(C)
C(Pay)	0.8800
C(Default)	0.0900
C(Breach)	0.0300

Table 6 provides the following interpretation: If the credit is approved under an optimistic scenario, there is an 88% chance that the applicant will make timely payments and be a good payer. The applicant must ensure they pay on time and meet the requirements. In case of any delay in payments, the value of the payments will be reduced to 3%. Additionally, there is a 9% chance that the client may fail to pay the loan, resulting in a high-risk loan.

3.2 Scenario 2

In this case, the probability distributions were assigned as follows P (Age ∈ [28, 33] & Income ∈ [50, 99.99] × 10^3). The BN algorithm determines an inferred probability as shown in Table 7.

Table 7. Results of scenario 2: In this case, Age is into the interval [28,33], and Income is into is into [50, 99.99] x 10^3. That scenario is less favorable in terms of probability than the previous example.

C	ϕ(C)
C(Pay)	0.7400
C(Default)	0.2000
C(Breach)	0.0600

According to Table 7, there is a 74% probability that the client will liquidate the credit. Furthermore, there is a 6% chance that the applicant may be late with the payment and a 20% chance that the client may default on the credit payment. These findings indicate a certain level of risk involved in this credit arrangement. Therefore, evaluating and implementing strategies to manage this risk is advisable.

3.3 Scenario 3

The probability distributions *P* (Age ∈ [20, 27] & Income ∈ [0, 49.999] × 10^3) can be seen in Figure 4. Table 8 presents the inferred probabilistic distribution for this case.

Table 8. Results of scenario 3: Corresponds to some scenario with someone very young (Age [20, 27]) and a small salary (Income ϵ [0, 49.999] x 10^3) concerning the two previous examples. In that sense, the probability of Default or Breach is more significant regarding previous scenarios.

C	ϕ(C)
C(Pay)	0.5600
C(Default)	0.3400
C(Breach)	0.1000

Based on the data presented in Table 8, it is evident that we need to exercise caution while granting credit. The figures indicate a relatively high chance (10%) of a credit applicant being late and failing to make timely payments. However, there is a slightly better chance (56%) that the applicant will complete their payments as agreed. Nonetheless, a worrying probability (34%) still exists that the applicant will breach the agreement and fail to pay the credit. These statistics

underscore the importance of careful consideration when deciding whether to grant credit to an applicant to avoid potential financial losses.

When assessing whether to approve a loan for a customer, the decision ultimately rests on the policies of the credit risk department within each bank. To demonstrate this process, we have categorized income and age into three groups, but it may be beneficial to increase the resolution for more complex scenarios. Furthermore, including other significant variables can help to minimize any potential bias in the inferred results.

4. Conclusions

Our research is focused on creating client profile scenarios to measure the probability of credit risk in a historical bank database. To achieve this, we proposed a Bayesian Network based on the relationship between factors such as Age, Income, Loan Grade, Joint Income, and Homeowner. By analyzing these factors, we identified different profiles: Pessimistic, Neutral, and Optimistic. We can assign a credit risk profile to a new applicant by knowing these profiles. This is a powerful computational tool based on statistical inference techniques, and the structural models between variables are still at the research frontiers. However, comparing the results of other models based on similar information will be necessary in a later work. Additionally, we can enhance the applicability of our work by enriching the presented probability tables with information from the credit market.

Acknowledgments

We thank the Supercomputing Laboratory of Apan Research Group "Energy Systems and Advanced Materials" for providing us access to supercomputing resources.

References

[1] Alkhawaldeh, A. A., Jaber, J. J., and Boughaci, D. (2019). A mortality approach for estimating the probability of default in credit risk. *Journal of Advanced Research in Law and Economics*, 10(8): 2233-2243. DOI: https://doi.org/10.14505/jarle.v10.8(46).01.

[2] Ankan, A. and Panda, A. (2015). pgmpy: Probabilistic graphical models using python. In *Proceedings of the 14th Python in Science Conference (SCIPY 2015)*. Citeseer. DOI:10.25080/Majora-7b98e3ed-001

[3] Assef, F., Steiner, M. T., Steiner Neto, P. J., and Franco, D. G. d. B. (2019). Classification algorithms in financial application: Credit risk analysis on legal entities. *IEEE Latin America Transactions*, 17(10):1733–1740.

[4] Bellini, T. (2019). Chapter 4 - LGD modelling. In Bellini, T., editor, *IFRS 9 and CECL Credit Risk Modelling and Validation*, pages 155–213. Academic Press.

[5] Buncic, D. and Melecky, M. (2013). Macroprudential stress testing of credit risk: A practical approach for policy makers. *Journal of Financial Stability*, 9(3):347–370. Part of special issue Central banking 2.0. https://doi.org/10.1016/j.jfs.2012.11.003

[6] Burova, A., Penikas, H., and Popova, S. (2021). Probability of default model to estimate ex ante credit risk. *Russian Journal of Money and Finance*, 80(3):49–72. DOI: 10.31477/rjmf.202103.49

