CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Tropical AgriSciences

Department of Economics and Development

Development of a scoring model based on fuzzy rules for Syrian Microfinance sector

Master's Thesis

Supervisor:

Vaclav Kozeny

Author:

Osama Darwish

Declaration

I hereby declare that I have done this thesis entitled Credit Scoring in Microfinance independently, all texts in this thesis are original, and all the sources have been quoted and acknowledged by means of complete references and according to Citation rules of the FTA.

In Prague 23rd August 2019

....
Osama Darwish

Acknowledgment

At first, I am grateful to my supervisor, Dr. Václav Kožený, for his meaningful time, insightfully supportive advice and comments. It is a pleasure to meet him and work with him, and I appreciate his guidance throughout the work with my genuinely thankful for his effort, and patience.

I also would like to sincerely thank Dr. Vladimir Verner, for his precious and invaluable recommendation and advice during the whole academic years I had in this university.

Secondly, I would like to express my gratitude to the internal grant agency of the faculty of Tropical AgriSciences for the financial assistance during the data collection period and the entire staffs of the faculty of Tropical AgriSciences who have contributed towards my professional development.

Thirdly, I would like to express my sincere gratitude to my parents for supporting me financially and in prayers as well.

Finally, I would like to extend my appreciation to my friend Mehyar Najla for hist time and support during my thesis work, and to my classmates Andrew Laitha, William Corredor, Valeriya Timoshenko, Sabina Kožichová, Kateřina Holmanová, and David Murcia Higuera, who we were like one big family during the studies. May the God of abundance grant you all your heart desires.

ABSTRACT

Microfinance sector has been growing immensely for the past two decades. Huge number of projects and programs to alleviate poverty, while microloans were the pioneer in fighting poverty and helping poor people to develop and sustain. Since MFIs operate on loans and micro-credits, it is crucial form them to curb the default rates in their loans and be able to identify the potential clients if they might default or not. For this purpose, the so-called credit scoring is used, where the banks build their own evaluative models to classify their clients and predict the probability of default. In order to obtain a successful scoring model, so many methods are used. Roughly, these methods can be classified as parametric statistical methods (or traditional methods) like LDA and LR… etc; and non-parametric statistical methods like CART, and soft computing method like artificial neural networks. Nevertheless, the existing statistical methods might be hard for MFIs' loan officers to use without having a deep background about these methods. Thus, in this thesis we aim to exploit the fuzzy rules that have the structure of "**IF, THEN**" rules to build up a verbal credit scoring model. This fuzzy rules-based credit scoring model can be used by any loan officer without the need of any knowledge related to statistical and mathematical methods. To this end, the problem of how to construct these fuzzy rules arises taking into account the enormous size of options that these rules might include. Therefore, we utilize the genetic algorithms that seem to be a proper approach to gradually design a set of fuzzy rules for default prediction. The dataset used for building the needed model are composed of 500 client's portfolios were collected from Tartous city in Syria. This data is demographical that contains, for example, age, gender and education of each client, to name a few. Moreover, the data contains the class of each client, a "good" or a "bad" one, based on his historical loan repayment. The GA are used to build a set of fuzzy rules that can classify this data and compare the prediction with the true class of each client. The resulted accuracy of the proposed model reaches 70,8% with a loss of, only, 7.6% compared with logistic regression-base statistical classification model.

Keywords: microfinance, credit scoring, fuzzy rules, genetic algorithms.

Contents

List of Figures:

List of tables:

List of Acronyms

CBS………………... commercial bank of Syria

CB…………………. Central Bank

MFI………………… Microfinance institution

GA………………… Genetic Algorithm

ANN……………… Artificial Neural Networks

LDA………………. Linear Discriminant Analysis

CART……………… Classification and Regression Trees

CBC………………. Credit and Monetary Council

SFBIs……………… Social Financial Banking Institutions

USD……………… United States Dollars

SP………………… Syrian Pounds

LR………………… Linear Regression

1. INTRODUCTION

Banks are one of the strongest pillars of the economy of any country, especially when It comes to their most important function, granting loans to both public and private sectors.

Basically, loans form half or more of the bank's assets and main source of their income, in order to generate that income and make a successful business, banks has to take some risks. Some might say taking risk should be avoided, but when it comes to banking industry, (Churchill & Coster 2001). The department of risk management is found to define the good risk from bad risk. Therefor it is crucial to restrain this risk coming from loaning credits to the customers. Most commonly used method in loan approval process is credit scoring which uses the following techniques: Linear Discriminant Analysis, Logit Analysis, K-nearest Neighbor Classifier, Classification and Regression Trees, Neural Networks (Vojtek & Koâenda 2006).

since we are talking about microcredits and poverty reduction, developing countries, are the first thing to come to the mind and people in the rural areas who are living under the poverty line, and in order to help those people it is assumed that by providing microloans would help increasing their income in a way to have a better living standards.

Therefor, Microfinance institutions were found, and the scope of outreach of this industry have grown significantly during the last two decades. While Asia was the pioneer in Microloans more specifically India were more than one million loans conducted according to UNDP (1999).

And due to the huge demand on microloans, MFIs has to come with a strategy regarding the high risk, especially when borrowing money to poor people and sometimes without any collaterals, which drives the repayment rate to become default, risk management includes both the prevention of potential problems and the early detection of actual problems when they occur the below three-step process defines the risk management as an ongoing process:

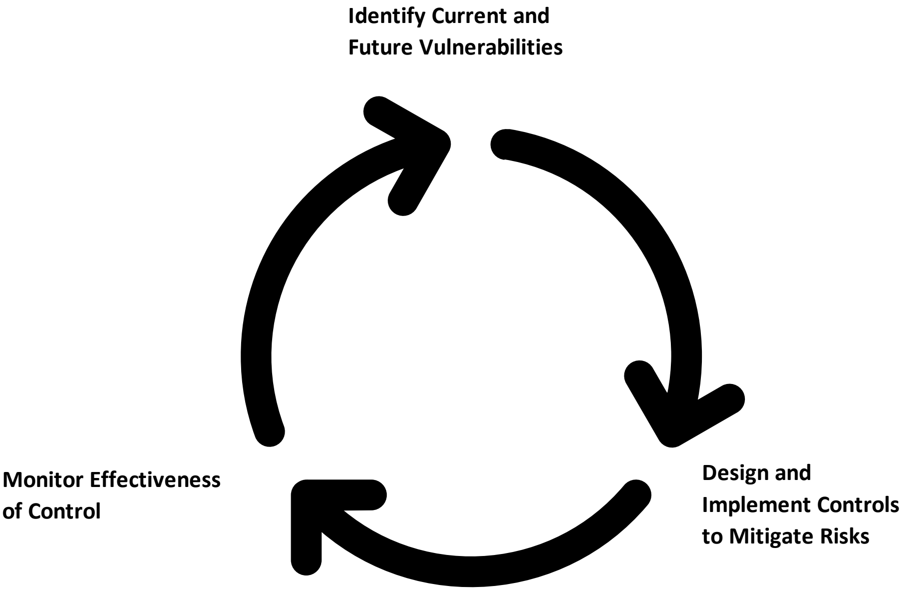


Figure 1: three-step risk management process

The need for credit scoring began to use since lending money to people and possibility of paying depts in the future, since that time financial institutions started to gather information about the creditors and catalogue them to make future decisions on lending money or not for future applicants (Louzada & Fernandes 2016).

According to Schreiner (2000) all the new technological innovations in microfinance, would not be able to replace the loan officers or. During his research in Colombia and Bolivia about credit scoring approach in microfinance, hi found that the scoring method will not be affective comparing to the wealthy countries, because the risk in microloans is not always related to the characteristics that can be quantified inexpensively.

2. LITERATURE REVIEW

2.1. History of Microfinance

Microfinance is the provision of financial services to the people with very low income who also does not have an easy access to financial resources, it was found to support those people in their small business by giving small loans with a low-interest rate.

It worth to mention that, different movements tried to provide small-scale farmers and poor people with the financial services, as we remember the Franciscan monks during the fifteenth century founded the community-oriented pawnshops

microfinance showed up in Europe after the incredible increases in poverty during 16th and 17th century, microfinance had a huge impact on now developed countries and some developing countries, particularly Asia, Germany, and Ireland. Almost from the onset, microfinance meant financial intermediation between micro savings and microcredit(Dieter 2005).

And last we cannot forget one of the main founders of the microcredit the economist Muhammad Yunus (2006 Nobel Prize for Peace), the one who developed the concept of the microcredit, he founded the Grameen Bank, which is a microfinance institution that makes small loans without requiring indemnity.

In the past thirty years, the fields of microfinance noticed an enormous progress, but still, the main concern is how to eliminate poverty around the world.

2.1.1. Ideologies of Microfinance

Microloans were mainly found to help the small-scale farmers with the low incomes to increase their income, start or enhance business to have a better life standard (Morduch & Morduchl 2007). Microfinance institutions are providing small-scale loans to individuals or a group of people which is more used nowadays in amounts of tens/hundreds of dollars for a short period of time usually around 12-24 months. Applicants are asked to provide a purpose for that loan and usually it is enterprise, enhancing a small-scale business or agricultural purposes. A team from the institution will conduct a small research about the client, financial background, collaterals if found, and reputation in the community. The procedure will take about 3-5 months to grant the credits to the borrower.

What is crucial for MFIs before lending the money to public, is to have a guaranteed plan in case of the borrower missed the payment date for more than 6 months, rules and actions in microfinance can differ from country to another and region to region but the ground rules are more likely to be similar (Fouillet et al. 2013).

2.1.2. interest rates in MFIs

The Microfinance institutions worldwide work with a strategy to maintain and enhance the financial services they provide throughout the time, by setting a high interest rate on their loans, as high as it requires to cover the costs of those loans otherwise it will cost losses. On the other hand, MFIs need some sort of funds from donors or government in order to keep their operations running. The three main costs that the institution should consider covering after granting the loans. **Cost of Money,** and **Cost of Loan Defaults** are the two crucial costs that constitute a proportion of the size of the loan for example, if the cost percentage of the credits is 10%, and there is a 1% default, therefore the sum of those costs is 11 USD for a 100 USD loan, and 55 USD for a 5500 USD loan, thus the interest rate is 11% for the loan amount to cover these two costs.

The third type of costs is the **Transaction Cost** it is a percentage of the credited loan. Whereas, the transaction cost for a 500 USD loan dose not differ much of a 100 USD loan, because both loans will require the same amount of time for the staff to conduct the meetings with the borrowers to determine the size of the loan, and the process of the granting the loan, as well as keep tracking with repayments from clients.

2.2. Finance sector in Syria

Back in 1986, the Central Bank, the Commercial Bank, and the specialized sector banks comprised the entire banking system in the country, up to date, all those banks operate under the CB supervision.

Public banks control the financial sector in Syria, the Syrian public banks have been reorganized to serve specific sectors, their main role is to focus on financing the public sector after the 1950s. the privet sector has been developing steadily over the two decades, the privet sector started to expand in their activities to serve privet enterprises.

The vast majority of micro and small enterprises that use banking services prefer public banks, as they are controlling the market, taking into account that nearly half of this group are using CBS (Commercial Bank of Syria).

The systemic and institutional deficits have resulted in inefficient banking systems among state banks, which play a small role until this day in financial intermediation for the sector. The limited availability of demand-driven financial services, particularly loans on fixed assets and working capital financing, means that private companies and individuals are the most affected by the lack of available credit (ICF 2008).

Table 1 is a List of the official financial services providers in Syria presented by ICF (2008).

Table 1 list of official financial institution

Commercial Bank of Syria (CBS)	It is the biggest bank in Syria, provides services such as letter of credits, managing accounts, credits, demand deposits, etc.
Real Estate Bank of Syria	It is specialized in funding real estate purposes. loans are limited by medium term (5-10 years) and long term (+15 years)
Saving Bank	It was restructured in 2000 to be a commercial bank instead of being only Postal Saving Bank.
Popular Credit Bank	It was found to support small businesses, but then it became as deposit and withdrawal financial institution, because of the low interest rates.
Agricultural Cooperative Bank	It provides short-term loans (one year) for agricultural purposes, and medium-term loans for breeding purposes, and long-term loans for bigger projects like irrigation and reclamation projects.
Industrial Bank	It is the smallest bank in the country, specialized with medium and long-term loans for the private sector for industrial purposes.

2.2.1. MFIs in Syria

According to the study done by Abu-Ismail, Abdel-Gadir, & El-Laithy (2011) supported by UNDP, they found that over two million Syrian live under the poverty line, even though, the percentage decreased comparing to the past years. Plenty of programs and projects internally and externally were implemented to curb the poverty in the country and help the poor households to improve their living standards.

Poverty is the one of the main factors that inhibit the economic growth and development in the society. Previously, Syrian banks did not provide loans to individuals who do not have adequate collateral. Thus, microcredit is not found in the Syrian banking system except in the form of consumer loans, which are partially used in 2007 for informal commercial purposes. However, on February 15, Syria came up with a new legislation "General Microfinance" which regulates the establishment of social financial banking institutions, this new legislation authorized the Central Bank's board of Money and Credit to issue a permit to the financial institutions to provide microcredit and other financial services, such as deposits. It reflected a significant opening of the financial markets. The decree allows the Credit and Monetary Council (CMC) of the Central Bank of Syria to license Social Financial Banking Institutions (SFBIs) with the objective of providing microfinance services for public (WB 2008).

Table 2 Is a List of the Microcredits providers in Syria, locals and international presented by (ICF 2008).

