# CHAPTER ONE

# Introduction

## 1.1 Background

The National Pension and Insurance Fund (Government Sector) seeks to meet pension needs. It supports finance for investment projects and provides loans to enable pensioners to face economic situations. The main aim of the national pensions fund is that to manage the system of subscriptions with which workers and the government contributed as well as providing social protection for pensioners through the monthly payment and other services offered by the system. The funds perform other functions through some specialized units. The social affairs unit which provides some social aid to pensioners represented in health care, assistance for patients, support for families, caring for siblings of pensioners among students and offering some assistance for pensioners on public occasions and festivals.

As a specialized institution, the Social Development Institution for Pensioners was established in the year 2000. It works for providing easy–term financing for small project operated by the pensioner or his family members in the aim of to generate additional income and providing employment opportunities for the family and to reintegrate them for effective participation in social life.

In addition to that, the institution provides commodities for the pensioner to purchase. All through its track since 2000 until now, the institution was keen on adopting many means to asses and evaluate work in aim of identifying the positive sides to support and promote them and knowing the negative sides to treat them. Those means were represented in the following: workshops to evaluate the experiment in

it’s all stages. The formation of some internal committees to evaluate the administrative and organizational work of the institution.

Evaluation was obtained by carrying out field studies and surveys to assess the economic and social impact of institution activities among pensioners by some economy experts and researchers. The most important comments and demands which were regression by any customer on financing are represented in:

- raising the amount of the funding,
- increasing the technical support for them
- enlightening them about the culture of financing and micro- projects,
- achieving broader geographic expansion and arrival to pensioners in the various states,
- finally, innovating current collective plans for pensioners.

## 1.2 Problem Definition:

With the huge amount of data stored in paper files, databases and other repositories in enterprises, it is increasingly important to develop powerful means of analysis and perhaps interpret these data and draw out interesting knowledge that can help in decision making.

The National Pension and Insurance Fund is one of these institutions that has a problem to determine the adequate budget for investments and loans.

## 1.3 Scope of Work:

The study is conducted on the National Pension and Social Insurance Fund (government sector) to analyze the large data in the databases and the part of the data for loans and investment.

## 1.4 Objectives:

The main objectives that the research seeks to identify are:

- Building a classification model for loan.
- Predicting the budget for investment and loans for the coming financial years.

## 1.5 Motivations of the research:

Data might be one of the most valuable assets of any corporation, but only if there is a way to reveal valuable knowledge hidden in raw data. Data mining allows extracting diamonds of knowledge from the historical data and predicting the outcomes of future situations.

## 1.6 Methodology:

The researcher will follow the steps of knowledge discovery (KDD) as described in the following diagram:

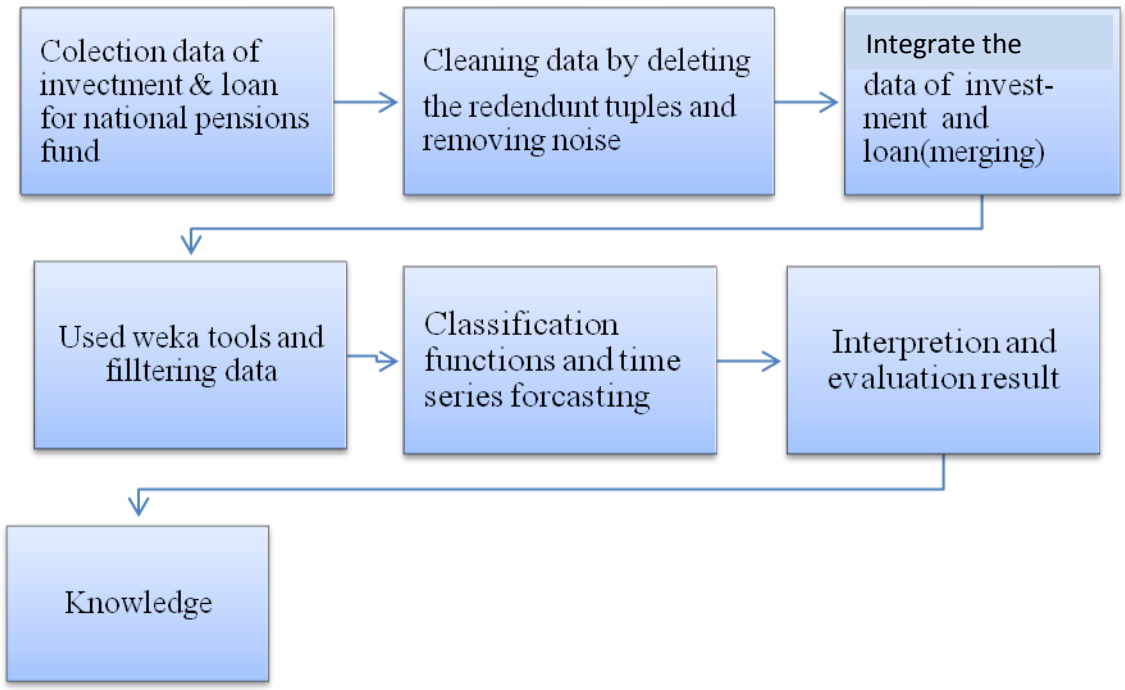



Figure 1.1 Research methodology

## 1.7 Thesis Organization

The study consists of five chapters. The first chapter deals with the general frame work of the research. The second chapter represents the literature review. The third chapter deals with the methodology from data collection, the tools & functions which were used. The fourth chapter presents the results and discussion. Finally, the fifth chapter discloses with the conclusion and recommendation.

# CHAPTER TWO

# Literature Review

## 2.1 Overview:

This chapter will give information regarding data mining and a general idea literature review which is related to the research work.

## 2.2 Data Mining

Data Mining (DM) is the process of discovering interesting knowledge from large amounts of data stored either in databases; data warehouse; or other information repositories. Data mining has been defined as: " the nontrivial extraction of implicit, previously unknown, and potentially useful information from data bases/data warehouses. It uses machine learning, statistical and visualization techniques to discover and present knowledge in a form which is easily comprehensible to humans." (Kamber, Jian Pei, 2012).

"Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases." (Kamber, Jian Pei, 2012).

While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is part of the knowledge discovery process. Figure 2.1 shows data mining as a step in an iterative knowledge discovery process.

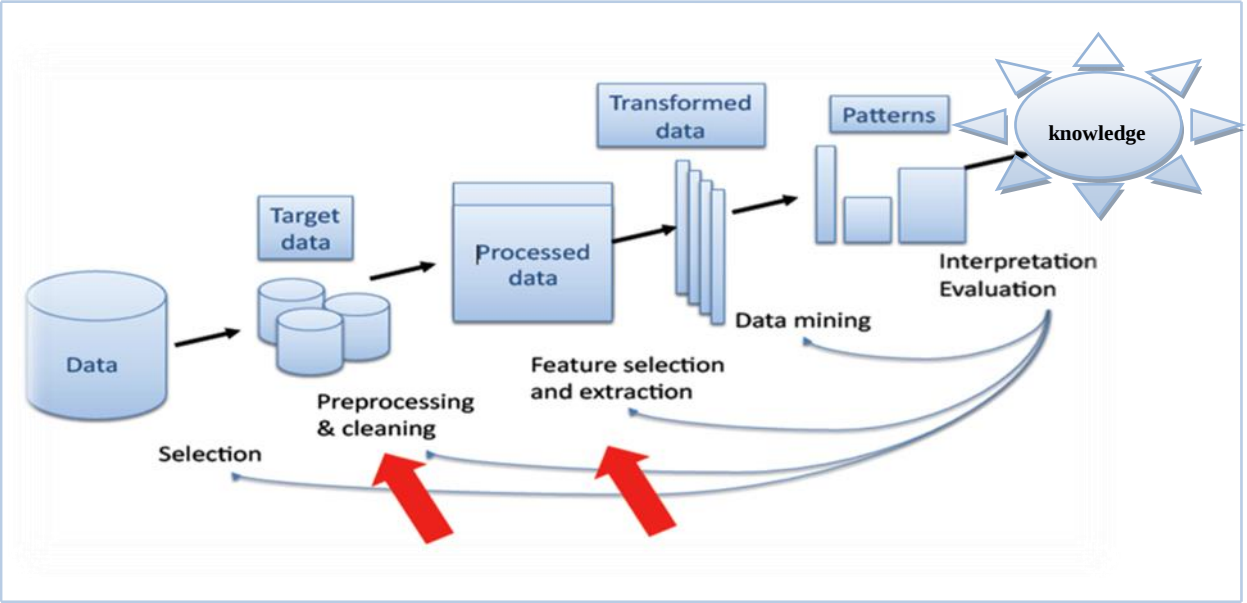



Figure 2.1 knowledge discovery processes

### 2.2.1 Data mining steps:

In the knowledge discovery process are as follows:

1. Data cleansing: The removal of noise and inconsistent data.
2. Data integration: The combination of multiple sources of data.
3. Data selection: The data relevant for analysis is retrieved from the database.
4. Data transformation: The consolidation and transformation of data into forms appropriate for mining.
5. Data mining: The use of intelligent methods to extract patterns from data.
6. Pattern evaluation: Identification of interesting patterns.