[7] Chatterjee, S. (2016). Modelos de riesgo de crédito. *CEMLA, Boletín, LXII*(3):273–300.
[8] Cole, R. A. (1998). The importance of relationships to the availability of credit. *Journal of Banking & Finance*, 22(6):959–977. DOI:10.1016/S0378-4266(98)00007-7.
[9] Dar, A. A., Anuradha, N., and Qadir, S. (2019). Estimating probabilities of default of different firms and the statistical tests. *Journal of Global Entrepreneurship Research*, 9:2251–7316. https://doi.org/10.1186/s40497-019-0152-8.
[10] Dimitris, C. (2000). *Managing credit risk, analysing rating and pricing the probability of default*. Euromoney Institutional Investor PLC.
[11] Enright, C. G., Madden, M. G., and Madden, N. (2013). Bayesian networks for mathematical models: Techniques for automatic construction and efficient inference. *International Journal of Approximate Reasoning*, 54(2):323–342. https://doi.org/10.1016/j.ijar.2012.10.004.
[12] Gao, L., and Xiao, J. (2021). Big data credit report in credit risk management of consumer finance. *Wireless Communications and Mobile Computing*, 2021:4811086. https://doi.org/10.1155/2021/4811086.
[13] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, United States of America.
[14] Hernán, M.A. and James M. R. (2023). *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC.
[15] Huang, S. J. and Yu, J. (2010). Bayesian analysis of structural credit risk models with microstructure noises. *Journal of Economic Dynamics and Control*, 34(11):2259–2272. Special Issue: 2008 Annual Risk Management Conference held in Singapore during June 30 - July 2, 2008. DOI:10.1016/j.jedc.2010.05.008.
[16] Leong, C. K. (2015). Credit Risk Scoring with Bayesian Network Models. *Computacional Economics*, 47:423–446. DOI:10.1007/s10614-015-9505-8.
[17] Marichal, C. (2008). Banking History and Archives in Latin America. *The Business History Review*, Vol. 82, No. 3, A Special Issue on Business in Latin America (Autumn, 2008), pp. 585-602. http://www.jstor.org/stable/40538504.
[18] Martínez-Sánchez, J. F., Martínez-Palacios, M.T V., and F Venegas-Martínez, F. (2016). An analysis on operational risk in international banking: A Bayesian approach (2007–2011). *Estudios Gerenciales*, 32(140), 208-220 DOI: 10.1016/j.estger.2016.06.004
[19] Martínez-Farías, F. J. (2023). Final state of bank loan: https://www.kaggle.com/dsv/5280325.
[20] Masmoudi, K., Abid, L., and Masmoudi, A. (2019). Credit risk modeling using Bayesian network with a latent variable. *Expert Systems with Applications*, 127:157–166. https://doi.org/10.1016/j.eswa.2019.03.014.
[21] Murillo-López, S., and Venegas-Martínez, F. (2011). Cobertura de los sistemas de pensiones y factores asociados al acceso a una pensión de jubilación en México. *Papeles de Población*, 17(67), 209-250. Disponible en: https://www.scielo.org.mx/pdf/pp/v17n67/v17n67a8.pdf
[22] Prieto, G. (2000). Estimación del riesgo crediticio en México. *Revista Internacional de Fondos de Pensiones*, (2):45-52.
[23] Risk Metrics Group Inc, J. M. (2007). *Credit Metrics*. RiskMetrics Group.
[24] Roisenvit, A. and Zárate, M. (2006). Hacia una cultura del risk management. superintendencia de entidades financieras y cambiarias. *Revista del BBanco Central de la RRepública de Argentina*.
[25] SHCP (1990). Ley de Instituciones de crédito: Gobierno de México and Secretaría de Hacienda y Crédito Público and Comisión Nacional Bancaria y de Valores.
[26] Kovalova, E. and Cug, J. (2015). Credit risk and LGD modelling. *Procedia Economics and Finance*, 23:439–444. 2nd Global Conference on Business, Economics, Management and Tourism. DOI:10.1016/S2212-5671(15)00379-2.
[27] Sucar, L. E. (2021). *Probabilistic Graphical Models*. Springer.

[28] Sánchez, J. F. M., García, S. C., and Martínez, F. V. (2020). Money laundering control in Mexico: A risk management approach through regression trees data mining. *Journal of Money Laundering Control*, 23(2):427– 439. DOI:10.1108/JMLC-10-2019-0083.

[29] Teles, G., Rodrigues, J., Rabelo, R. A. L., and Kozlov, S. A. (2020). Artificial neural network and Bayesian network models for credit risk prediction. *Journal of Artificial Intelligence and Systems*, 2:118–132. DOI:10.33969/AIS.2020.21008.

[30] Transparency, G. (2015). Revisions to the standardised approach for credit risk. *BIS,Transparency Group of the Basel Committee on Banking Supervision*.

[31] Transparency, G. (2021). Fintech and the digital transformation of financial services: implications for market structure and public policy. *BIS, Transparency Group of the Basel Committee on Banking Supervision*.

[32] Turner, B. M. and Van Zandt, T. (2018). Approximating Bayesian inference through model simulation. *Trends in Cognitive Sciences*, 22(9):826–840. https://doi.org/10.1016/j.tics.2018.06.003.

[33] Vargas Sánchez, A. and Mostajo Castelú, S. (2014). Medición del riesgo crediticio mediante la aplicación de métodos basados en calificaciones internas. *Investigación & Desarrollo*, 2:5 – 25. DOI: 10.23881/idupbo.014.2-1e

[34] Varotto, S. (2012). Stress testing credit risk: The great depression scenario. *Journal of Banking & Finance*, 36(12):3133–3149. https://doi.org/10.1016/j.jbankfin.2011.10.001.

[35] Wang, F., Ding, L., Yu, H., and Zhao, Y. (2020). Big data analytics on enterprise credit risk evaluation of e-business platform. *Information Systems and e-Business Management*, 18(3), 311-350. https://doi.org/10.1007/s10257-019-00414-x