Table 2 list of local and international MFIs

Local institutions	
Unemployment agency	Found in 2001. And by the end of 2006 the borrowers were estimated around 57.000 client. The agency operates in four types of loans: 1- micro and small loans. 2- family microloans. 3- microloans for individual small businesses. 4- microloans for enhancing an exists business.
Paradise agency	Found in 2001. After three years of operating, estimated clients were around 3.000 client.
Women empowerment association in economic development	Found in 2003, it aims to encourage women around the country mostly in rural areas to be part of the economic development of the country. It provides consultation and technical support besides workshops and seminars for the female investors.
International institutions	
Aga khan Foundation	It started to operate in Syria in 2002, by the end of 2005 it reached 13.370 distributed loans in Syria.
Center of Commercial projects in Syria	It started in 1996 in Syria, by focusing on the private sector. It provides consulting and technical services for individual investors for the private sector as well as the small enterprise.

United Nation Relief and Works Agency	The program of microloans started in 2003 in the middle east focusing on the Palestinian refugees in Syria as well as the Syrians.

2.3. The mechanism of Microfinance

In general Microfinance institutions (MFIs) have their main function of providing financial services to the low-income households who have long been deemed ‘unbankable”, including the self-employed and customers without collateral assets (Ibtissem & Bouri 2013). With a high risk of repayment, risk management is one of the main pillars of any banking system, while most of the financial institutions are facing the main problem represented by the default risk. Default risk can be defined as the delay in the payments.

Differences between microfinance and commercial banking should be considered when evaluating risk management processes, in the case of MFIs, the officers should focus on credit risk, as the loan portfolio is their main asset. However, this does not assure that microloan portfolios will always fail. contrariwise, credit risk is well controlled by a good methodology.

To avoid this default, the microfinance sector is well known by the strength of its analysis, lending, tracking, and collection procedures, which can be a powerful source of security (Epic Org 2009).

Meanwhile, in banking industry there is so many techniques used to reduce the credit risks and default rate. One of them is the credit scoring approach.

2.3.1. Risk Management

People who has nothing into business and finance will always perceive risk as a bad thing and need to avoid it, but when we talk about risks in finance industries precisely banks, taking risk is sort of a crucial thing to make in order to have a better reward. And this is what is called good risk (Stulz.René M 2015).

Risk is a probability of a potential threat, liability or loss caused by external or internal vulnerabilities, it can also be defined as the exposure to change, the wider and regular the variability, the greater the risk, risk also measures the uncertainty that an investor is willing to take to realize a gain from an investment, risk management, from the point of view of financial institutions, is a successful method in banking environment and therefore these institutions should focus on having a remarkable management of risk. The successful financial institutions are, and will increasingly be those that develop focused strategies, lower their overhead ratios, ingeniously exploit their advantages and know how to calculate their risks (Mwirigi 2006). Types of risks have been introduced by Steinwand (2000). Most risks are common to all financial institutions, mainly financial risks, operational risks, and strategic risks, as shown in **table 3** below.

Table 3 Major Risk Categories(Steinwand 2000).

Financial Risks	**Operational Risks**	**Strategic Risks**
1.Credit Risk 2.Liquidity Risk 3.Market Risk	1.Transaction Risk 2.Fraud Risk	1.Governance Risk 2.Reputation Risk 3.External Business Risks

2.3.1.1. Financial Risks

MFIs are obligated to secure loan facilities and equity from shareholders and extending these as loans to clients as well as pursuing other business objectives. The financial intermediation as financial liabilities (savings, loans funds) are used to make the financial assets (loans and investments) assuming that those assets will generate earnings to pay back the cost of the liabilities (Stepri & Accra 2014).

Credit Risk: is basically seen as the uncertainty or the risk to earnings or capital will reduce due to the late or non-repayment of a loan obligations. It is classified into two main categories, Transaction Risk and Portfolio Risk. (Fernando 2008).

Liquidity Risk: meaning of the inability to meet the current cash duties on time, due to the late re-payments of loans and savings withdrawals institutions won’t be able to meet the market demands.

Market Risk: is represented by three main fields Interest Rate Risk, Foreign Exchange Risk and Investment Portfolio Risk.

1.Interest Rate Risk is known as the risk of financial losses from changes in market interest rate. In MFIs environment, when the cost of funds goes up faster than the institution can adjust its lending rates this is considered as a great interest rate risk (Steinwand 2000).

2.Foreign Exchange Rate it occurs when the institution borrows in one currency and lend in another, so the fluctuations in currency values can make a potential loss.

3.Investment Portfolio Risk generally refers to long-term investment decisions, it must balance credit risks, income goals and timing to meet liquidity needs (Steinwand 2000).

2.3.1.2. Operational Risks

It arises from the errors within daily product delivery and services, those errors come from operational activities from humans or computers entries. Two main types of Operational risks: Transaction risks and Fraud risks. The key drivers of operational risks and their mitigating practices as presented by Stepri & Accra (2014).

Table 4 Drivers of operational risks and their mitigating practices(Stepri & Accra 2014).

	Drivers of Operational Risks	**Mitigates of Operational Risks**
People	*Understaffing, High turnover *Manual processing	*Recruitment, training, motivation *Straight Through Processing
Systems	*Different platforms *Integration	*Project planning *Testing *Contingency Planning
Processes	*Lack of documentation *Unclear responsibilities	*Policies and procedures *Clear responsibilities
External Events	*Service Providers *Changes in regulation *Natural disasters	*Service Level Agreements *Back up service providers *Business continuity

1.**Transaction Risks**: it exists in the delivery of all product and service, Microfinance environment is rich with transaction risk as they must handle a high amount of small transactions daily, meanwhile banks have a highly trained responsible staff and high level of cross-checking system.

2.**Fraud Risks**: or known as integrity risk, is the risk of loss of earnings or capital because of intentional deception by an employee or client or both. Bribes, kickbacks and misleading financial statements are the faces of fraudulent activities in any financial institution, direct theft of funds by loan officers or other staff members is the most common type of fraud in MFIs (Steinwand 2000).

2.3.1.3. Strategic Risks

Strategic risks have internal and external factors, some pernicious business decisions or the wrong implementation of those decisions; poor leadership, or ineffective governance and oversight seen as internal factors, while external risks could be defined as changes in the business or competitive environment.

2.3.2. Effective risk management

MFIs worldwide are considered as a huge revelation in poverty reduction and one of the development approaches, in order to contain and preserve this reputation, MFIs had to find a strong management and great governance to face the risks in this environment, compared to the strategies that have been used in banks and other financial institutions (Mersland and Øystein Strøm 2009). One of the methods used and still been used to reduce the credit risks, is to have at least two employed guarantors, in case of the borrower would stop paying back the loan, the institution will be in contact with the guarantors to cover the loan.

2.4. Loan approval process

Banks worldwide set up a policy for lending money to the public, and it shows them the requirements of a loan, which is regularly under developing to make it more accurate and far from default. Therefore, MFIs needed to increase their efficiency in their processes, minimize their costs, and control their credit risks, to be able to compete with the commercial banks.

Loans around the world are recognized as engines of economic growth. Two type of loans, secured like mortgage loans and Unsecured like consumer loan, credit card, and overdraft, small and medium enterprise loan. Every bank build criteria for loan approval to prevent it from default, the criteria might differ between banks in one or two items, but the whole concept is one. What matters the banks is that the consumer has sufficient income, legal responsibility, no negative history, integrity, reliability, and ability to repay, thus they need to segregate between bad borrowers from the good borrowers, and mostly the ability to predict the percentage of the bad ones.

The scheme below shows the process of a Retail credit process in the Czech Republic, starting from the moment the consumer applies for the loan.

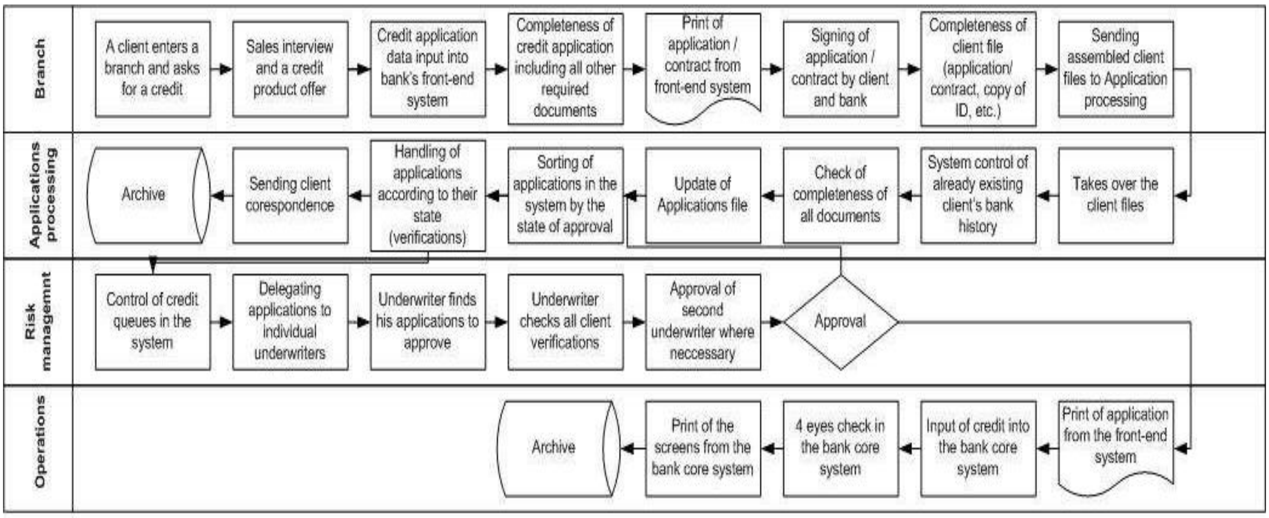

Branch
A client enters a branch and asks for a credit
Sales interview and a credit product offer
Credit application data input into bank's front-end system
Completeness of credit application including all other required documents
Print of application / contract from front-end system
Signing of application / contract by client and bank
Completeness of client file (application/ contract, copy of ID, etc.)
Sending assembled client files to Application processing
Applications processing
Archive
Sending client corespondence
Handling of applications according to their state (verifications)
Sorting of applications in the system by the state of approval
Update of Applications file
Check of completeness of all documents
System control of already existing client's bank history
Takes over the client files
Risk managemnt
Control of credit queues in the system
Delegating applications to individual underwriters
Underwriter finds his applications to approve
Underwriter checks all client verifications
Approval of second underwriter where neccessary
Approval
Operations
Archive
Print of the screens from the bank core system
4 eyes check in the bank core system
Input of credit into the bank core system
Print of application from the front-end system

Figure 2 A Retail process in the Czech Republic (Kožený and Srnec 2010)

2.4.1. The use of credit scoring

This method was introduced in the late 1960s, before that, the loan approval process used to take a period of at least 6 months to make the decision done. Credit scoring is basically a numerical rating used by lenders in the loan approval decision process, it might be also helpful to set the interest rate a creditor can get on his loan. It is a method of evaluating the risk of loan application, it is based on historical data and statistical techniques(Mester 1997).

It has become a very important task as banking industries can gain profit from improving cash flows, reducing possible risks, developing a better managerial decision, enables faster credit decisions, and reduces the losses. It can be defined as a way to segregate good borrowers from bad borrowers in terms of their creditworthiness(Kinematic and Planning 2011).

The objective of credit scoring models is to categorize credit applicants to either a good credit group that is likely to repay financial obligation or a bad credit group whose application will be denied because of his high possibility of defaulting(Lee et al 2002).

To build up the score for the borrowers, banks need demographical data provided by consumers, such as income, age, gender, educational level, housing, marital status, and more. Several statistical methods are used in building the credit scoring model, including Linear probability, Logistic Regression, Linear Discriminant Analysis (LDA), probit models. Those are standard statistical techniques for estimating the probability of default based on historical data on loan performance of the borrower, and all these methods are using parametric methodologies.

Gorzałczany & Rudzi (2016) stated that the classification models are normally tested with three main aspects: their accuracy, their transparency and interpretability, and their computational efficiency meaning of the speed of classification.

Biological inspiration methods or advanced methods have been widely used in artificial intelligence, and non-parametric methods as Neural Networks and Genetic Algorithms are the newer methods where they have the potential to be more useful and accurate in developing the model for commercial loans. In this work, we are trying to apply these methods in the microfinance environment.

2.4.2. Existing techniques used in Credit Scoring

Statistical techniques are commonly used in building the scoring models. Some of these methods are non-linear as well. The scoring models were found for predictive purposes. Some of the conventional statistical techniques, such as logistic regression, linear discriminant analysis, logit analysis and classification and regression trees.

2.4.2.1. Logistic Regression

It Is a binary classification and one of the best methods used, because it is a simple algorithm that perform very good on a wide range of problems. It is used when we know that the data can be separated linearly, and the outcome is 0 or 1. The difference between linear regression and logistic regression is that the first one’s outcome is contentious, while logistic regression has the outcome as limited numbers of possible values.

Application of logistic regression is broad as the results binary 0 or 1. Some of the application fields are:

1. Prediction whether a student will pass or not.

2. Prediction of a loan approval based on the credit score

3. Prediction of a firm failure.

2.4.2.2. Linear Discriminant Analysis (LDA)

The discriminant analysis was first presented by Fisher (1936) whose idea was to find the best way to separate two groups of borrowers using linear combination of variables, this approach is still one of the most broadly established techniques, it was found to mainly classify customers as good credit or bad credit (Mpofu & Mukosera 2014).