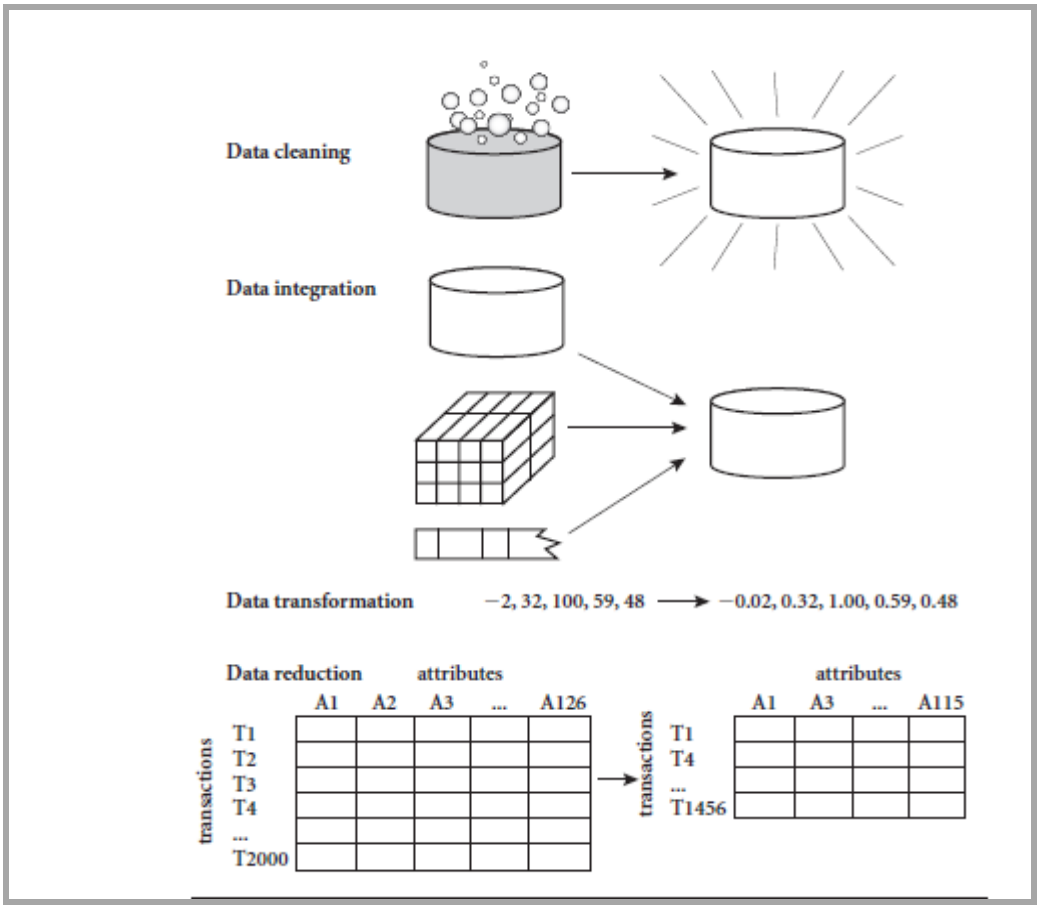



Figure 2.2 data mining steps

### 2.2.1.1 Data cleansing

It is the removal of noise and inconsistent data; it has multi-task such as filling missing values on the data, identification of outliers and smooth out noisy data, correction of inconsistent data, resolving redundancy caused by data integration, data acquisition and metadata, unified date format and converting nominal to numeric.

#### 2.2.1.1.1 Missing data

Missing Values and its problems are very common in the data cleaning process. Several methods have been proposed to process missing data in datasets and avoid problems caused by it.

Missing data may be due to equipment malfunction, inconsistent with other recorded data and thus deleted, data not entered due to misunderstanding, certain data may not be considered important at the time of entry, not register history or changes of the data.

Missing Data handle by ignoring the tuples; it is usually done when the class label is missing. Also, it is handled by filling in the missing value manually, Fill in it automatically with a global constant: e.g., "unknown". Also, it is handled by Imputation by using the attribute mean to fill in the missing value or using the attribute mean for all samples belonging to the same class, in order to fill in the missing value, this is called smattering the attribute mean. And using the most probable value to fill in the missing value: inference-based such as Bayesian formula or decision tree.

#### 2.2.1.1.2 Noisy data

Noise is an unavoidable problem, which affects the data collection, data preparation processes in data mining applications where errors commonly occur; also it's a random error or variance in a measured variable — other data problems which require data cleaning such as duplicate records, incomplete data and inconsistent data.

### 2.2.2 Intelligent methods to extract knowledge

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of the four types of relationships are sought:

- Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order; this information could be used to increase traffic by having daily specials. Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large(Bharati M. Ramageri, 2010). Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, (Bharati M. Ramageri, 2010) make the software that can learn how to classify the data items (Data Mining Techniques).
- Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- **Sequential patterns**: Data is mined to anticipate behaviour patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

### 2.2.3 Classification techniques

Many classification techniques are used in data mining like Decision trees, Nearest neighbour method, Rule induction and Data visualization; a brief description will be given below:

Decision trees are tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments

using chi-square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

Nearest neighbour method: A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) which are most similar to it in a historical dataset (where *k* 1) is called the *k*-nearest neighbour technique .

Rule induction: The extraction of useful if-then rules from data based on statistical significance. Rule induction using sequential covering algorithm, sequential Covering Algorithm can be used to extract IF-THEN rules form the training data, do not require generating a decision tree first. In this algorithm, each rule for a given class covers many of the tuples of that class. Some of the sequential Covering Algorithms are AQ, CN2, and RIPPER. As per the general strategy the rules are learned one at a time. For each time rules are learned, a tuple covered by the rule is removed and the process continues for the rest of the tuples. This is because the path to each leaf in a decision tree corresponds to a rule.

Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

## 2.3 Previous Studies:

Data mining in finance typically follows a set of task steps such as problem understanding, data collection and refining, building a model, model evaluation and deployment. (Adedara, 2012) Data mining approach covers empirical models and regularities derived directly from data and almost only from data with little domain knowledge explicitly involved. Historically, in many domains, deep field-specific theories emerge after the field accumulates enough empirical regularity. As (Mitchell, 1997) perceive that the future of data mining in finance would be to generate more empirical regularities and combine them with domain knowledge via generic analytical data mining approach.

Risk taking capability of a person in the financial market is based on many factors including demographic factors like age, education, marital status etc. In (Sakshi Singh,2014) in order to analyze the various investment instruments used by

the people of different profiles; (Sakshi Singh) have applied fuzzy data mining technique to the demographic factors of a human being. After the division into fuzzy clusters, membership of the person to the clusters is calculated. After the memberships, the derived memberships are used to find the people those have the memberships in the similar range. In further processing the FP growth is applied to find the most recurring patterns. The result was in the form of investment patterns of similar people.

The banking industry has started realizing the need of the techniques like data mining which can help them to compete in the market. Leading banks are using Data Mining (DM) tools for customer segmentation and profitability, credit scoring and approval, predicting payment default, marketing, detecting fraudulent transactions, etc. (Moin K. a., 2012) provides an overview of the concept of DM and highlights the applications of data mining to enhance the performance of some of the core business processes in banking industry.

People of different profiles have different investment strategies as per the humanly attributes. (Joo2, 2004),(Veld, 2008)the work in the paper concerns the analysis of the investment patterns which are affected by the attributes of a person it carries in the given population. The classification of the population into groups as per the profile attributes and then analysis of the patterns in investment needs to be done. For any person, to find the investment patterns of the people in the population, the association of the person to the group of similar people need to be found out and then the patterns above a set threshold to be reported.

The Weka workbench is an organized collection of state of the art machine learning algorithms and data pre-processing tools. The basic way of interacting with these methods is by invoking them from the command line. However, convenient interactive graphical user interfaces are provided for data exploration, for setting up large-scale experiments on distributed computing platforms, and for designing configurations for streamed data processing. These interfaces constitute an advanced environment for experimental data mining. Classification is an important data mining technique with broad applications. It classifies data of various kinds. (Kalmegh, 2015)has been carried out to make a performance evaluation of REP Tree, Simple

Cart and Random Tree classification algorithm. (Kalmegh, 2015) sets out to make comparative evaluation of classifiers REP Tree, Simple Cart and Random Tree in the context of dataset of Indian news to maximize true positive rate and minimize false positive rate. For processing Weka API were used. The results in the paper on dataset of Indian news also show that the efficiency and accuracy of Random Tree is the best one.

Bank direct marketing and business decisions are more important than ever for preserving the relationship with the best customer. The purpose of (Elsalamony, January 2014) was to increase the campaign effectiveness by identifying the main characteristics that affect the success by evaluating and comparing the classification performance of four different data mining techniques' models; MPLNN, TAN, LR and C5.0. Classification was done on the bank direct marketing data set to classify for bank deposit subscription. Experimental results have shown the effectiveness of models. C5.0 has achieved slightly better performance than MLPNN, LR and TAN. Importance analysis has shown that attribute "Duration" in C5.0, LR, and MLPNN models have achieved the most important attribute; however, the attribute Age is the only assessed as more important than the other attributes by TAN.

(Chawan, 2013) described different data mining techniques used in financial data analysis. Financial data analysis is used in many financial institutes for accurate analysis of consumer data to find defaulter and valid customer. In (Study of Data Mining Techniques used for Financial Data Analysis)(Chawan,2013) study about loan default risk analysis, Type of scoring and different data mining techniques like Bayes classification, Decision Tree, Boosting, Bagging, Random forest algorithm and other techniques. The system using data mining for loan Default risk analysis enables the bank to reduce the manual errors involved in the same. Decision trees are preferred by banks because they are a white box model. The discrimination made by decision trees is obvious and the people can understand its working easily. This enables the banks and other financial institutions to provide an account for accepting or rejecting an applicant. Boosting has already increased the efficiency of decision trees. The assessment of risk will enable banks to increase profit and can result in reduction of interest rate.

In South Africa, almost 50% of the people who take loans cannot afford it. Previously, lenders were able to make deductions from a borrower's pay slip but this practice is no longer allowed. Consequently, lenders are far more vulnerable to default particularly if these loans were no longer being backed by any form of meaningful collateral. The aim of (Jonah Mushava, 2018) was to investigate the predictive power of some classification techniques currently in use with specific attention to predicting the propensity for a borrower who are 90 days or more in arrears on an unsecured loan to pay over a fixed window period at least 30% of the total amount due. Results show that these classification techniques perform best for predicting payment patterns over a future horizon period between 3 and 12 months. It is also found that generalized additive models (especially using a generalized extreme value link function), which have not been extensively explored within the credit scoring literature, outperformed all the other classifiers considered in this study. (Jonah Mushava ∗, 2018).

Table 2.1 literature summary:

| **Research paper /book** | **Area** | **Tasks** | **Techniques** | **Results** |
|---|---|---|---|---|
| (Sakshi Singh, 2014) | Financial market | Analyze the various investment instrum-ents. | Clusters (fuzzy) And FP growth (association) | Investment patterns of the similar people. |
| (Moin K. a., 2012) | Leading banks. | Enhance the perfo-rmance of core bus-iness processes in banking industry. | Customer segmentation and profitability, credit scoring and approval, predicting payment default , marketing, detecting fraudulent transactions | Increase performance of business process |
| (Joo2, 2004), (Veld, 2008) | Population | Analysis of the investment patterns | Classification and association techniques | Investment patterns |
| (Kalmegh, 2015) | Indian news Data set. | Comparative evaluation of classifiers | Classification techniques (REP Tree, Simple Cart and Random Tree) | Random Tree is efficient and accurate than |

| | | | | |
|---|---|---|---|---|
| | | REP Tree, Simple Cart and Random Tree | | REPTree, and Simple Cart |
| (Elsalamony, January 2014) | Bank direct marketing data set | Increasing the campaign effectiveness to classify for bank deposit subscription. | Classification techniques (MPLNN,TAN, LR and C5.0) | That most important is attribute "Duration" in C5.0, LR, and MLPNN, the attribute Age is assessed as more important than the other attributes by TAN |
| (Chawan, 2013) | Banks | Study about loan default risk analysis, Type of scoring and different data mining techniques | Scoring Bayes classification, Decision Tree, Boosting, Bagging and Random forest algorithm. | Enables the bank to reduce the manual errors involved. Decision trees are preferred |
| (Jonah Mushava ∗, 2018) | South Africa's unsecured lending market | Investigate the predictive power of classification techniques | Logistic regression, binary generalized extreme value additive model, Linear discriminant analysis, generalized LDA, Random forests (RF) | Classification techniques perform best for predicting payment |

## 2.4 Summary:

After providing an overview of the exploration of the data and addressing the previous studies related, it was evident that the majority used the classification technique as apparent in table 2.1. The steps of prospecting in the data will be implemented in the next section.

# CHAPTER THREE

# Methodology

## 3.1 Introduction

This chapter contains the main steps of implementation which are used in the research, figure (3.1) represent the steps used.

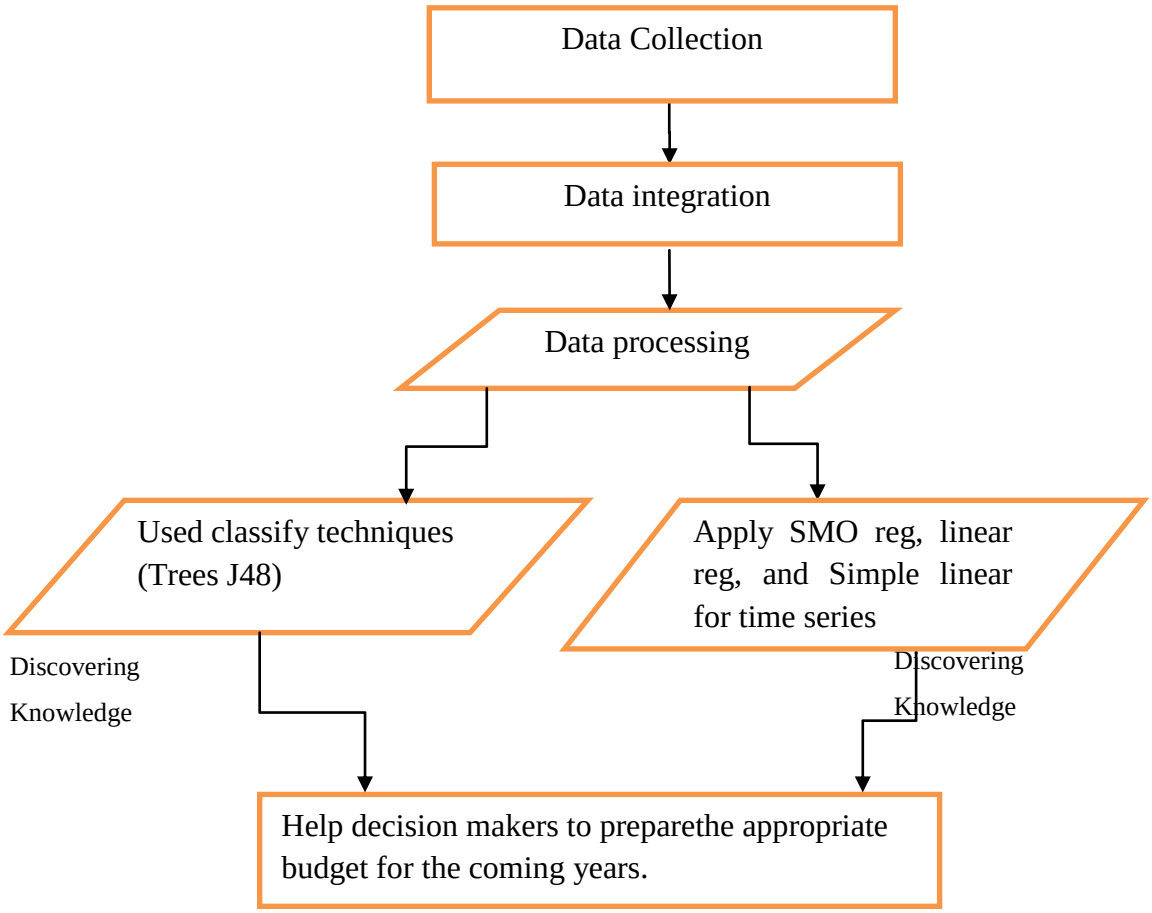



Figure 3.1 MethodologySteps

The researcher started by collecting data from the Oracle database of the National Pension and Insurance Fund (Government Sector) which contains different databases of the employees, adminstrative, departments, pensioners, the loans and investments. The data of loans and investment was chosen, the data was transformed into an excel file format.The data of loans contains 34.879 records, where each row has ten attributes; Pensioners ID, Gender, State ,State number, State of paid, Date of loan, Amount of loan, Type of loan, Reason, Reason ID. This attributed data is illustrated in figure 3.2.

| رقم الملف | رقم الولايه | تاريخ السلفيه | المبلغ | نوع السلفيه | النوع | الوحدة | السبب رقم | السبب | الولايه(الصرف) |
|---|---|---|---|---|---|---|---|---|---|
| 11283151 | 31 | 27/01/10 | 250 | 0 | 0 | التخطيط العمراني | 2 | بلوغ السن | الخرطوم |
| 10332721 | 31 | 27/01/10 | 250 | 0 | 0 | التربية والتعليم | 3 | إختيارى | الخرطوم |
| 11228121 | 31 | 27/01/10 | 250 | 0 | 0 | الاحصاء | 2 | بلوغ السن | الخرطوم |
| 11231571 | 31 | 27/01/10 | 250 | 0 | 1 | مستشفى ام درمان | 2 | بلوغ السن | الخرطوم |
| 11260201 | 31 | 27/01/10 | 250 | 0 | 0 | الصحة الإتحادية | 2 | بلوغ السن | الخرطوم |
| 20358721 | 31 | 27/01/10 | 200 | 0 | 1 | الصحة الإتحادية | 11 | وفاة بالخدمة | الخرطوم |
| 20358541 | 31 | 27/01/10 | 250 | 0 | 0 | السكة حديد | 21 | وفاة بالمعاش | الخرطوم |
| 11184291 | 31 | 27/01/10 | 250 | 0 | 0 | مستشفى سوبا | 2 | بلوغ السن | الخرطوم |
| 11068031 | 31 | 27/01/10 | 250 | 0 | 0 | التربية والتعليم | 2 | بلوغ السن | الخرطوم |
| 10929111 | 31 | 27/01/10 | 250 | 0 | 0 | الطباعة والنشر | 2 | بلوغ السن | الخرطوم |
| 10933971 | 31 | 27/01/10 | 250 | 0 | 0 | النقل النهرى | 2 | بلوغ السن | الخرطوم |
| 10929251 | 31 | 27/01/10 | 250 | 0 | 0 | مياة ولاية الخرطوم | 7 | فصل | الخرطوم |
| 10948221 | 31 | 27/01/10 | 250 | 0 | 0 | مياه ولاية الخرطوم | 3 | إختيارى | الخرطوم |
| 10931101 | 31 | 27/01/10 | 250 | 0 | 1 | النقل الميكانيكى | 4 | إلغاء وظيفة | الخرطوم |
| 10937311 | 31 | 27/01/10 | 250 | 0 | 0 | مياه ولاية الخرطوم | 2 | بلوغ السن | الخرطوم |
| 10937111 | 31 | 27/01/10 | 250 | 0 | 1 | محلية الشهداء | 2 | بلوغ السن | الخرطوم |
| 10950981 | 31 | 27/01/10 | 250 | 0 | 1 | محلية الامير | 3 | إختيارى | الخرطوم |
| 10928831 | 31 | 27/01/10 | 250 | 0 | 0 | البريد والبرق | 2 | بلوغ السن | الخرطوم |
| 10697331 | 31 | 27/01/10 | 250 | 0 | 0 | الزراعة | 2 | بلوغ السن | الخرطوم |
| 10684031 | 31 | 27/01/10 | 250 | 0 | 0 | التربية | 8 | إستقالة | الخرطوم |
| 10738561 | 31 | 27/01/10 | 250 | 0 | 0 | بلدية الخرطوم | 2 | بلوغ السن | الخرطوم |
| 20114131 | 31 | 27/01/10 | 250 | 0 | 0 | مجلس منطقة الخرطوم | 2 | بلوغ السن | الخرطوم |
| 10304871 | 31 | 27/01/10 | 250 | 0 | 1 | التربية | 3 | إختيارى | الخرطوم |