Below figure explains how LDA separate the data set into two groups

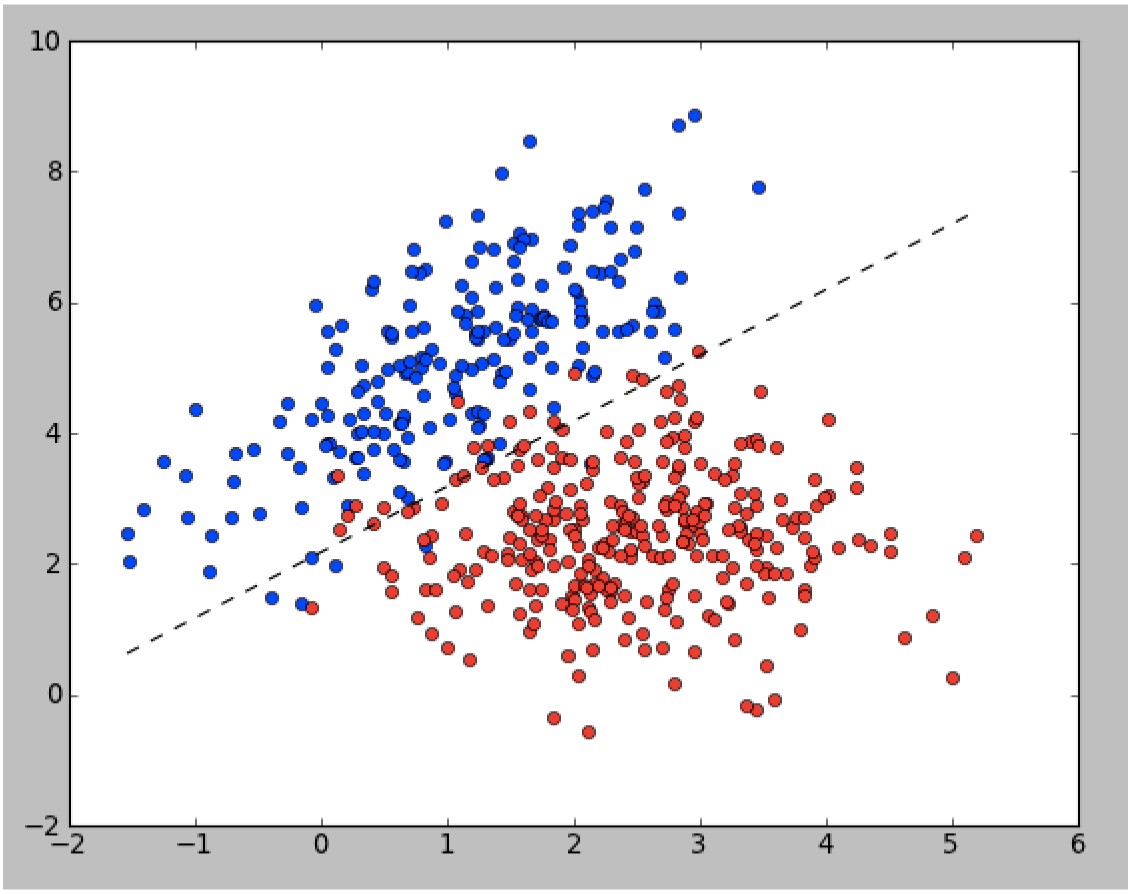


Figure 3 LDA Data classification Source: http://stat-mzhong.blogspot.com/2012/09/python-linear-discriminant-analysis.html

Some writers criticized this method like Eisenbeis (1977) by stating that the rule is optimal only for a small class of distribution. However, Hand & Henley (1997) claim that "*if the variables follow a multivariate ellipsoidal distribution, then the linear discriminant is optimal*". Meanwhile, Vojtek & Koâenda (2006) in their short paper about credit scoring methods, said that the disadvantage of LDA that it requires normally distributed data which is not the case of credit data.

2.4.2.3. Logit Analysis

Logit model is used as an extension of the Linear Discriminant Analysis model because of the non-normality of the credit information data, this extension allows for some parametric distribution (Vojtek & Koâenda 2006).

2.4.2.4. Classification and Regression Trees (CART)

Known also as Decision trees, it is another classification technique used for creating credit scoring models, it is a non-parametric method used to analyze dependent variables as a function of continuous explanatory variables as Breiman (1984) stated in his book.

Below figure can simplify how CART decision is made, an example about credit risk for bank loan.

There is a set of questions we need to ask before starting with CART analysis, for instance let us take this example about credit risk, what are we trying to predict here? Whether the borrower will pay back the loan or not. We need to set the label or output which is in this case the creditability of a borrower. What are the 'if questions' or properties that you can use to predict?

An applicant's demographic profile: age, gender, marital status, salary, education, occupation, etc. these are the features, to build a classifier model, you extract the features of interest that most contribute to the classification

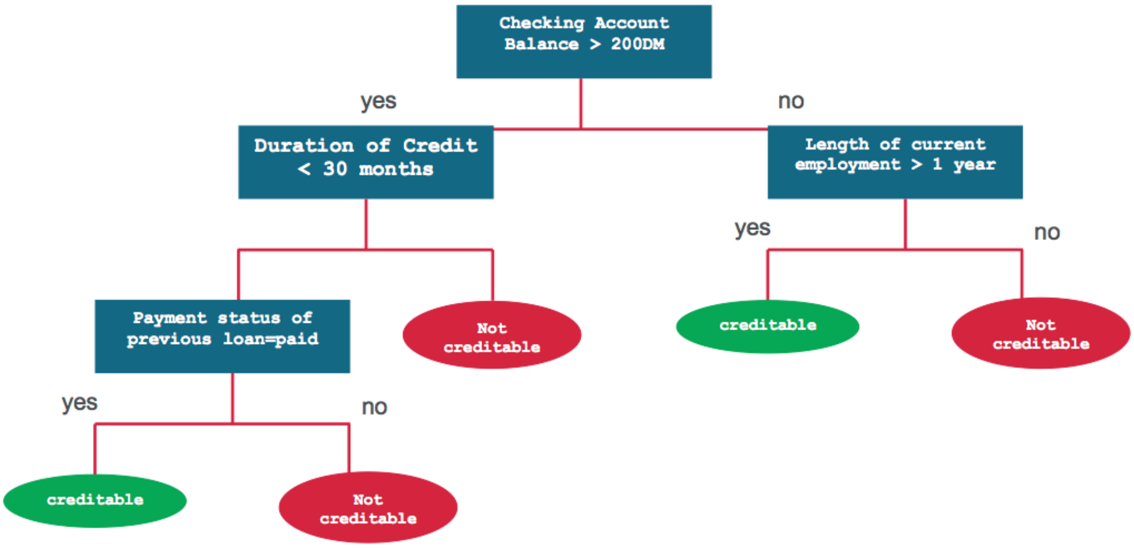

Checking Account Balance > 200DM
yes
no
Duration of Credit < 30 months
Length of current employment > 1 year
yes
no
Payment status of previous loan=paid
Not creditable
creditable
Not creditable
yes
no
creditable
Not creditable

Figure 4 Decision Trees. Source : https://mapr.com/blog/predicting-loan-credit-risk-using-apache-spark-machine-learning-random-forests/

2.4.2.5. Neural Networks

Neural networks are artificial intelligence tools, that allow through a learning process to build up a connection between borrower characteristics and the probability of a default and to determine which characteristics affect more the default-case prediction accuracy. Neural networks are more powerful comparing with the previously-mentioned statistical techniques, as the assumptions do not have to be made about the functional form of the relationship between characteristics and default probability or about the distributions of the variables(Mester 1997). Instead, a well-designed neural network is able to automatically build this relation.

The neural networks are a step forward over the LDA and logistic regression, especially where the dependent and independent variables exhibit complex non-linear relationship. However, it has also been criticized by the difficulty of finding an optimal neural network architecture, as well as the needed long training process and the neural networks' black box nature (Blanco et al 2013).

However, the main drawback of these existing statistical methods for developing credit scoring models, is its dependency on understanding mathematical classification models, that might be complicated or even hard for MFIs loan officers. This is the main motivation for this dissertation theses, in which we aim to create a verbal classification model that can be easily used and even memorized by MFIs' loan officers. To this end, we exploit fuzzy rules that are based on "**IF, THEN**" constraints to build up our verbal classification model. Thus, in the next section, we summarize the principle of fuzzy rules.

3. BASICS OF FUZZY RULES AND GENETIC ALGORITHMS

In this section, we summarize the principles and basic concept of fuzzy rules. Then, the genetic algorithms and their needed steps are illustrated in detail.

3.1. Fuzzy rules

The fuzzy rules are based on "**IF, THEN**" instructions, that state in which situation(s) which decision should be made. To be more precise, a rule has the structure of "**IF** condition is satisfied **THEN** do this action". It is important to mention that the condition in a fuzzy rule can contain a logical **AND**, **OR** and **NOT** to fit a specific case that corresponds to a specific output (i.e., a specific action to be executed).

It was firstly found by professor Lotfi A. Zadeh from university of California, fuzzy logic in the broad sense is a logical system based on a wider generalization of the classical binary logic which is based solely on 0 or 1, in order to infer uncertain circumstances by using inputs between 0 and 1, he noted that correctness and error are not sufficient to represent all logical forms, classical logic is only based on 0 or 1, and this is what many relations depend on while other relationships exist where the position in which it can be considered partly true or partly false at the same time, In two-valued logical systems, a proposition P is either true or false. In multivalued logical systems, a proposition may be true or false or have an intermediate truth value, which may be an element of a finite or infinite truth value set T. For example, if T is the unit interval, then a truth value in fuzzy logic, for example, "very true," may be interpreted as a fuzzy subset of the unit interval. In this sense, a fuzzy truth value may be viewed as an imprecise characterization of a numerical truth value (Zadeh 1988).

A simple example presented by (Gorzałczany and Rudzi 2016), showing the Fuzzy Rule-based credit's knowledge base in the context of financial applications is now presented. Four fuzzy rules enable us to classify applicants as either "good credit risk" ones or "bad credit risk" ones based on two categorical attributes: "status of exciting checking account" ("status", for short) and "credit history" as well as two numerical attributes: "credit amount" and "installment rate in percentage of income" ("installment", for short), in the following way:

IF ["status"is"no checking account"**]THEN** "good credit risk",
IF ["credit history"is"other credit existing(not at this bank) "**] THEN** "good credit risk",
IF ["credit amount"is"Large"**]THEN**"bad credit risk",
IF ["status"is"Less than 0"**]AND[**"installment"is"Large"**]THEN** "bad credit risk"

However, in this thesis, such a rules structure is designed to classify borrowers to a good or bad borrower as a credit scoring verbal model as described in section 5.4. Nevertheless, while thinking about having such rules, a question arises about how to build up these rules taking into account the high dimensional search space in which the options of these rules take place. To solve this problem, we rely on the powerful genetic algorithms to design the needed fuzzy rules-based classification model as illustrated in section 5.5. As a refreshment, the basics of the genetic algorithms are summarized in the next subsection.

3.2. Genetic algorithms

At the beginning of the 1960s, GA was found by John Holland (Holland 1975). in the United States of America and has been developed by his own students in Michigan University between 1960 and 1980.

Genetic Algorithms (GA) are a well-performing tool to find high-quality solutions in large search spaces optimizing a specific criteria. The main principle of the GA is based on Darwin's theory of evolution, where children have a DNA that has, mainly, parents' genes. Considering this concept, a GA is designed to start with a random population of solutions to a specific problem, where each individual (solution) in this population is evaluated based on a criterion defined by the user giving this individual a specific fitness score. By choosing a subset of individuals (solutions) with high score from this population and breeding those chosen individuals (so-called parents); a new population with expected higher score is produced. By repeating these steps iteratively, the GA moves, gradually, towards the optimal solution of the targeted problem.

In a simplified way, The GA is a method of random search that will be used in computing in order to find the correct or closest solutions to the optimal solution among the range of possible solutions that constitute the so-called search space through a series of steps that rely on comparison and find the "distance" between solutions And then select the appropriate solutions and rely on them to shape other solutions (second generation) more appropriate and closer to the optimal solution (Whitly 1994).

The operation of GA needs a big number of variables, GA uses a genetic inspired operator to evolve an initial population into a new population. Each population comprises of chromosomes that represent a genetically encoded individual solution to a specific problem. Each individual has a fitness score assigned to them, which represents its ability in terms of a solution. A new population is evolved by using operators of crossover, mutation, and selection, where the selection is based on the individual's fitness and influences its ability to reproduce into the next generation(Kozeny 2015).

Getting into more details, the genetic algorithm starts with the first population of solutions randomly generated and presented in the form of chromosomes of genes (see Figure 5).

Gene 1	Gene 2	. . . . .	Gene N

Figure 5 A chromosome of genes representing and individual (solution) in the population

Each chromosome is given a score that defines how good is this chromosome (individual), in terms of solving our problem. The chromosomes with highest score are chosen as parent (Michalewicz 1996). The breeding process includes mixing the genes of two parents giving a new chromosome presenting a child. The process of mixing the genes of the parents is called "Crossover". The breeding process is repeated until we have a new population of new chromosomes. To avoid losing the probability of a specific gene occurring in the future generations, the genes in new generation's chromosomes are mutilated with a specific probability to a random possible gene.

Note that the parents are added to the new population which is called "Elitism". The elitism aims to avoid the possibility of getting a low-efficiency individual in the new generation as a result of the breeding and the crossover process. The principle of elitism relies on the copying of the parents that are high-efficient individuals from the previous generation to the new generation, to ensure that they will be used in the second generation if no better individuals (chromosomes) are reproduced (Whitly 1994).

The new population (generation) is evaluated according to the fitness score function and then the process is repeated until one of the individuals satisfies a pre-defined stopping criteria.

Figure 6, shows the flowchart of the steps included in any genetic algorithm.

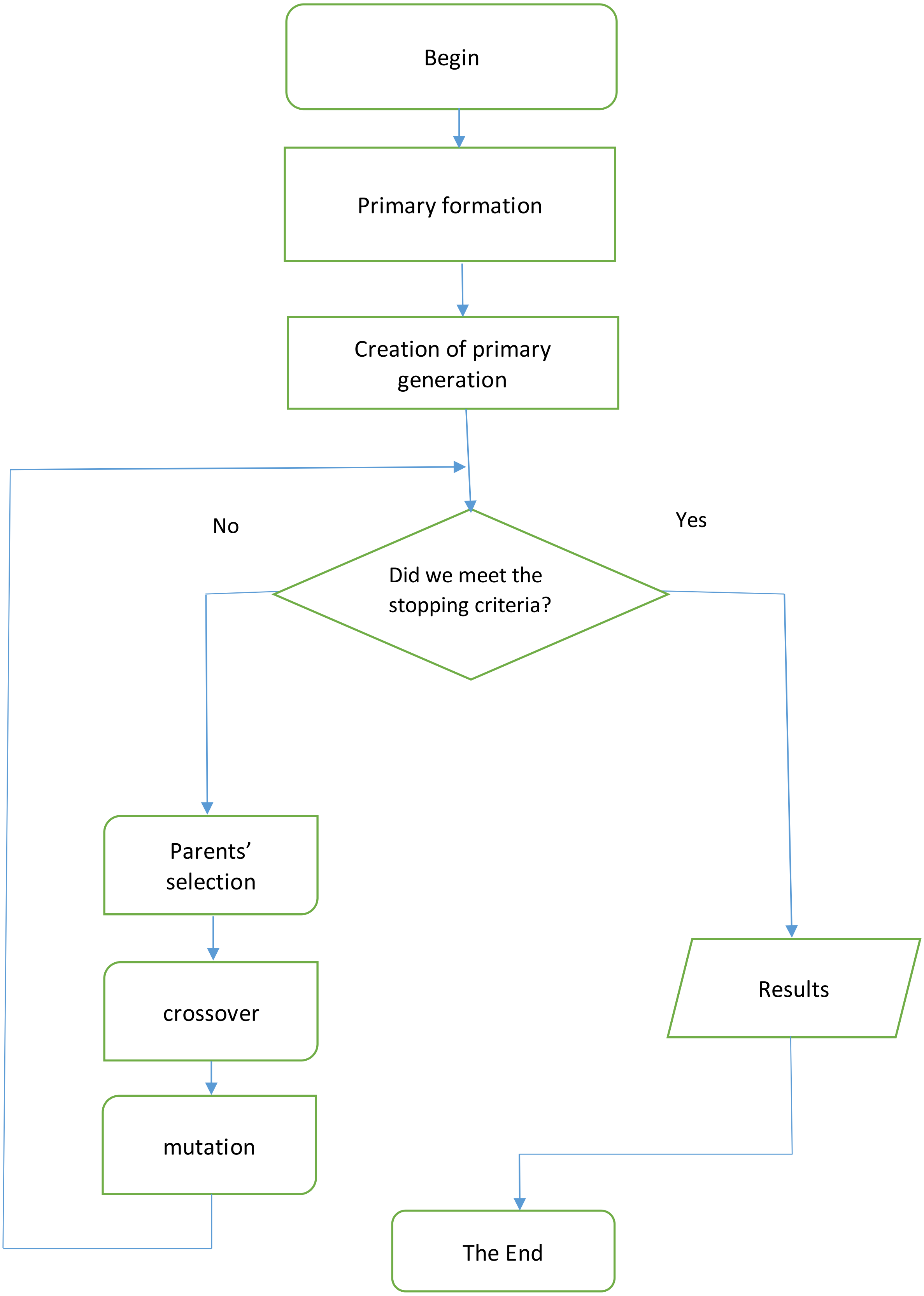


Figure 6 GA Process Flowchart

4. OBJECTIVES

The aim of this research is to assist the MFIs worldwide to have verbal model in order to maintain the credit risks and to develop a new classification model for credit scoring in microfinance that needs no prior knowledge in dealing with advanced statistical classification models. For this purpose, we exploit the fuzzy rules that are based on "**IF, THEN**" constraints to perform the needed classification. Moreover, to compose the fuzzy rules structure, the genetic algorithms are used to develop these rules maximizing the classification accuracy.

The main objectives of this thesis are as the followings:

1. To develop and design a simple scoring model based on fuzzy rules that can be used by MFIs' loan officers where no knowledge or understanding of statistical classification models is needed. The rules target to classify a set of financial data samples from the Syrian microfinance sector.

2. To exploit the genetic algorithms to build up a set of "**IF, THEN**" rules in order to optimize the classification accuracy.

3. To evaluate the performance of the proposed classification model based on fuzzy rules in terms of classification accuracy; and to compare it with statistical classification methods.

5. METHODOLOGY

First, this section describes the study area and the reason why it is chosen. Then, the proposed fuzzy rules-based method to classify creditors in order to eliminate the default, is illustrated. and limitations which occurred during the data collection.

5.1. Study area description

The data collection took a place in Tartus city, Syria. Were the loans have been distributed mostly in Duraykish, Baniyas, Safita, and Ash-shaykh Badr as rural areas of Tartus governorate.

The study area is chosen due to its consideration as one of the safest cities in the Syria during the crises, and consequently, a high number of the borrowers exists comparing to other regions, despite the fact that this region is based on agricultural industries. Therefore, small-scale farmers and unemployment rate is high as well as the amount of distributed loans in the region.

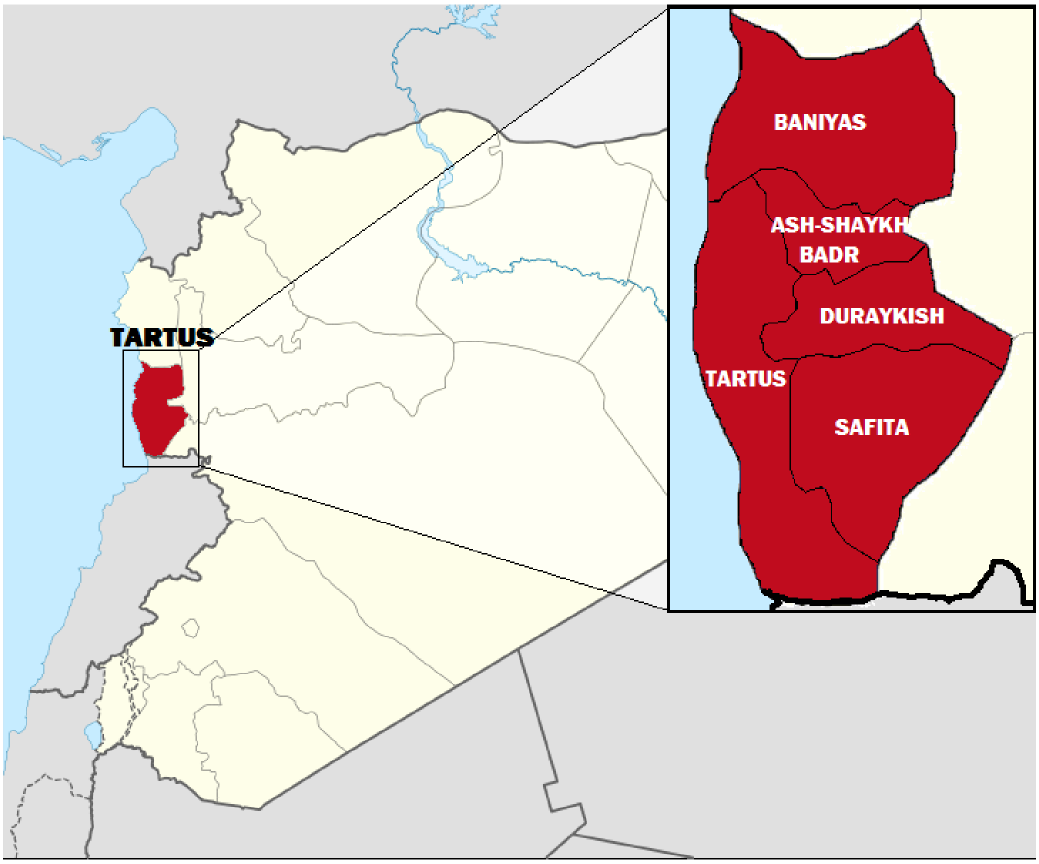


Figure 7 Tartous Governorate with Districts. Source: Wikimedia Commons

5.2. Data collection

A secondary data is collected from **Community service jobs l.l.c** institution located in this area. The institution offered a dataset of loans from 2012 – 2015 with a respective repayment history. Each data sample is composed of twelve independent variables as inputs and one dependent variable as an output. The independent variables (or the so-called "data features") represent the borrower characteristics (e.g., age, gender…etc) as shown in later in **Table 8**. The dependent variable which is the classification output, includes two possible classes: "good" and "bad" based on borrowers' repayment history.

Out of the offered dataset, 500 cases are chosen based on simple sampling method, the data is conducted by the institution in a report method as well as questionnaires provided by them.

5.2.1. Samples from the data

In this subsection, **Table 5** shows how the data is formatted into MS Microsoft excel sheet, it is archived by the institution as hand write reports in hard copies forms. In **Table 6,** the primary data collected by the institution is shown. These data, in a form of the reports, are written by the staff of the institution based on conducted interviews with potential clients. Moreover, a second report is made by the staff in the location of the project claimed by the client; where the staff aim to check on the client's assets. The decision if the client gets the credits or not is based on those reports and other factors. The loan repayment characteristics include two possible cases. The first case represents a "good" borrower who have paid the installments on time or within six months. The second case includes "bad" borrowers who are due for more than six months. As a result, the collected dataset consists of 283 cases representing "good" borrowers and 217 out of 500 considered as "bad" borrowers based on the categorization of the institution.

Table 5 Sample of the Data

AGE	GENDER	EDUCATION	MARITAL STATUS	NUMBER OF FAMILY MEMEBERS	HOUSING	INCOME	OCCUPATION	PURPOSE OF LOAN	PROPOSED AMOUNT	OTHER PROPERTIES	Ability to pay	LOAN REPAYMENT
47	Male	primary school	married	6	owned	35,000	freelance	grocery shop	120,000	small shop	6,000 SP	good
22	female	high school	married	3	owned	30,000	freelance	buying a cow	100,000	non	5,000 SP	good
36	male	primary school	married	4	owned	20,000	freelance	home garden	100,000	land	5,000 SP	good
31	male	high school	married	4	owned	30,000	freelance	commercial support	120,000	shop	5,000 SP	bad
42	male	high school	married	5	owned	16,000	employee	educational support	100,000	non	5,000 SP	bad
47	female	primary school	widow	4	owned	25,000	freelance	commercial support	100,000	shop	6,000 SP	good
40	male	university	married	5	owned	50,000	freelance	service support	100,000	shop	5,000 SP	good
45	female	university	married	5	rented	25,000	freelance	crafts support	100,000	non	5,000 SP	good
33	female	high school	married	5	owned	35,000	freelance	grocery shop	100,000	shop	5,500 SP	good
28	female	high school	married	3	owned	15,000	freelance	grocery shop	100,000	shop	5,000 SP	good

Table 6 is showing the loan application forms which was prepared by the institution and all potential borrowers must fulfill in order to be a subject of further procedures.