Figure 3.2 loans data after transformation from the Oracle database

The investments database contains (12.380) record and has 12 attributes; Pensioners ID, Sector, Gender, State, StateID, Date of loan, Date of implement loan ,Amount of loan, Amount of loan with profit, Type of loan, Reason, Reason ID. Figure 3.3 illustrates part of the data set.

| المبلغ | الولايه رقم | تاريخ التنفيذ | تاريخ التقديم | القطاع | الاجمالي | رقم الملف | الولايه | سبب التقاعد | سبب التقاعد | الوحده | النوع |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3050 | 31 | 7/2/2011 | 3/2/2011 | الخدمى | 3599 | 10055571 | الخرطوم | 2 | بلوغ السن | الصحة | 0 |
| 4700 | 31 | 9/5/2012 | 18/04/12 | الخدمى | 5452 | 10070201 | الخرطوم | 6 | تلجين | التربية والتعليم | 0 |
| 3000 | 31 | 24/06/10 | 6/5/2010 | مشاركة المزارع تجارى | 3540 | 10069471 | الخرطوم | 6 | تلجين | التربية والتعليم | 0 |
| 11400 | 51 | 3/11/2017 | 9/10/2017 | القطاع الخدمي | 13110 | 10742461 | شمال كردفان | 3 | إختيارى | التربية والتعليم | 0 |
| 3750 | 31 | 28/04/10 | 22/03/10 | الخدمى | 4425 | 20158401 | الخرطوم | 3 | إختيارى | التربية | 0 |
| 4950 | 31 | 2/9/2012 | 21/05/12 | الخدمى | 5742 | 10090851 | الخرطوم | 7 | فصل | التربية والتعليم | 0 |
| 9650 | 31 | 1/6/2015 | 26/04/15 | صيانة وبناء المنازل | 11099 | 10090871 | الخرطوم | 2 | بلوغ السن | التربية والتعليم | 1 |
| 9500 | 31 | 1/5/2015 | 2/3/2015 | الخدمى | 10925 | 10091171 | الخرطوم | 7 | فصل | التربية والتعليم | 0 |
| 2700 | 31 | 11/10/2012 | 9/10/2012 | الخدمى | 2916 | 10092131 | الخرطوم | 3 | إختيارى | التربية والتعليم | 0 |
| 1950 | 31 | 6/3/2010 | 22/02/10 | الخدمى | 2126 | 10092131 | الخرطوم | 3 | إختيارى | التربية والتعليم | 0 |
|  | 43 |  | 12/3/2017 | القطاع التجاري |  | 10092691 | النيل الأبيض | 4 | إلغاء وظيفة | الزراعة | 0 |
|  | 43 |  | 21/09/16 | القطاع التجاري |  | 10092691 | النيل الأبيض | 4 | إلغاء وظيفة | الزراعة | 0 |
| 14100 | 31 | 12/10/2017 | 26/09/17 | القطاع الخدمي | 16215 | 10139641 | الخرطوم | 3 | إختيارى | وزارة التربية | 1 |
| 0 | 31 |  | 12/3/2014 | احتياجات منزلية | 0 | 10139641 | الخرطوم | 3 | إختيارى | وزارة التربية | 1 |
| 3600 | 31 | 23/12/10 | 8/12/2010 | الخدمى | 4248 | 10139641 | الخرطوم | 3 | إختيارى | وزارة التربية | 1 |
| 9500 | 31 | 3/9/2015 | 31/08/15 | صيانة وبناء المنازل | 10925 | 10138691 | الخرطوم | 3 | إختيارى | إدارة القوى العاملة | 0 |
| 4008 | 31 | 2/6/2010 | 22/05/10 | التجارى | 4729.44 | 10138691 | الخرطوم | 3 | إختيارى | إدارة القوى العاملة | 0 |
| 2650 | 31 | 21/06/10 | 19/04/10 | التجارى | 3127 | 10075271 | الخرطوم | 7 | فصل | الحكم الشعبى المحلى | 0 |
| 4000 | 41 | 5/1/2017 | 11/12/2016 | القطاع الخدمي | 4000 | 10089711 | الجزيرة | 7 | فصل | التربية والتعليم | 0 |
| 9500 | 31 | 9/12/2015 | 18/11/15 | القطاع التجاري | 10925 | 10090071 | الخرطوم | 7 | فصل | التربية والتعليم | 0 |
| 5000 | 31 | 9/9/2012 | 24/07/12 | الخدمى | 5800 | 10090071 | الخرطوم | 7 | فصل | التربية والتعليم | 0 |
| 3350 | 31 | 8/8/2010 | 29/07/10 | مشاركة المزارع تجارى | 3953 | 10086871 | الخرطوم | 7 | فصل | التربية | 0 |
| 5000 | 31 | 13/09/15 | 6/9/2015 | الخدمى | 5450 | 10086961 | الخرطوم | 7 | فصل | التربية | 0 |
| 3000 | 31 | 7/3/2011 | 20/02/11 | الخدمى | 3540 | 20159851 | الخرطوم | 2 | بلوغ السن | الارشاد الفوري | 0 |

Figure 3.3 investment data after transformation from Oracle data base

## 3.2 Data preprocessing:

Data were preprocessed following the procedures explained in the previous chapter. The choice of deleting the attribute didn't help in achieving the research objectives (Amount of loan with profit) and remove redundancy (Reason, Reason ID, State,State ID). Nominal data was converted into numeric data in order to conform with the Weka tool environment. Figure 3.4 demonstrate changing the unit elements to numbers and figure 3.5 iluustrate the changing of sectors elements to numbers. Figures 3.6, 3.7 display the transformed loans and investments data respectively.

| A | B |
|---|---|
| الوحدة | unit |
| استثنائى | 0 |
| 1 | 1 |
| الخدمات الصحية | 1 |
| الصحه | 1 |
| وزارة الصحه | 1 |
| الصحه الشماليه | 1 |
| الصحه سنار | 1 |
| الصحه نهر النيل | 1 |
| الصحه غ كردفان | 1 |
| الصحه | 1 |
| الصحه كسلا | 1 |
| سنار الصحه | 1 |
| الصحه الجزيره | 1 |
| الجزيره الصحه | 1 |
| الصحه الولائية | 1 |
| وزارة الصحه الاتحادية | 1 |
| الصحه الخرطوم | 1 |
| الصحه ولايه الخرطوم | 1 |
| 2 | 2 |
| التربية والتعليم | 2 |
| التربة والتعليم | 2 |
| التعليم | 2 |
| التربي كردفان | 2 |

Figure 3.4 converting unit's type

| A | B |
|---|---|
| Sector | القطاع |
| 1 | الخدمى |
| 1 | الخدمى |
| 21 | مشاركة المزارع تجارى |
| 1 | القطاع الخدمي |
| 1 | الخدمى |
| 1 | الخدمى |
| 3 | صيانة وبناء المنازل |
| 1 | الخدمى |
| 1 | الخدمى |
| 1 | الخدمى |
| 4 | القطاع التجاري |
| 4 | القطاع التجاري |
| 1 | القطاع الخدمي |
| 8 | احتياجات منزلية |
| 1 | الخدمى |
| 3 | صيانة وبناء المنازل |
| 4 | التجارى |
| 4 | التجارى |
| 1 | القطاع الخدمي |
| 4 | القطاع التجاري |
| 1 | الخدمى |
| 21 | مشاركة المزارع تجارى |
| 1 | الخدمى |