Table 6 Loan Application Form

مؤسسة فرص العمل لخدمة المجتمع		**Community service jobs l.l.c**
شهرت بالقرار رقم /2767/ تاريخ 2009/11/26		
طلب قرض		**loan Application Form**
الاسم:		name:
العمر		age:
المؤهل الدراسي:		education:
رقم الهاتف:		phone number:
جوال:		cellphone number:
عنوان السكن:		home address:
الحالة الاجتماعية:		marital status:
عدد افراد الاسرة:		number of family members:
طبيعة السكن:		housing:
الرقم الوطني:		birth number:
تاريخ المنح:		date of credits:
عدد العمال المتوقع:		expected number of workers in the project:
اسم المشروع:		project name:
العمل الحالي:		occupation:
الغرض من الدعم:		purpose of loan:
قيمة الدعم المطلوب:		proposed amount of money:
هل لديك قروض من جهات اخرى:		do you have other loans from different organizations:
عنوان المشروع:		project address:

5.3. Data Analysis

In this section we will be talking about the process of how the data was analyzed and the techniques that has been used,

5.3.1. Interpretation of the collected data

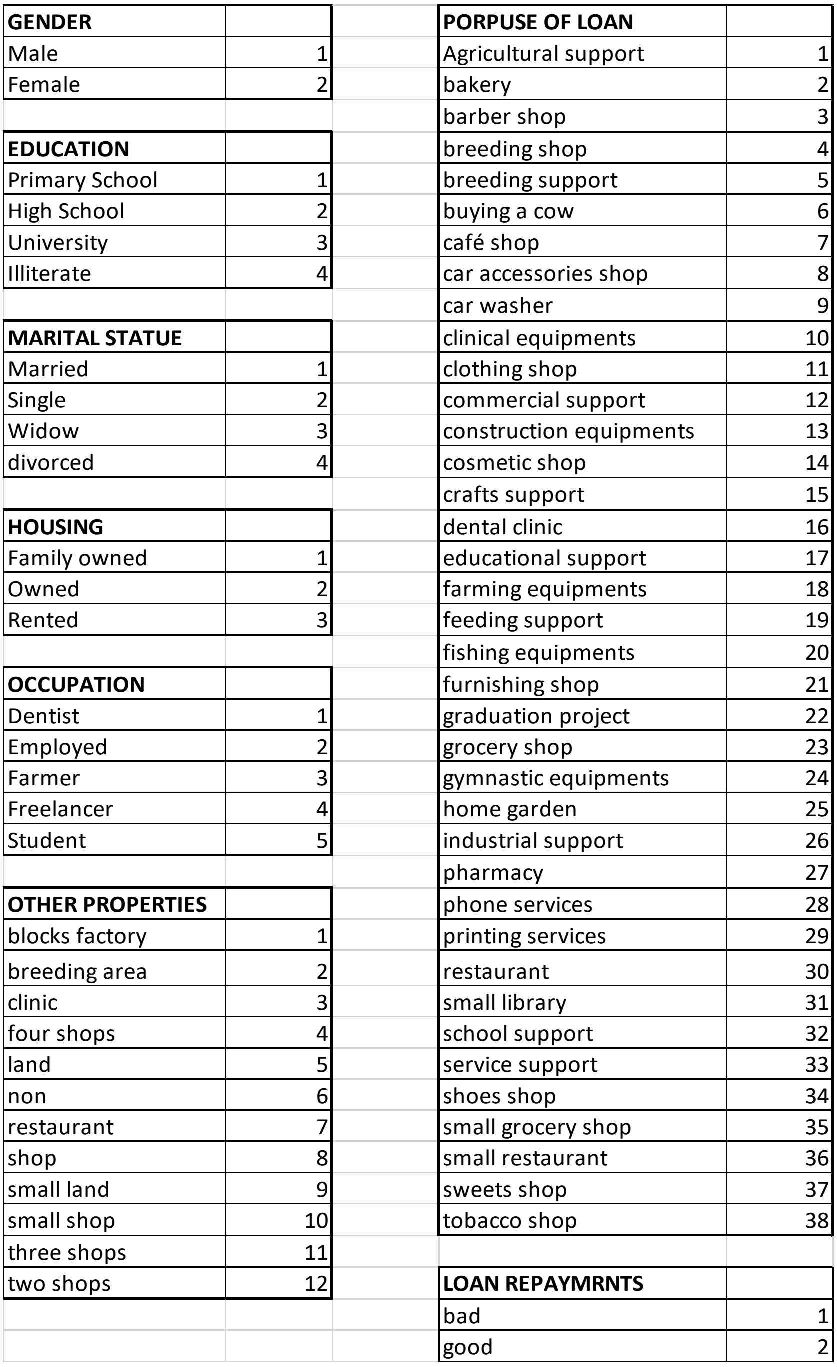
In order to build a set of fuzzy rules classifying the collected dataset using Matrix Laboratory (MATLAB) software, the date is processed in two sequential steps.

First, the data inputs and outputs are changed to a numerical format. **Table 7** shows a set of data samples after reformulating data inputs in MS Microsoft excel. Obviously, the data inputs are presented either by the accurate number in the case of a continuous infinite options of a feature (e.g., borrower's income); or by a number that refers to a specific state of the categorical features (e.g., education is 1 when the borrower's highest education is a primary school graduate). However, **Table 8** illustrates the legend that corresponds to the features with finite options in the data set and the way it is represented numerically.

Table 7 Formulated Data

AGE	GENDER	EDUC-ATION	MARITAL STATUS	NUMBER OF FAMILY MEMBERS	HOUSING	INCOME	OCCUP-ATION	PURPOSE OF LOAN	PROPOSED AMOUNT	OTHER PROPE-RTIES	ABILITY TO PAY	LOAN REPAY-MENT
36	2	2	1	4	2	15000	4	1	100000	5	4500	2
26	2	2	4	2	1	14000	4	6	100000	6	6000	2
43	2	2	1	7	2	17500	5	6	100000	6	6000	2
41	1	2	1	5	2	10000	4	3	100000	8	4000	2
40	2	2	1	3	2	19800	4	6	100000	6	5000	2
31	1	1	1	4	2	30000	4	12	120000	8	5000	1
47	2	1	3	4	2	25000	2	12	100000	8	6000	1
40	1	1	1	5	2	50000	2	33	100000	8	5000	1

Table 8 Legend for the Formulated Data



GENDER	
Male	1
Female	2

EDUCATION	
Primary School	1
High School	2
University	3
Illiterate	4

MARITAL STATUE	
Married	1
Single	2
Widow	3
divorced	4

HOUSING	
Family owned	1
Owned	2
Rented	3

OCCUPATION	
Dentist	1
Employed	2
Farmer	3
Freelancer	4
Student	5

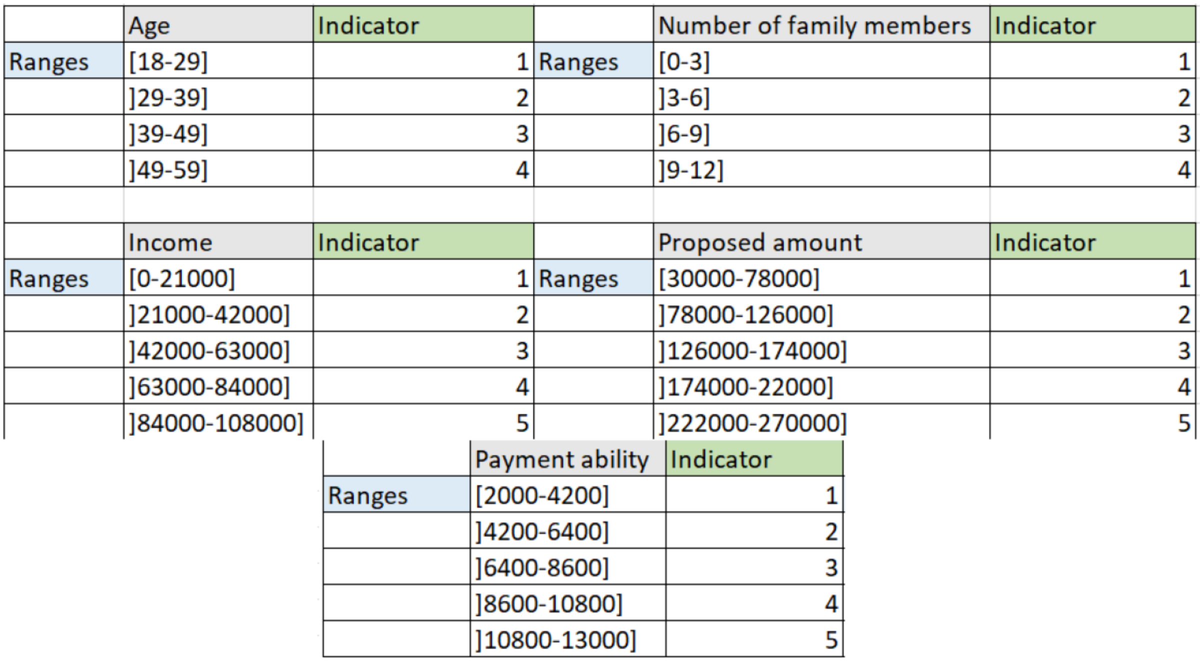
OTHER PROPERTIES	
blocks factory	1
breeding area	2
clinic	3
four shops	4
land	5
non	6
restaurant	7
shop	8
small land	9
small shop	10
three shops	11
two shops	12

PORPUSE OF LOAN	
Agricultural support	1
bakery	2
barber shop	3
breeding shop	4
breeding support	5
buying a cow	6
café shop	7
car accessories shop	8
car washer	9
clinical equipments	10
clothing shop	11
commercial support	12
construction equipments	13
cosmetic shop	14
crafts support	15
dental clinic	16
educational support	17
farming equipments	18
feeding support	19
fishing equipments	20
furnishing shop	21
graduation project	22
grocery shop	23
gymnastic equipments	24
home garden	25
industrial support	26
pharmacy	27
phone services	28
printing services	29
restaurant	30
small library	31
school support	32
service support	33
shoes shop	34
small grocery shop	35
small restaurant	36
sweets shop	37
tobacco shop	38

LOAN REPAYMRNTS	
bad	1
good	2

The second step of processing the data is transforming the features that can have infinite options into finite number of options. More precisely, age, number of family members, income, proposed amount and payment ability can take any continuous value, and thus, we transform each one of them into ranges that can be presented by a finite limited set of integers. In other words, those features are separated into ranges of values. Each range is, then, given a specific number as an indicator that tells to which range this data sample belongs. **Table 9** shows the dataset after changing infinite options' features into ranges, and consequently, finite number of options.

Table 9 Ranges



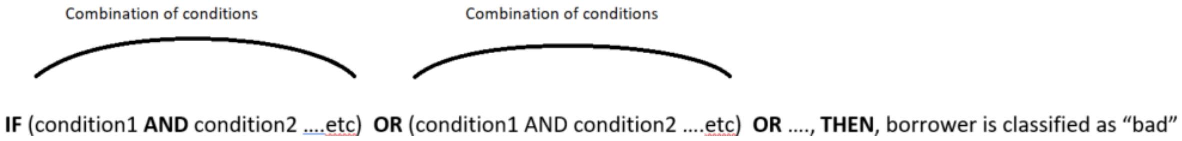
	Age	Indicator		Number of family members	Indicator
Ranges	[18-29]	1	Ranges	[0-3]	1
	]29-39]	2		]3-6]	2
	]39-49]	3		]6-9]	3
	]49-59]	4		]9-12]	4
	Income	Indicator		Proposed amount	Indicator
Ranges	[0-21000]	1	Ranges	[30000-78000]	1
	]21000-42000]	2		]78000-126000]	2
	]42000-63000]	3		]126000-174000]	3
	]63000-84000]	4		]174000-22000]	4
	]84000-108000]	5		]222000-270000]	5

	Payment ability	Indicator
Ranges	[2000-4200]	1
	]4200-6400]	2
	]6400-8600]	3
	]8600-10800]	4
	]10800-13000]	5

5.4. Credit scoring model based on GA-aided Fuzzy Rules

The main objective of this thesis is to build up a credit scoring fuzzy rules-based verbal model that classifies borrowers into good and bad borrowers based on their characteristics. The idea of using the fuzzy rules is to replace the complex statistical models so that all loan officers of MFIs are easily able to understand it and memorize it during their daily job.

Taking into account that the problem presented in this thesis is a binary classification problem (i.e., output is either good or bad), we target to build up a set of rules that can recognize a bad borrower only. However, if the rules are not satisfied, the borrower is considered to be good. Based on this, the **IF, THEN** rules solving our problem follow the following structure as shows in **figure 8**



IF (condition1 **AND** condition2 ....etc) **OR** (condition1 AND condition2 ....etc) **OR** ...., **THEN**, borrower is classified as “bad”

Figure 8 Structure of rule

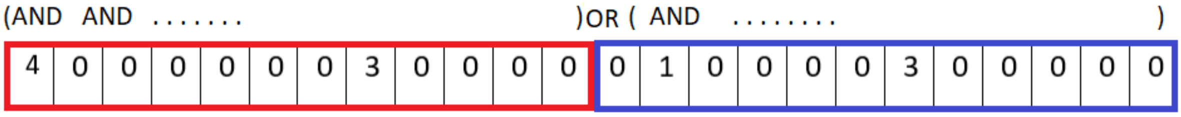
In the previously illustrated rules structure, a condition is a specific feature and a corresponding number. The number represents the corresponding range in which the value of this feature takes place for the borrower we want to classify. Multiple conditions can compose together a combination of constraints needed for the borrower to be bad. This combination is achieved by using the logical **AND** operator. Moreover, more than one combination of conditions can lead to the “bad” classification of the borrower by using the logical **OR** operator.

However, traying to construct those **IF, THEN** based on the existing twelve features leaves us with an enormous number of possible constraints. Therefore, a smart method is needed to obtain a set of rules maximizing the classification accuracy. Hence, we intend to use the Genetic Algorithm method which is able to solve combinational optimization problems gradually, based on natural selection of process that mimics biological evolution. The motivation towards this idea is the fact that the pre-illustrated fuzzy rules structure can be transformed into a chromosome of genes where every gene presents a condition. In other words, as every condition is a number presenting the range of a specific feature, this number can be considered as a gene. Then, the multiple conditions with an AND among each other can be seen as a set of genes. Note that a gene can take the value of zero if this feature is not one of these multiple conditions. As a result, twelve genes representing the range to which each of the twelve features we have belongs, compose together the combination of conditions with an **AND**. Then, another twelve genes can be added to present another combination of conditions, where an **OR** is considered to separate these two combinations of conditions.