Figure 3.5 converting sectore's type

| A | B | C | D | E | F |
|---|---|---|---|---|---|
| Pensioners ID loans | Date of loans | Type of loans | Uit of loan | Loans amount | Gender |
| 10069471 | 12/23/2012 | 0 | 2 | 300 | 0 |
| 10069471 | 1/19/2014 | 0 | 2 | 1000 | 0 |
| 10069471 | 3/7/2014 | 0 | 2 | 300 | 0 |
| 10069471 | 3/22/2015 | 1 | 2 | 1500 | 0 |
| 10069471 | 6/19/2016 | 1 | 2 | 2000 | 0 |
| 10070201 | 5/18/2011 | | 2 | 250 | 0 |
| 10070201 | 12/16/2012 | 0 | 2 | 300 | 0 |
| 10070201 | 4/23/2015 | 1 | 2 | 1500 | 0 |
| 10070201 | 9/22/2016 | 1 | 2 | 2000 | 0 |
| 10070201 | 10/17/2017 | 1 | 2 | 3000 | 0 |
| 10070501 | 7/21/2015 | 1 | 2 | 1500 | 0 |
| 10071981 | 8/22/2017 | 1 | 2 | 3000 | 0 |
| 10075271 | 3/24/2010 | 0 | 6 | 250 | 0 |
| 10075401 | 4/6/2015 | 1 | 2 | 1500 | 0 |
| 10075401 | 7/24/2016 | 1 | 2 | 2000 | 0 |
| 10082421 | 6/7/2014 | 0 | 4 | 1000 | 0 |
| 10082421 | 2/14/2016 | 1 | 4 | 2000 | 0 |
| 10082421 | 3/16/2017 | 1 | 4 | 3000 | 0 |
| 10084161 | 2/28/2013 | 0 | 460 | 350 | 0 |
| 10085151 | 10/17/2016 | 1 | 3 | 2000 | 0 |
| 10085771 | 7/15/2015 | 1 | 2 | 1500 | 0 |
| 10085771 | 4/8/2016 | 1 | 2 | 2000 | 0 |
| 10085771 | 3/10/2017 | 1 | 2 | 2000 | 0 |
| 10087211 | 3/6/2010 | | 462 | 250 | 0 |
| 10087211 | 9/14/2011 | | 462 | 250 | 0 |
| 10087211 | 11/11/2012 | 0 | 462 | 300 | 0 |

ass22

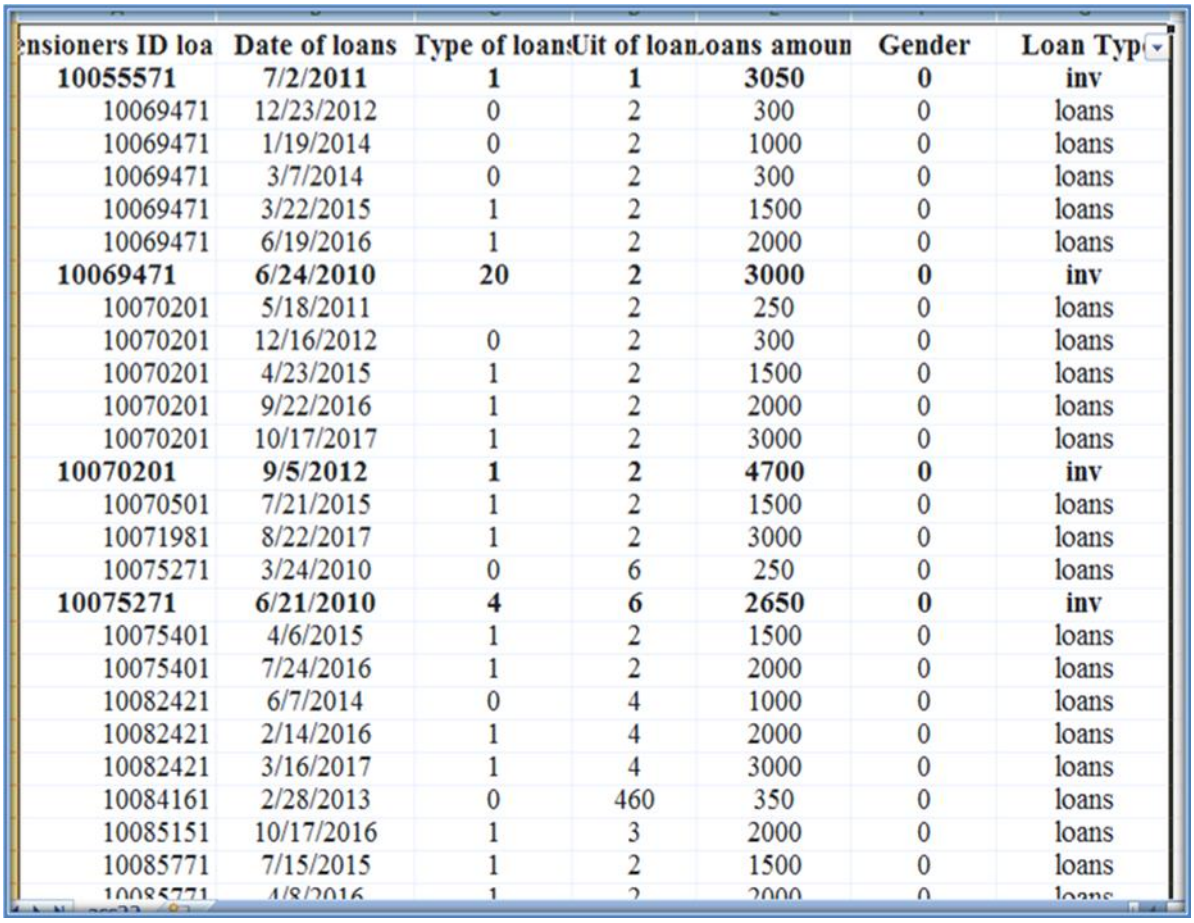
Figure3.6 the loans data after preprocessing

| Pensioners ID loans | Date of loans | Type of loans | Uit of loan | Loans amount | Gender |
|---|---|---|---|---|---|
| 10055571 | 7/2/2011 | 1 | 1 | 3050 | 0 |
| 10069471 | 6/24/2010 | 20 | 2 | 3000 | 0 |
| 10070201 | 9/5/2012 | 1 | 2 | 4700 | 0 |
| 10075271 | 6/21/2010 | 4 | 6 | 2650 | 0 |
| 10085771 | 5/30/2013 | 8 | 2 | 3500 | 0 |
| 10085771 | 10/8/2010 | 200 | 2 | 3350 | 0 |
| 10086871 | 8/8/2010 | 20 | 2 | 3350 | 0 |
| 10086961 | 9/13/2015 | 1 | 2 | 5000 | 0 |
| 10087211 | 12/18/2017 | 60 | 462 | 5000 | 0 |
| 10087211 | 1/6/2016 | 60 | 462 | 4000 | 0 |
| 10087211 | 2/6/2015 | 8 | 462 | 3500 | 0 |
| 10087211 | 4/14/2011 | 1 | 462 | 1727 | 0 |
| 10087851 | 12/12/2011 | 1 | 2 | 3144 | 0 |
| 10088931 | 6/13/2015 | 10 | 2 | 8800 | 0 |
| 10089331 | 1/12/2016 | 1 | 9 | 10450 | 0 |
| 10089491 | 3/21/2011 | 4 | 3 | 3350 | 0 |
| 10089541 | 11/8/2015 | 8 | 4 | 3500 | 0 |
| 10089711 | 5/1/2017 | 1 | 2 | 4000 | 0 |
| 10090071 | 9/12/2015 | 4 | 2 | 9500 | 0 |
| 10090071 | 9/9/2012 | 1 | 2 | 5000 | 0 |
| 10090701 | 12/15/2010 | 4 | 1000 | 3700 | 0 |
| 10090851 | 2/9/2012 | 1 | 2 | 4950 | 0 |
| 10090871 | 1/6/2015 | 3 | 2 | 9650 | 1 |
| 10091171 | 1/5/2015 | 1 | 2 | 9500 | 0 |
| 10092131 | 11/10/2012 | 1 | 2 | 2700 | 0 |
| 10092131 | 6/3/2010 | 1 | 2 | 1950 | 0 |

Figure 3.7 the investment data after preprocessing

## 3.3 Data Integration:

Loans data with investment data were merged, and a new attribute named loan type (loan or investment) was added. The number of attributes became seven attributes (Pensioners ID- loans, Date of loans, Type- loans, Unit_loan, Loan Type, Loans amount, Gender). The total instances equal 47.257 of the merged investment &loans databases. Figure 3.8 express all attributes in the correct type format.