To clarify how the fuzzy rules are transformed into chromosomes of genes, an example is needed. Let us say we have a set of rules saying:

IF (age is 4 **AND** occupation is 3) **OR** (gender is 1 **AND** income is 3) **THEN** the borrower is "bad"

This constraint can be read as: if the age belongs to the fourth range in the age feature and if the occupation corresponds to the one equivalent to number 3 in the legends table (TABLE 8), then the borrower is "bad'. Moreover, if the gender is 1 (**Table 8)** and the income belongs to the third range in the income feature, the borrower is also "bad". If none of the two combination of rules mentioned in the example is satisfied, the borrower is good. This example of **IF, THEN** rules can be presented as a chromosome of genes as follows.



4	0	0	0	0	0	0	3	0	0	0	0	0	1	0	0	0	0	3	0	0	0	0	0

Figure 9 Example of IF, THEN rule presented as a chromosome of genes

Note that in the chromosome, every twelve genes follow the same order of features from **Table 7**.

As the fuzzy rules we are targeting can be transformed to a chromosome of genes, in the next subsection we propose a genetic algorithm to construct a set of rules to maximize the classification accuracy.

5.5. GA to construct Fuzzy Rules for credit scoring

In order to construct the fuzzy rules for credit scoring, the genetic algorithms can be exploited as an optimization method to reach a well-performing set of **IF, THEN** rules. As mentioned in previous section, the fuzzy rules can be transferred to a form of a chromosome of genes where the genes represent the conditions that compose these rules (see **Figure 9**).

However, in this thesis, we also care about the simple utilization of the fuzzy rules by MFIs loan officers. Thus, we aim to reach a set of rules that are relatively short in order to make these rules easy to memorize and to be used by bankers. For this purpose, we assume a fixed number of **OR** operators in the targeted fuzzy rules. In this thesis, without loss of generality, we set this fixed number of **OR** operators to three, as this number seem to be reasonable for a banker to remember. Nevertheless, we also want these rules to perform well in terms of classification accuracy, which is the goal of the GA to achieve.

Getting more into the details of the implemented GA for constructing the targeted fuzzy rules, the GA steps are as follows.

1. The population of chromosomes: In this step, a population of 200 chromosomes (individuals) is generated randomly representing the primary population. Each chromosome (individual) is a random set of rules composed of 12 x 4 = 48 genes. These 48 genes are simply four combination of conditions with three **OR** operators in between as explained earlier in this section. To shorten the **IF, THEN** needed rules, we give each gene a probability of 50 % to have a value of zero, which means that this feature presented by this gene is not part of the corresponding combination of conditions. The left 50 % is divided equally among the other options that the feature of this gene can take. For example, the first gene presents the age feature, and it can take one of four possible options as it is divided into four ranges. Thus, in every individual, the first gene of every combination out of the four possible combination of conditions can take the value of 0 with 50 % probability, while it can take the value of 1,2,3 or 4 with 50/4 % probability for each. The resulted population of chromosomes (rules) is shown in **Figure 10**

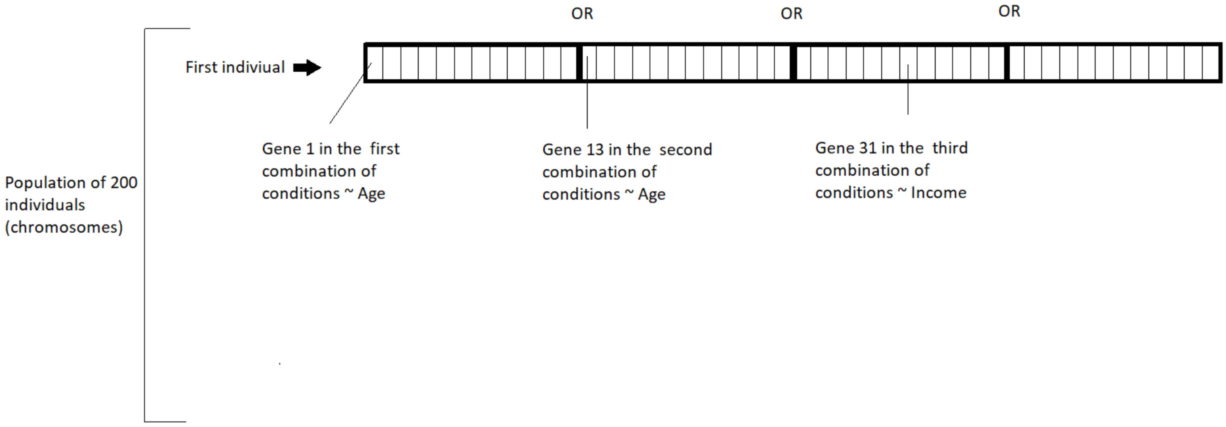

OR
OR
OR
First indiviual
Gene 1 in the first combination of conditions ~ Age
Gene 13 in the second combination of conditions ~ Age
Gene 31 in the third combination of conditions ~ Income
Population of 200 individuals (chromosomes)

Figure 10 Population of chromosomes (rules)

2. Selection: this step aims to select the best chromosomes from the population to set them as parents for the next generation. Thus, first, each chromosome is given a score

based on a chosen fitness function. In this thesis, the fitness function is the classification accuracy that this chromosome corresponding rules give when implementing them on the dataset. In other words, if we implement the chromosome rules on the dataset, the set of the borrowers correctly classified can be considered as S where $S \subset X$ and X is the whole dataset. Then, the score of this individual (chromosome) is $F = 100 \times \frac{|S|}{|X|}$.

By giving each individual in the population of rules a score, we choose the best individuals with highest scores to be considered as parents for the next generation. We set the number of parents to four that are moved directly to the next generation (so-called Elitism) to keep the mixture of genes with high score that exists in the parents. As the population is composed of 200 individuals, the new generation should contain 196 children in addition to the chosen four parents.

Note that if the best individual of the population satisfies the GA stopping criteria, neither breeding nor new population are needed, and instead, the best individual is the targeted solution of our problem (i.e., the fuzzy rules we need for credit scoring).

3. Crossover: This step includes the breeding process to generate the children in the new generation (the new population) from the parents chosen from the previous population. The crossover starts by choosing two parents randomly out of the four parents. The two parents are two chromosomes of genes that should be mixed together to generate one new chromosome, a child, that has a mixture of genes from its two parents. The process of mixing the two parents genes to compose the chromosome of the child is called "Crossover". In this thesis, we consider the random point crossover that can be explained as follows. After choosing the two parents, a random gene out of the 48 genes of each parent is chosen. Note that the same random gene applies for both parents. The genes before the random gene from the first parent are aligned with the genes that are located after the random gene from the second parent and vice versa. This way, two children are the result of the crossover of the two parents. **Figure 11** clarifies the random single point crossover.

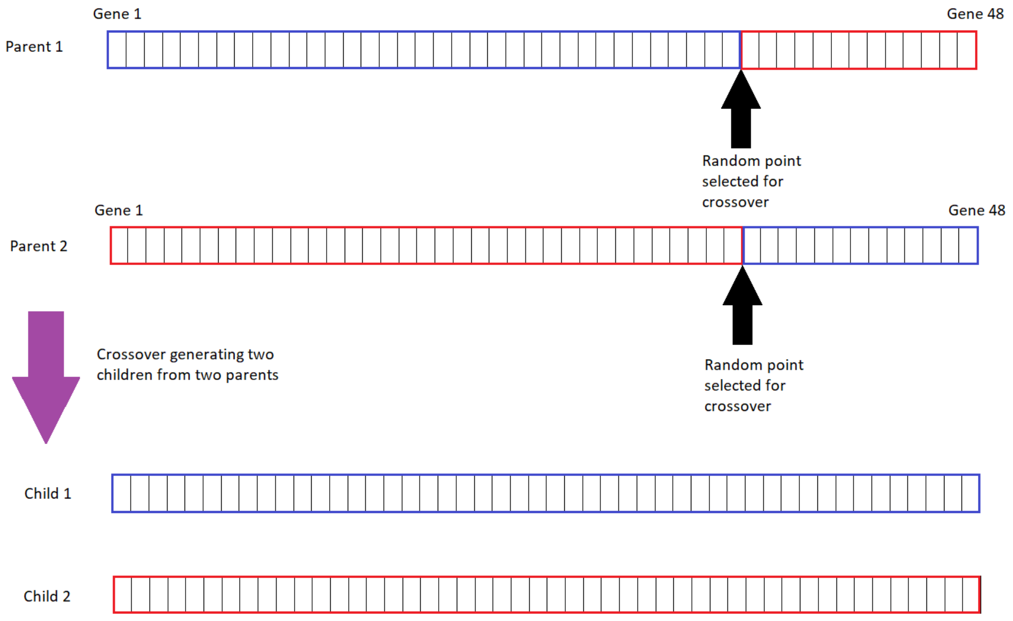

Gene 1
Gene 48
Parent 1
Random point selected for crossover
Gene 1
Gene 48
Parent 2
Crossover generating two children from two parents
Random point selected for crossover
Child 1
Child 2

Figure 11 Random single point crossover

The crossover process is repeated 196/2 times to generate 196 children in addition to the four chosen parents from the previous generation to compose a new generation with a new population of 200 individuals (each individual is a set of rules).

4. Random Mutation: After each crossover, two children are created. However, as the parent's number is four and the children having a mixture of genes from the genes of those parents, some options (values) that a gene can take, might disappear. To prevent the elimination of some options that do not exist in the parents genes; each gene in a child chromosome has a probability of 15 % to mutate randomly to any random possible option that this gene can take.
5. Repeat by going to step number 2.

Note that the stopping criteria we consider is having an accuracy on the good class and an accuracy on the bad class, both, higher than a threshold. We set this threshold to a value, and then gradually increase it till the GA is not able to find a solution satisfying the threshold.

It is important to mention that the GA parameters, such as number of parents in each iteration, percentage of mutation …etc., is set based on trial and error approach.

The flowchart of the proposed GA for building our fuzzy rules for credit scoring is shown in **Figure 12**

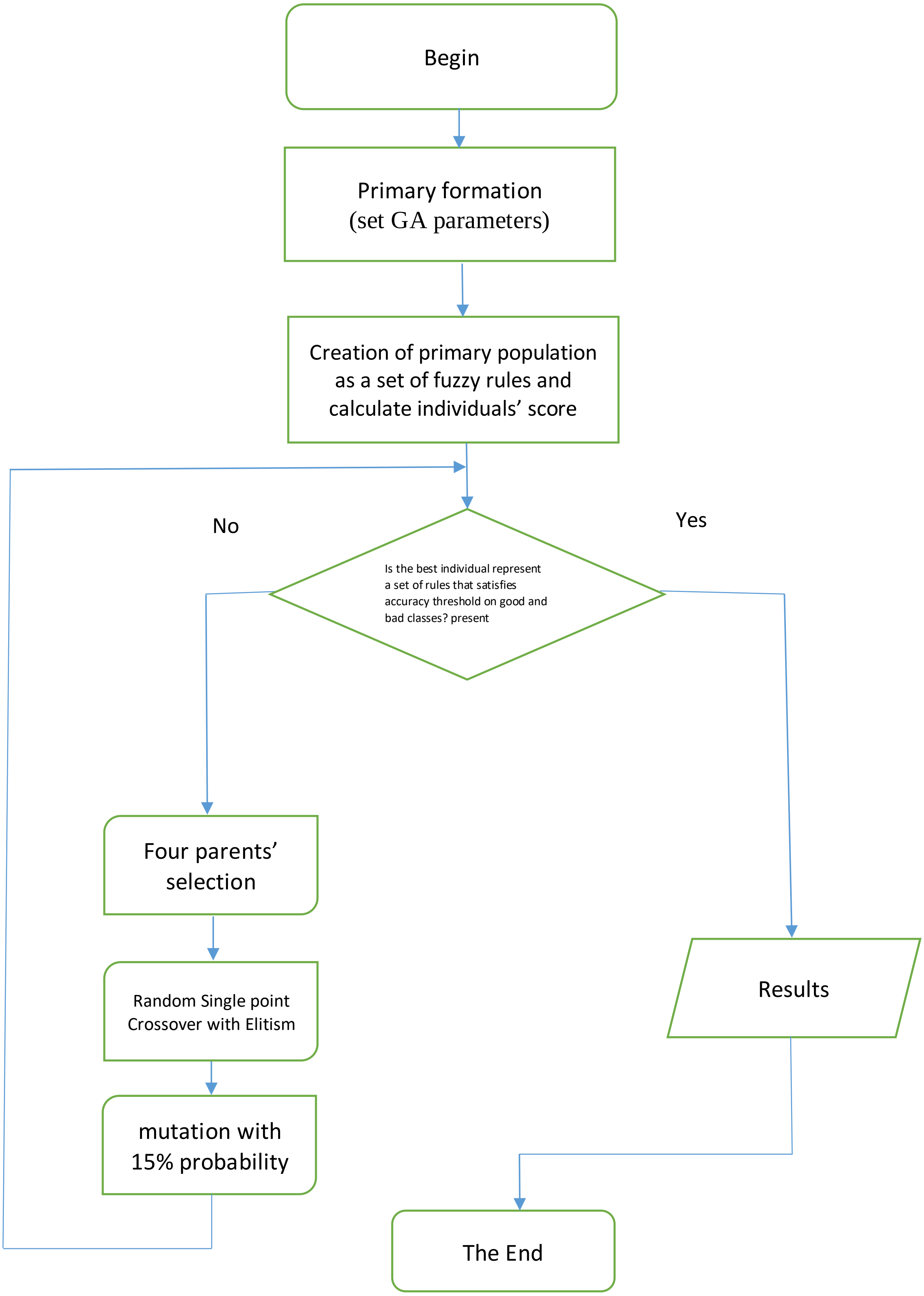

Begin
Primary formation
(set GA parameters)
Creation of primary population as a set of fuzzy rules and calculate individuals' score
No
Yes
Is the best individual represent a set of rules that satisfies accuracy threshold on good and bad classes? present
Four parents' selection
Random Single point Crossover with Elitism
mutation with 15% probability
Results
The End

Figure 12 proposed GA flowchart for building our fuzzy rules

5.6. Limitation of the research

The main limitation of this thesis is the small size of the dataset. This small number of samples leads to a no-guarantee state regarding these data following a specific pattern to be extracted in the form of rules. Moreover, even some rules are possible to be constructed, these rules cannot be generalized to match other samples from different bank or area.

In addition, dividing the data into ranges can be done by multiple ways that need to be tried in order to choose a well performing divisions of this data into ranges. Similarly, GA parameters are, also, set based on experimental measurements of the resulted classification accuracy for each combination of these parameters.

Moreover, there is no specific idea about how long the rules should be to achieve a good performance on the dataset. In other words, the number of logical operators, **AND** and **OR**, that each set of rules representing an individual should contain.

6. Results

In this section we evaluate the performance of the fuzzy rules-based model for credit scoring after building up those rules with the GA.

As described in the previous section, the GA is implemented iteratively till an individual in a population satisfies the stopping criteria. The stopping criteria is having a set of rules with an accuracy that is above 67% on, both, the “good” and the “bad” classes. Nevertheless, as the fitness function is presented by the total accuracy on both classes, the total accuracy of the best individual in each iteration is either equal or larger than the total accuracy of the best individual in the previous iteration. This increment can be shown in **Figure 13**, in which the fitness function value for the best individual in each iteration is shown over the GA iterations.

Figure 13 also shows that the fitness function value almost saturates after the tenth iteration to a value of 70.8%. After the eighth iteration, the increment in the fitness function value is almost negligible. Nevertheless, the saturation of the total accuracy does not mean that the accuracy on each class is saturated, and thus, the GA continues iterating as the stopping criteria relies on the accuracy on each class separately.

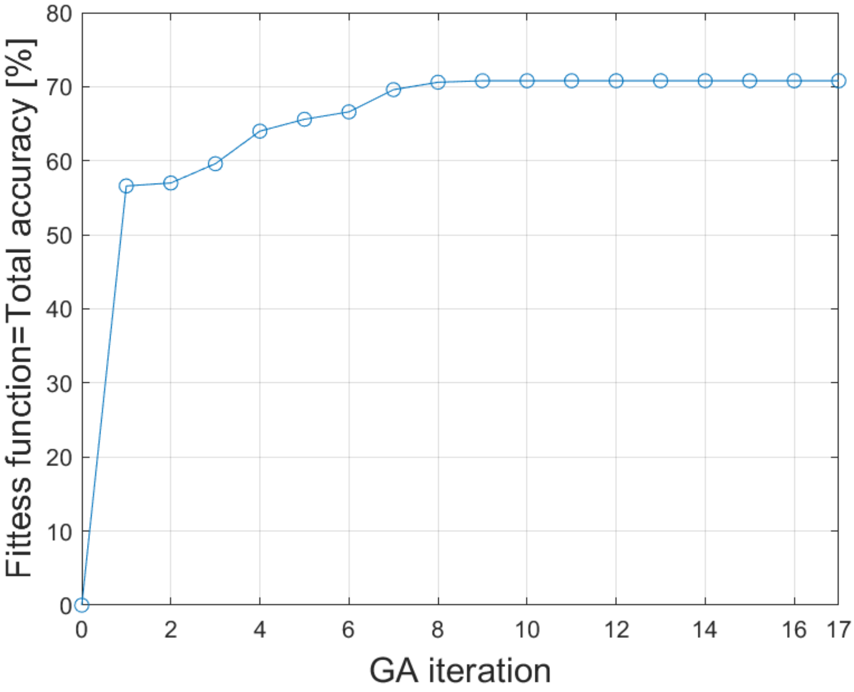
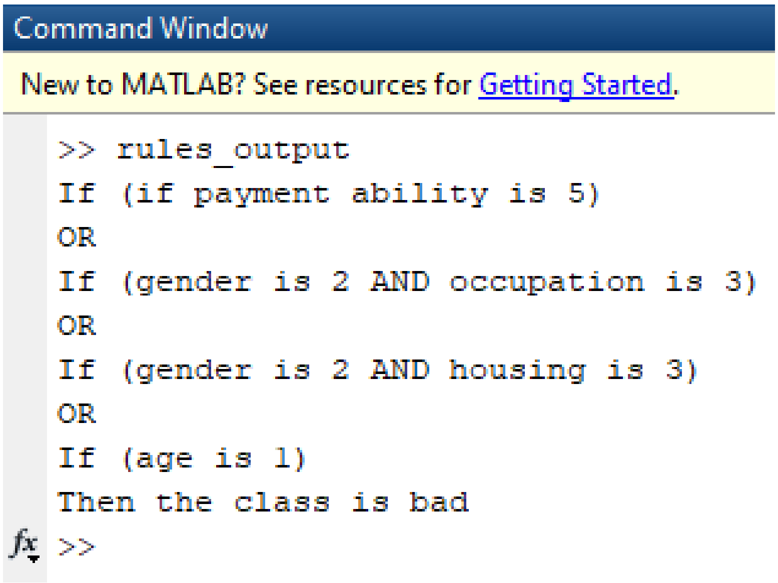


Figure 13 Total accuracy [%]

Figure 13 also shows that in the 17th iteration, the stopping criteria is satisfied, and the resulted rules are shown in **Figure 14**. We can see that the resulted rules rely on the age, gender, occupation, payment ability, and housing features. In other words, a correlation between these features is extracted to indicate which combination of them leads to a default (a borrower to be classified as "bad"). If a borrower does not satisfy those rules shown in **Figure 14**, the borrower is considered to be "good".

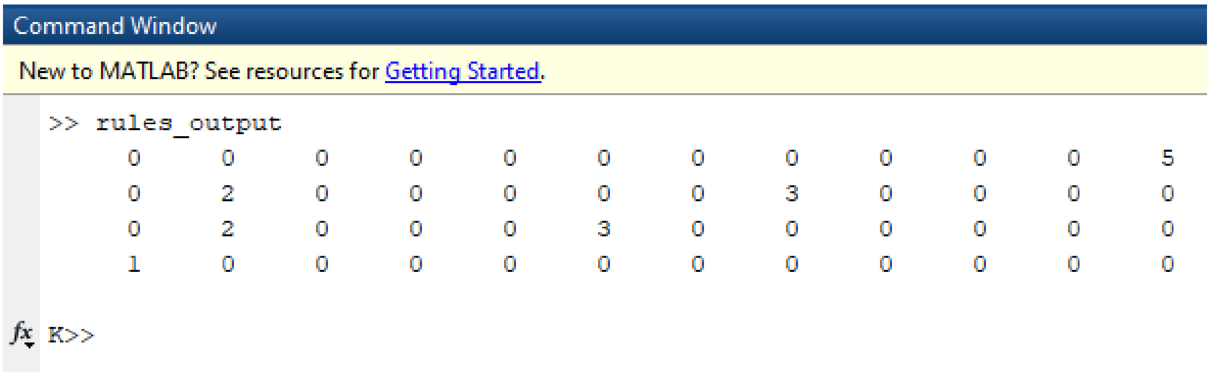


New to MATLAB? See resources for Getting Started.

```
>> rules_output
If (if payment ability is 5)
OR
If (gender is 2 AND occupation is 3)
OR
If (gender is 2 AND housing is 3)
OR
If (age is 1)
Then the class is bad
>>
```

Figure 14 Rules

In **Figure 15**, we can see the corresponding rules as chromosomes, where each column represents one of the features from the dataset (see **Table 7**).



New to MATLAB? See resources for Getting Started.

```
>> rules_output
     0     0     0     0     0     0     0     0     0     0     0     5
     0     2     0     0     0     0     0     3     0     0     0     0
     0     2     0     0     0     3     0     0     0     0     0     0
     1     0     0     0     0     0     0     0     0     0     0     0

K>>
```

Figure 15 Rules as chromosomes

Going back to the GA process, **Figure 16** shows how the accuracy on the good and **Figure 17** shows how the accuracy on the bad changes with the GA iterations. It is obvious that the accuracy on the good class and the bad class does not have to increase over GA iterations, while the total accuracy on both classes should. The reason, as explained previously, is that the total accuracy on both classes is the accuracy that represents the fitness function (the best individual score). In **Figures 16** and **17**, we see that the accuracy on the bad and the good class swings up and down over the GS iterations. However, although **Figure 13** shows that the total accuracy of the classification saturates after the eighth iteration, **Figure 18** illustrates that the individual accuracy on each of the two classes does not saturate after the eighth iteration. Moreover, the stopping criteria which is having an accuracy above 67% on each of the classes, is satisfied only in the last iteration (the 17th iteration) where the GA stops.

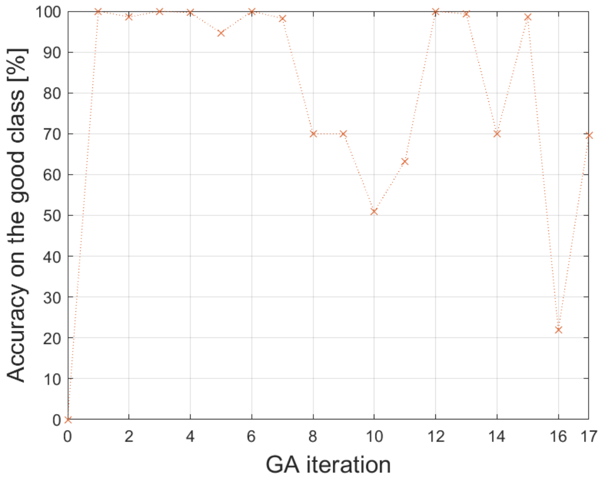


Figure 16 Accuracy on the good class

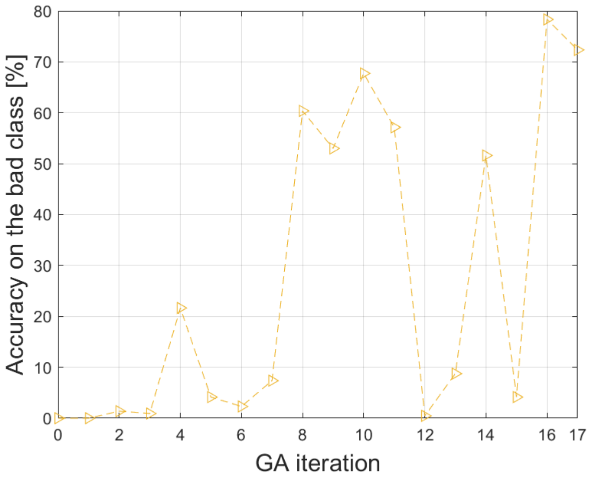


Figure 17 Accuracy on the bad class

Figure 18 summarizes the accuracy behavior combining, all, total, “good” and “bad” accuracies

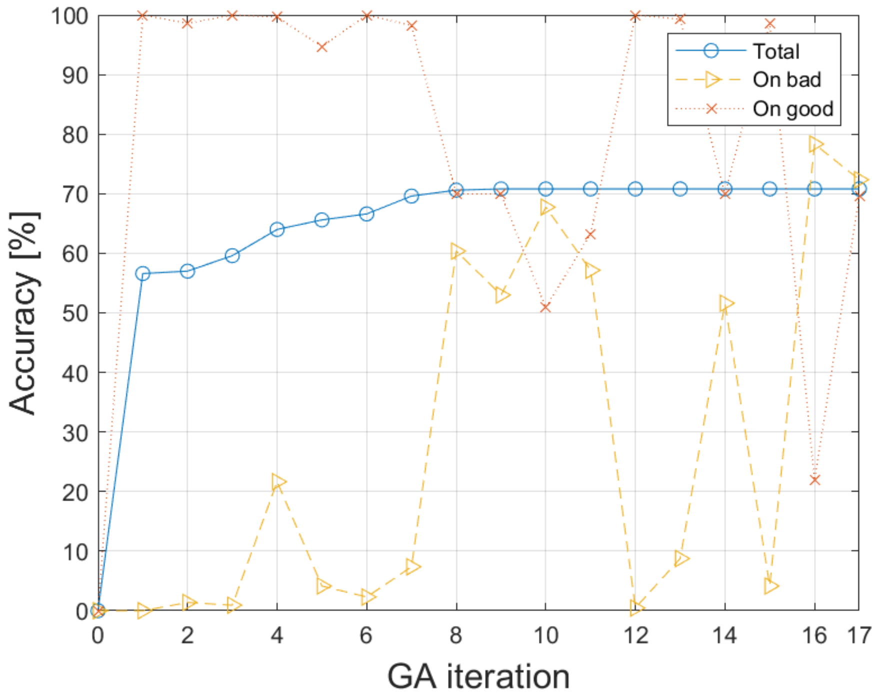
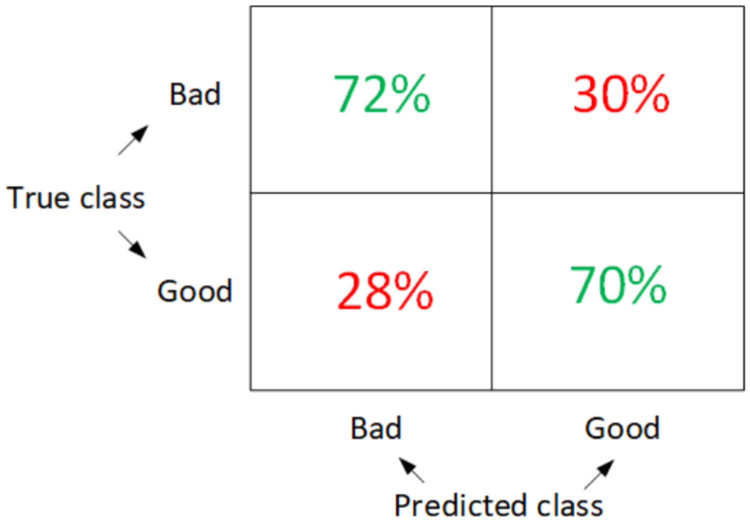