| ensioners ID loa | Date of loans | Type of loan | Uit of loan | oans amoun | Gender | Loan Typ |
|---|---|---|---|---|---|---|
| **10055571** | **7/2/2011** | **1** | **1** | **3050** | **0** | **inv** |
| 10069471 | 12/23/2012 | 0 | 2 | 300 | 0 | loans |
| 10069471 | 1/19/2014 | 0 | 2 | 1000 | 0 | loans |
| 10069471 | 3/7/2014 | 0 | 2 | 300 | 0 | loans |
| 10069471 | 3/22/2015 | 1 | 2 | 1500 | 0 | loans |
| 10069471 | 6/19/2016 | 1 | 2 | 2000 | 0 | loans |
| **10069471** | **6/24/2010** | **20** | **2** | **3000** | **0** | **inv** |
| 10070201 | 5/18/2011 | | 2 | 250 | 0 | loans |
| 10070201 | 12/16/2012 | 0 | 2 | 300 | 0 | loans |
| 10070201 | 4/23/2015 | 1 | 2 | 1500 | 0 | loans |
| 10070201 | 9/22/2016 | 1 | 2 | 2000 | 0 | loans |
| 10070201 | 10/17/2017 | 1 | 2 | 3000 | 0 | loans |
| **10070201** | **9/5/2012** | **1** | **2** | **4700** | **0** | **inv** |
| 10070501 | 7/21/2015 | 1 | 2 | 1500 | 0 | loans |
| 10071981 | 8/22/2017 | 1 | 2 | 3000 | 0 | loans |
| 10075271 | 3/24/2010 | 0 | 6 | 250 | 0 | loans |
| **10075271** | **6/21/2010** | **4** | **6** | **2650** | **0** | **inv** |
| 10075401 | 4/6/2015 | 1 | 2 | 1500 | 0 | loans |
| 10075401 | 7/24/2016 | 1 | 2 | 2000 | 0 | loans |
| 10082421 | 6/7/2014 | 0 | 4 | 1000 | 0 | loans |
| 10082421 | 2/14/2016 | 1 | 4 | 2000 | 0 | loans |
| 10082421 | 3/16/2017 | 1 | 4 | 3000 | 0 | loans |
| 10084161 | 2/28/2013 | 0 | 460 | 350 | 0 | loans |
| 10085151 | 10/17/2016 | 1 | 3 | 2000 | 0 | loans |
| 10085771 | 7/15/2015 | 1 | 2 | 1500 | 0 | loans |
| 10085771 | 4/8/2016 | 1 | 2 | 2000 | 0 | loans |

Figure 3.8 Merge of investments and loans

Finally, the data became ready for implementing Weka tools. Filtered data by replacing all missing values to nominal and numerical attributes in a dataset by the modes and means of the training data. Figure 39 and table 3.1explain the data in details.

| No. | 1: Pensioners ID loans | 2: Date of loans | 3: Type of loans | 4: Uit of loan | 5: Loan Type | 6: Loans amount | 7: **Gender** |
|---|---|---|---|---|---|---|---|
| | Numeric | Nominal | Numeric | Numeric | Nominal | Numeric | Numeric |
| 1 | 1.0069471E7 | 12/23/2012 | 0.0 | 2.0 | loans | 300.0 | 0.0 |
| 2 | 1.0069471E7 | 1/19/2014 | 0.0 | 2.0 | loans | 1000.0 | 0.0 |
| 3 | 1.0069471E7 | 3/7/2014 | 0.0 | 2.0 | loans | 300.0 | 0.0 |
| 4 | 1.0069471E7 | 3/22/2015 | 1.0 | 2.0 | loans | 1500.0 | 0.0 |
| 5 | 1.0069471E7 | 6/19/2016 | 1.0 | 2.0 | loans | 2000.0 | 0.0 |
| 6 | 1.0070201E7 | 5/18/2011 | | 2.0 | loans | 250.0 | 0.0 |
| 7 | 1.0070201E7 | 12/16/2012 | 0.0 | 2.0 | loans | 300.0 | 0.0 |
| 8 | 1.0070201E7 | 4/23/2015 | 1.0 | 2.0 | loans | 1500.0 | 0.0 |
| 9 | 1.0070201E7 | 9/22/2016 | 1.0 | 2.0 | loans | 2000.0 | 0.0 |
| 10 | 1.0070201E7 | 10/17/2017 | 1.0 | 2.0 | loans | 3000.0 | 0.0 |
| 11 | 1.0070501E7 | 7/21/2015 | 1.0 | 2.0 | loans | 1500.0 | 0.0 |
| 12 | 1.0071981E7 | 8/22/2017 | 1.0 | 2.0 | loans | 3000.0 | 0.0 |
| 13 | 1.0075271E7 | 3/24/2010 | 0.0 | 6.0 | loans | 250.0 | 0.0 |
| 14 | 1.0075401E7 | 4/6/2015 | 1.0 | 2.0 | loans | 1500.0 | 0.0 |
| 15 | 1.0075401E7 | 7/24/2016 | 1.0 | 2.0 | loans | 2000.0 | 0.0 |
| 16 | 1.0082421E7 | 6/7/2014 | 0.0 | 4.0 | loans | 1000.0 | 0.0 |
| 17 | 1.0082421E7 | 2/14/2016 | 1.0 | 4.0 | loans | 2000.0 | 0.0 |
| 18 | 1.0082421E7 | 3/16/2017 | 1.0 | 4.0 | loans | 3000.0 | 0.0 |
| 19 | 1.0084161E7 | 2/28/2013 | 0.0 | 460.0 | loans | 350.0 | 0.0 |
| 20 | 1.0085151E7 | 10/17/2016 | 1.0 | 3.0 | loans | 2000.0 | 0.0 |
| 21 | 1.0085771E7 | 7/15/2015 | 1.0 | 2.0 | loans | 1500.0 | 0.0 |
| 22 | 1.0085771E7 | 4/8/2016 | 1.0 | 2.0 | loans | 2000.0 | 0.0 |
| 23 | 1.0085771E7 | 3/10/2017 | 1.0 | 2.0 | loans | 2000.0 | 0.0 |
| 24 | 1.0087211E7 | 3/6/2010 | | 462.0 | loans | 250.0 | 0.0 |
| 25 | 1.0087211E7 | 9/14/2011 | | 462.0 | loans | 250.0 | 0.0 |
| 26 | 1.0087211E7 | 11/11/2012 | 0.0 | 462.0 | loans | 300.0 | 0.0 |
| 27 | 1.0087211E7 | 6/4/2014 | 0.0 | 462.0 | loans | 1000.0 | 0.0 |
| 28 | 1.0088461E7 | 3/11/2013 | 0.0 | 2.0 | loans | 500.0 | 0.0 |
| 29 | 1.0088561E7 | 2/22/2012 | | 2.0 | loans | 300.0 | 0.0 |
| 30 | 1.0088561E7 | 12/23/2012 | 0.0 | 2.0 | loans | 300.0 | 0.0 |
| 31 | 1.0088561E7 | 4/7/2013 | 2.0 | 2.0 | loans | 400.0 | 0.0 |
| 32 | 1.0088561E7 | 9/17/2013 | 1.0 | 2.0 | loans | 2890.0 | 0.0 |
| 33 | 1.0088561E7 | 8/17/2014 | 0.0 | 2.0 | loans | 1000.0 | 0.0 |
| 34 | 1.0088561E7 | 2/7/2015 | 1.0 | 2.0 | loans | 1500.0 | 0.0 |
| 35 | 1.0088561E7 | 7/18/2016 | 2.0 | 2.0 | loans | 3600.0 | 0.0 |
| 36 | 1.0088681E7 | 2/29/2012 | | 3.0 | loans | 300.0 | 0.0 |
| 37 | 1.0088681E7 | 10/3/2015 | 1.0 | 3.0 | loans | 1500.0 | 0.0 |
| 38 | 1.0088761E7 | 10/3/2010 | 0.0 | 2.0 | loans | 250.0 | 0.0 |

Figure 3.9 data with missing value

Table 3.1 noisy data

| **Name** | **Pensioner ID loans** | **Date of loans** | **Type of loans** | **Unit of loan** | **Loan type** | **Loan amount** | **Gender** |
|---|---|---|---|---|---|---|---|
| Type | Numeric | Nominal | Numeric | Numeric | Nominal | Numeric | Numeric |
| Missing | 0 | 1900(4%) | 1366 (3%) | 8112(17%) | 0 | 1436(3%) | 0(0%) |
| Distinct | 26660 | 1879 | 23 | 677 | 2 | 1003 | 2 |
| Unique | 15206 (32%) | 175(0%) | 1(0%) | 125(0%) | 0 | 469(1%) | 0(0%) |

## 3.4 Classification Techniques:

Then use Classification techniques to classify the database of the training set and the values (class labels) in a classifying attribute (Type of loan) and used it in classifying new data to build the model. Testing by choosing Cross Validation option (66%)with trees j48, as explained in figure 3.10.And testing by choosing Percentage

Split options (66%)with trees j48,the Type of loan represents the classifier, as explained in figure 3.11.

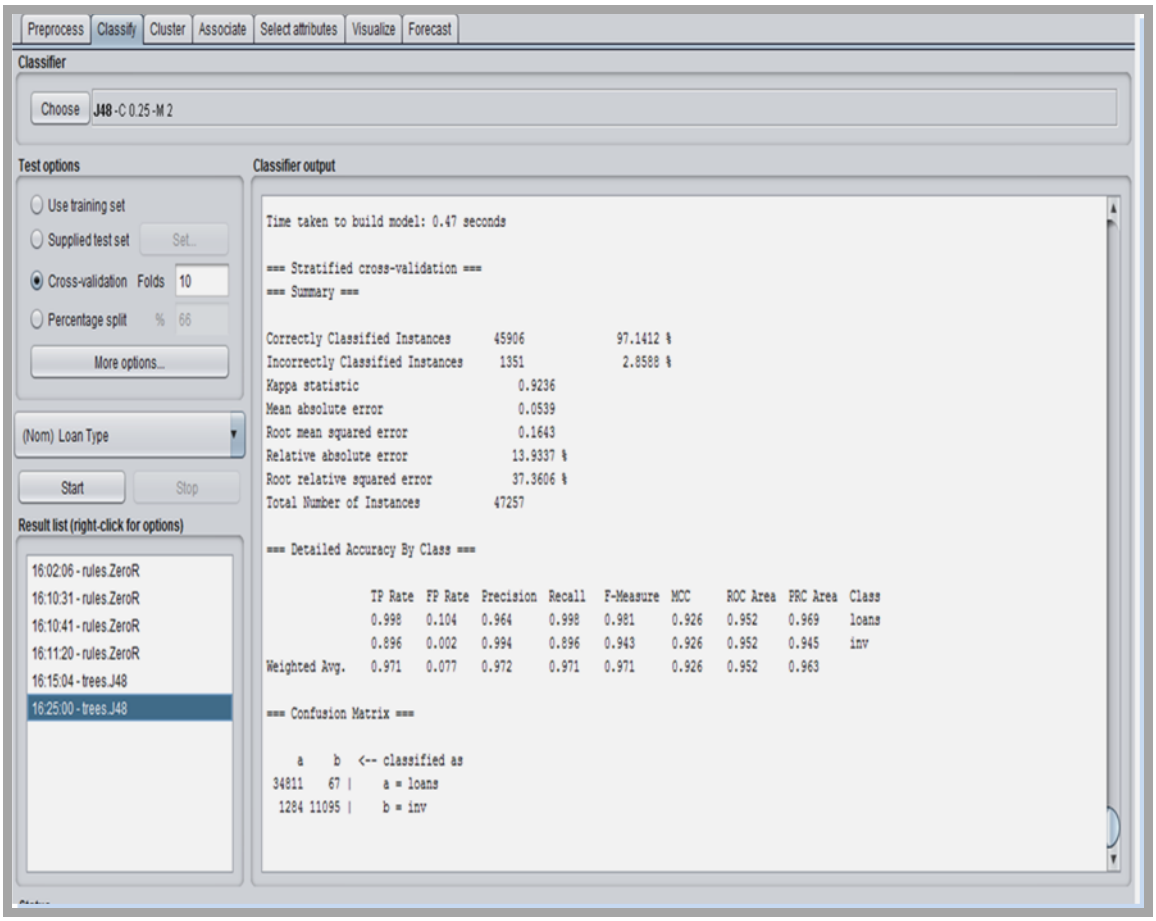



Figure 3.10 Testing by Cross-validation Split options -trees j48

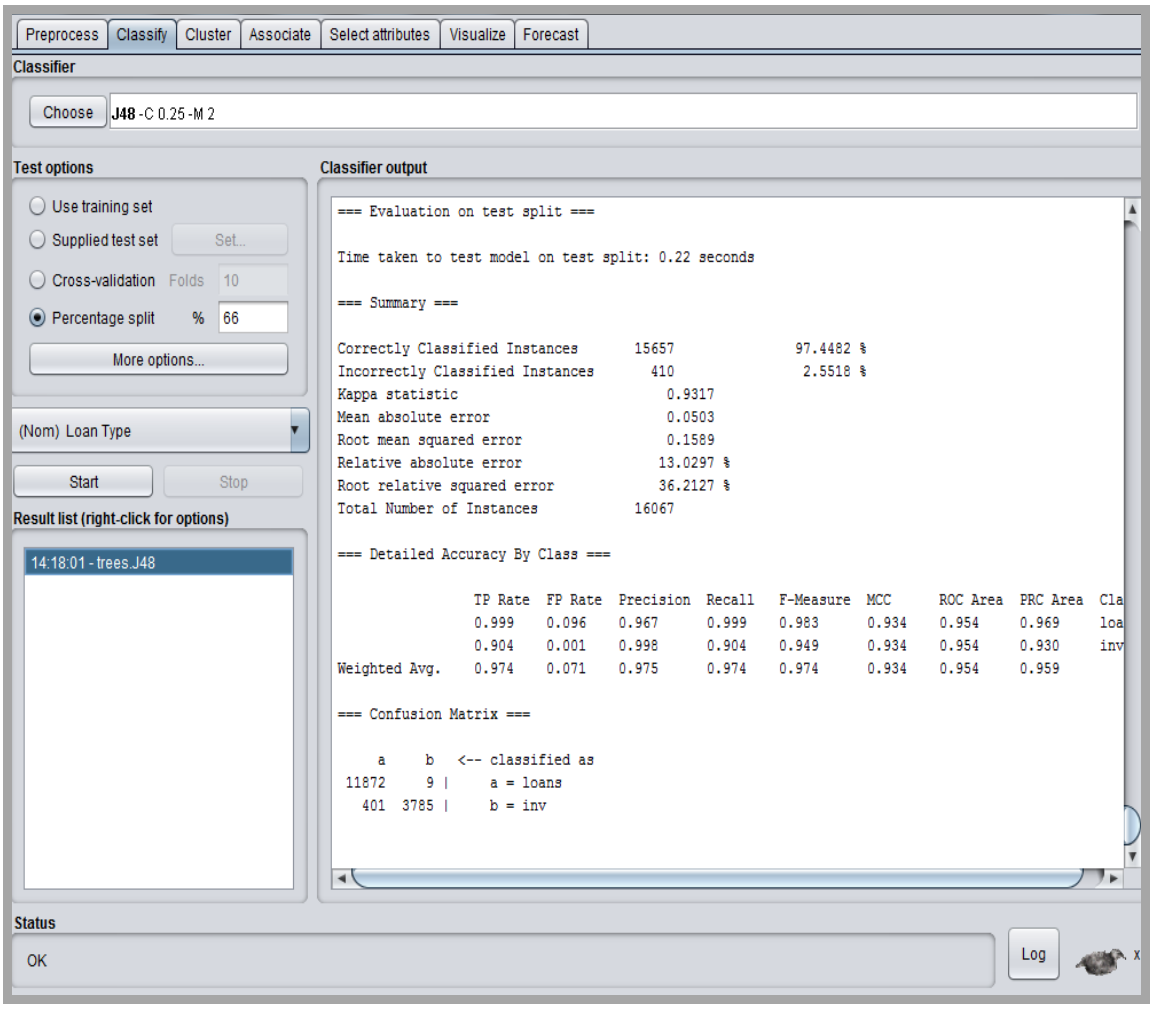



Figure 3.11 Testing by using Percentage split options - TreesJ48

## 3.5 Applying regression

Because the nature of the data contains series dates, the researcher has generated two new dataset that depends on three attribute (pensioners ID, date of loans and loan amount) as explained in figure 3.12,so new data were created; as shown in table 3.2.

| F | G | H | I |
|---|---|---|---|
| Pensioners ID loan | Date of loans | Loans amount | |
| 10069471 | 12/23/2012 | 300 | |
| 10069471 | 1/19/2014 | 1000 | |
| 10069471 | 3/7/2014 | 300 | |
| 10069471 | 3/22/2015 | 1500 | |
| 10069471 | 6/19/2016 | 2000 | |
| 10070201 | 5/18/2011 | 250 | |
| 10070201 | 12/16/2012 | 300 | |
| 10070201 | 4/23/2015 | 1500 | |
| 10070201 | 9/22/2016 | 2000 | |
| 10070201 | 10/17/2017 | 3000 | |
| 10070501 | 7/21/2015 | 1500 | |
| 10071981 | 8/22/2017 | 3000 | |
| 10075271 | 3/24/2010 | 250 | |
| 10075401 | 4/6/2015 | 1500 | |
| 10075401 | 7/24/2016 | 2000 | |
| 10082421 | 6/7/2014 | 1000 | |
| 10082421 | 2/14/2016 | 2000 | |
| 10082421 | 3/16/2017 | 3000 | |
| 10084161 | 2/28/2013 | 350 | |
| 10085151 | 10/17/2016 | 2000 | |
| 10085771 | 7/15/2015 | 1500 | |
| 10085771 | 4/8/2016 | 2000 | |
| 10085771 | 3/10/2017 | 2000 | |
| 10087211 | 3/6/2010 | 250 | |
| 10087211 | 9/14/2011 | 250 | |

Figure 3.12 main attribute to predict number of borrowers and budget

Table 3.2 Amount of money, number of borrower over years2007-2017

| Years | Borrowers | Amount of money |
|---|---|---|
| 2007 | - | 24330 |
| 2008 | - | 25951 |
| 2009 | - | 27864 |
| 2010 | 1684 | 2937677 |
| 2011 | 1364 | 2574792 |
| 2012 | 2198 | 4489164 |
| 2013 | 2845 | 5325592 |
| 2014 | 3853 | 6487326 |
| 2`015 | 10958 | 31158315 |
| 2016 | 10918 | 42682559 |
| 2017 | 11807 | 56304844 |

Then separate data , one contained number of years 2007-2017 for amount of money, as explained in figure 3.13.

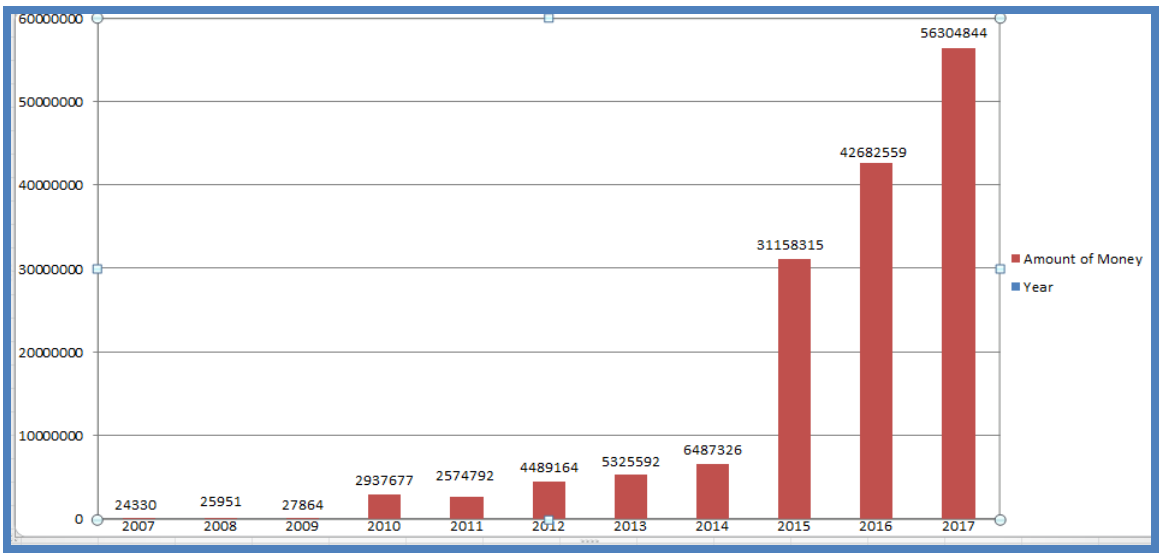



Figure 3.13Amount of money over years (2007-2017)

The other one contained the years from 2010 to 2017forborrowers, as explained in figure 3.14.