Figure 18 Accuracy behavior for "good" and "bad"

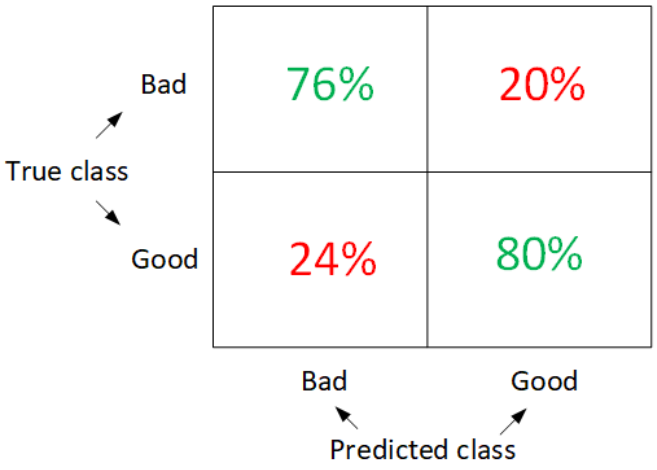
For more details, **Figure 18** shows the confusion matrix that shows the accuracy on the “good” class and on the “bad” class, in addition to the classification error percentage on, also, “good” and “bad” classes. In more details, as the class can be either “good” or “bad” and can be predicted as either “good” or “bad”, there are four possible cases that might happen. The four cases are either to predict a “good” class correctly as a “good” class or wrongly as a “bad” class and vice versa. In Figure 18, we see that the “good” class is predicted correctly by our fuzzy rules-based model in 70% of the cases and wrongly in 30% of the cases. Similarly, on the “bad” class the probability of a correct classification is 28%.


True class
Predicted class

	Bad	Good
Bad	72%	30%
Good	28%	70%

Figure 19 Confusion matrix when Fuzzy rules implemented

For comparison purposes, **Figure 19** shows the confusion matrix when the logistic regression is implemented on the dataset to classify it into “good” and “bad” classes (which is a binary classification problem). We can see that the logistic regression, as a statistical model, is able to reach a higher accuracy on, both, “good” and “bad” classes comparing to the proposed fuzzy rules. More precisely, the logistic regression is able to reach an accuracy of 80% and 76% on the “good” and the “bad” classes, respectively. The logistic regression misclassifies the prediction with a probability of 20% on the “good” class and 24 % on the “bad” class. Nevertheless, with a loss of 10% on the “good” class and 4% on the “bad” with respect to the logistic regression, the fuzzy rules show a promising performance taking into account the simplicity of their verbal nature for bankers’ usage.


True class
Predicted class

	Bad	Good
Bad	76%	20%
Good	24%	80%

Figure 20 Confusion matrix when Logistic regression implemented

More detailed comparison between GA-aided fuzzy rules credit scoring model and the statistical logistic regression-based model is shown in **Figure 20**. In addition to the difference in “good” and “bad” accuracies between logistic regression model and fuzzy rules model, the difference in the total final accuracy is shown. We can see that the logistic regression can reach a total accuracy of 78.4%, while the fuzzy rules total accuracy is 70.8% as mentioned before. In other words, the total classification accuracy loss the fuzzy rules suffer in comparison with the logistic regression is 7.6%, which is considered as a low loss when comparing a words-based classification model with a statistical one.

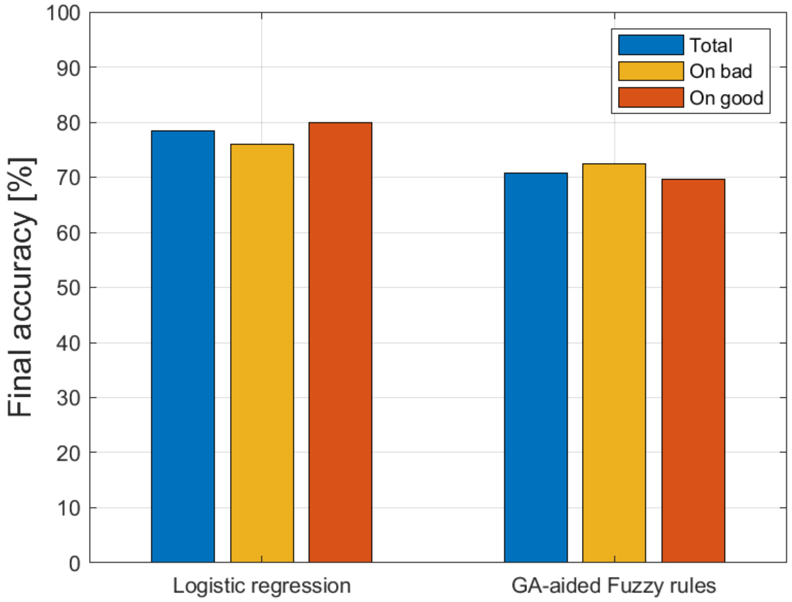


Figure 21 Comparison between GA-aided fuzzy rules model and the statistical logistic regression-based model

7. CONCLUSION AND RECOMMENDATIONS

In this thesis, we target building a credit scoring with the fuzzy rules of the structure "**IF, THEN**". The model aims to classify a set of borrowers into bad and good borrowers based on the characteristics of these borrows (e.g., age, gender, education.. etc). The main benefit of this idea is the simplicity of the fuzzy rules as these rules are expressed verbally and can be used and memorized by any banker without the need of any statistical knowledge. In other words, a banker can decide whether the client is a good or a bad creditor based on a set of "**IF, THEN**" rules that are pre-build based on a collected sample of data. However, building such rules presents the problem of how to come up with a set of rules that maximizes the classification accuracy. For this purpose, we propose a genetic algorithm (GA) that is able to solve this problem. The reason for choosing the GA is the fact that the GA is considered as smart optimization methods, especially in combinational optimization problems. Thus, a set of fuzzy rules can be transformed into a combination of genes composing a chromosome. Then, by having a population of chromosomes where each chromosome present a set of **"IF, THEN"** rules, The GA can be used traditionally to move from one population to another optimizing the fitness function presented by the classification accuracy of the rules (the chromosomes).

The results show a promising performance achieved by our proposed GA-aided fuzzy rules credit scoring model with a total accuracy of 71%. The proposed model reaches an accuracy that is close to the logistic regression which is a typical statistical method for classification. However, although our proposed fuzzy rules-based model includes verbal constraint only, it loses only 7.4% comparing to the logistic regression accuracy on the same data set. Nevertheless, the fuzzy rules-based model outperforms the logistic regression model in terms of simplicity of usage without any needed knowledge of any statistical method.

As a conclusion, a fuzzy rules-based credit scoring model can be efficient in terms of default discovery and risk elimination. On the other hand, results show that the achieved performance is close to the statistical models that might be complicated for bankers to use. Moreover, the genetic algorithms seem to be able to help building the needed fuzzy rules quickly and efficiently.

However, during this research, the small size of data set has caused difficulties in terms of classifying the data and understanding its behavior. For the future, it is recommended to collect a larger data set that can be used to extract a more general model. In other words, collecting larger data from different resources can lead to a strong model that can be used for different microfinance institutions and for variety of creditors categories in the targeted area.

Moreover, the microfinance institutions are recommended to have their own criterias in classifying their customers which can highly help engineers to develop the needed fuzzy rules models based on computer science methods like the genetic algorithm. This collaboration can give a step forward towards default discovery with quick decision-making regarding loan approval, high prediction accuracy. on the other hand, this process can be done by any ordinary loan officer in any bank.

REFERENSES

Abu-Ismail, K., Abdel-Gadir, A., & El-Laithy, H. (2011). *Arab Development Challenges Report Background Paper 2011/15 - Poverty and Inequality in Syria (1997-2007)*. 1–43. Retrieved from http://www.undp.org/content/dam/rbas/doc/poverty/BG_15_Poverty and Inequality in Syria_FeB.pdf

Blanco, A., Pino-Mejías, R., Lara, J., & Rayo, S. (2013). Credit scoring models for the microfinance industry using neural networks: Evidence from Peru. *Expert Systems with Applications*, *40*(1), 356–364. https://doi.org/10.1016/j.eswa.2012.07.051

Churchill, C., & Coster, D. (2001). *M ICROFINANCE*.

Dieter, H. (2005). *www.econstor.eu*.

Epic Org. (2009). *Credit Scoring*.

Fernando, N. A. (2008). *رغصلا يهانتملیومتلا رطاخم ةرادإ تاحارتقلااو تاظحلاملا ضعبرغصلا يهانتم*
34–1 .*ضعب تاظحلاملا تاحارتقلااو ةرادإ رطاخم ليومتلا.*

Fisher, R. A. (1936). *THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS Table I*.

Fouillet, C., Hudon, M., Harriss-White, B., & Copestake, J. (2013). Microfinance Studies: Introduction and Overview. *Oxford Development Studies*, *41*(SUPPL 1). https://doi.org/10.1080/13600818.2013.790360

Framework, R. (2008). *Policy and Regulatory Framework for Microfinance in Syria*. (January). Retrieved from http://documents.worldbank.org/curated/en/126971468132270230/pdf/434330WP0Syria10box032736801PUBLIC1.pdf

Gorzałczany, M. B., & Rudzi, F. (2016). *A multi-objective genetic optimization for fast , fuzzy rule-based credit classification with balanced accuracy and interpretability*. *40*, 206–220. https://doi.org/10.1016/j.asoc.2015.11.037

Henley, W. E. (1997). *Statistical Classi ® cation Methods in Consumer Credit Scoring : a Review*. 523–541.

Ibtissem, B., & Bouri, A. (2013). Credit Risk Management in Microfinance: the Conceptual

Framework. *ACRN Journal of Finance and Risk Perspectives*, *2*(1), 9–24.

ICF. (2008). *رغصلا يهانتملا ليومتلا قوس ميقت : ايروس يئاهن ريرقت 2008 ويلوي*.

Kinematic, A. N., & Planning, C. T. (2011). *Short Papers*. *2005*(402), 1–8. https://doi.org/978 0 7340 3893 7

Kozeny, V. (2015). Genetic algorithms for credit scoring: Alternative fitness function performance comparison. *Expert Systems with Applications*, *42*(6), 2998–3004. https://doi.org/10.1016/j.eswa.2014.11.028

Lee, T., Chiu, C., & Lu, C. (2002). Credit scoring using the hybrid neural discriminant technique.....file danneggiato, guardare nella cartella origine. *Expert Systems with Applications*, *23*, 245–254.

Louzada, F., & Fernandes, G. B. (n.d.). *Classification methods applied to credit scoring : A systematic review and overall comparison*.

Mersland, R., & Øystein Strøm, R. (2009). Performance and governance in microfinance institutions. *Journal of Banking and Finance*, Vol. 33, pp. 662–669. https://doi.org/10.1016/j.jbankfin.2008.11.009

Mester, L. J. (1997). What Is the Point of Credit Scoring? *Business Review (Federal Reserve Bank of Philadelphia)*, (February 1997).

Michalewicz, Z. (1996). Michalewicz Z. Genetic Algorithms + Data Structures = Evolution Programs (3ed).PDF. *Artificial Intelligence in Medicine*. https://doi.org/10.1016/S0933-3657(96)00378-8

Morduch, J., & Morduchl, J. (2007). *The Microfinance Promise*. *37*(4), 1569–1614.

Mpofu, T. P., & Mukosera, M. (2014). *Credit Scoring Techniques : A Survey*. *3*(8), 2012–2015.

Mwirigi, P. K. (2006). An Assessment of Credit Risk Management Techniques Adopted By Microfinance Institutions In Kenya. *MBA Project, University of Nairobi*, (November).

Schreiner, M. (2000). Credit Scoring for Microfinance: Can It Work? *Journal of Microfinance*, *2*(2), 105–118. Retrieved from https://ojs.lib.byu.edu/spc/index.php/ESR/article/view/1404

Steinwand, D. (2000). A Risk Management Framework for Microfinance Institutions. *Development, Financial Systems Services, Banking*, (July), 1–70.

Stepri, C., & Accra, H. (2014). *Risk Management Training Programme for Money Lenders*.

Stulz.René M. (2015). *Applied cor porate finance*. 7–19.

UNDP. (1999). Essential Microfinance: A Synthesis of Lessons Learned. *Evaluation Office*, (3), 1–12. Retrieved from www.undp.org/eo

Vojtek, M., & Koâenda, E. (2006). *Short papers*. *2005*(402), 152–167.

Whitly, D. (1994). *Genetic algorith tutorial*.

Zadeh, L. A. (1988). Fuzzy logic. *Computer*, *21*(4), 83–93. https://doi.org/10.1109/2.53