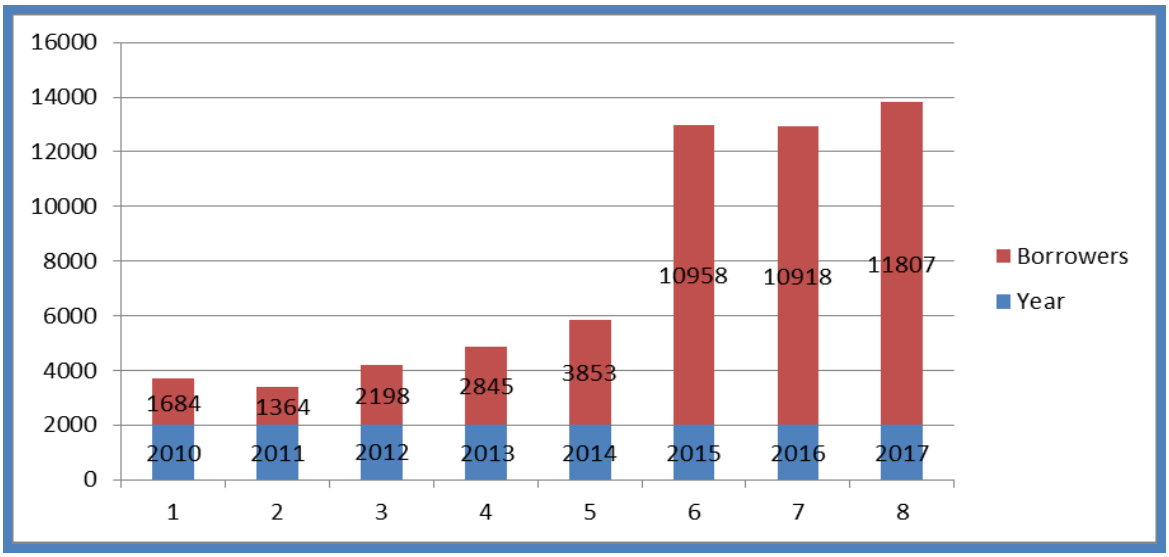



Figure 3.14Number of borrowers over years (2010-2017)

After that time series forecasting was used by applying the SMO regression, linear regression and simple linear regression. Time Series forecasting tools was installed from Package Manager in the tools to forecast the number of the borrowers' in 2018-2019, and amount of money which decision maker need to borrow pensioners in 2018-2019.

Sequential Minimal Optimization (SMO) regression was used to predict the amount of money for years 2018-2019, as explained in figure 3.15.

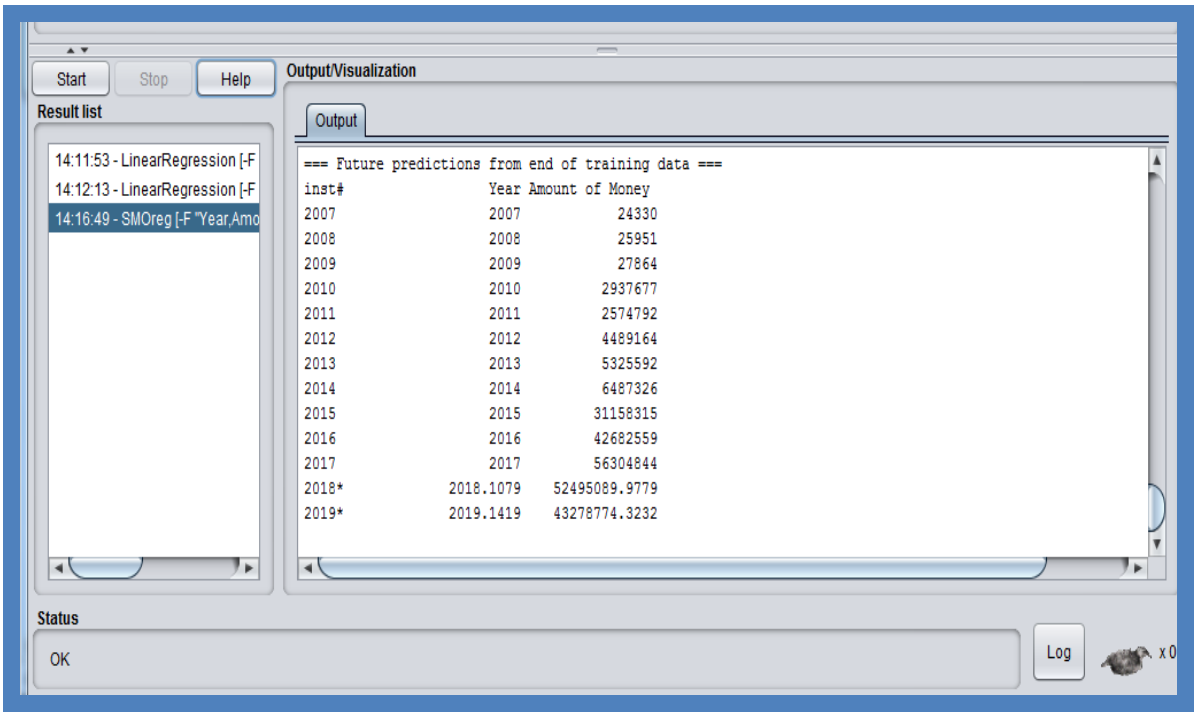



Figure 3.15 predicting amount of money 2018-2019(SMO regression)

Forecasting the amount of money for years 2018-2019 from years (2007 to 2017) was performed by using linear regression with confidence 95%, as explained in figure 3.16.

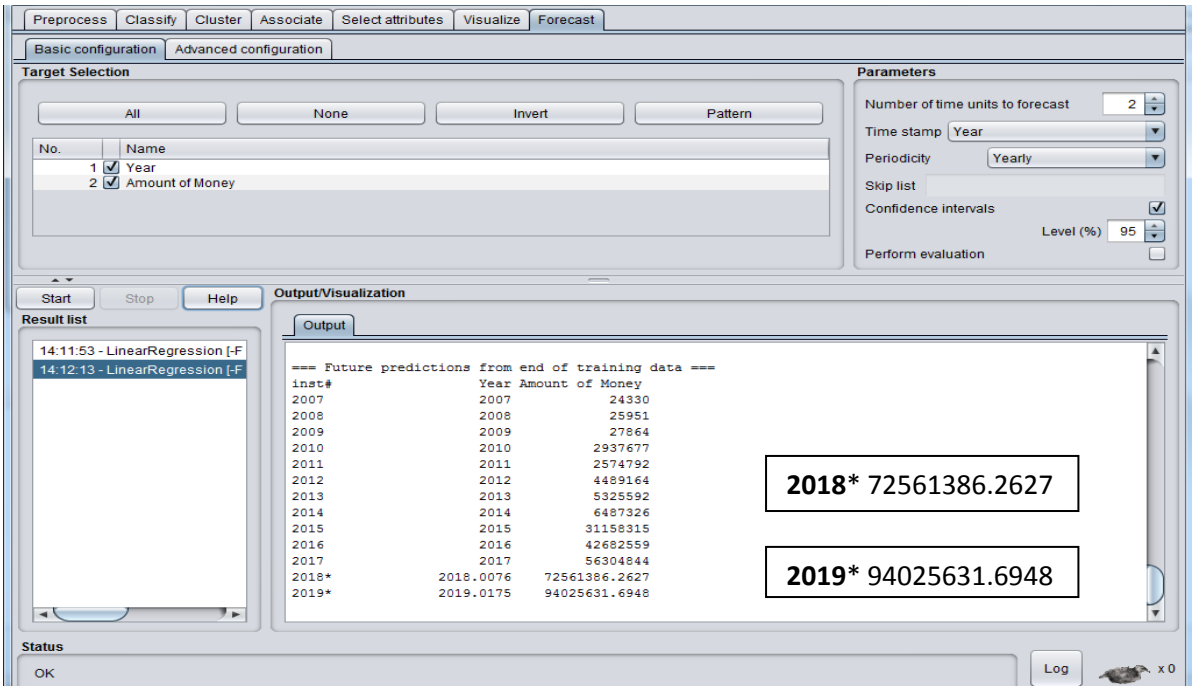



Figure 3.16 predicting amount of money using Liner regression

Forecasting the amount of money for years 2018-2019 was achieved by using Simple linear regression, as explained in figure 3.17.

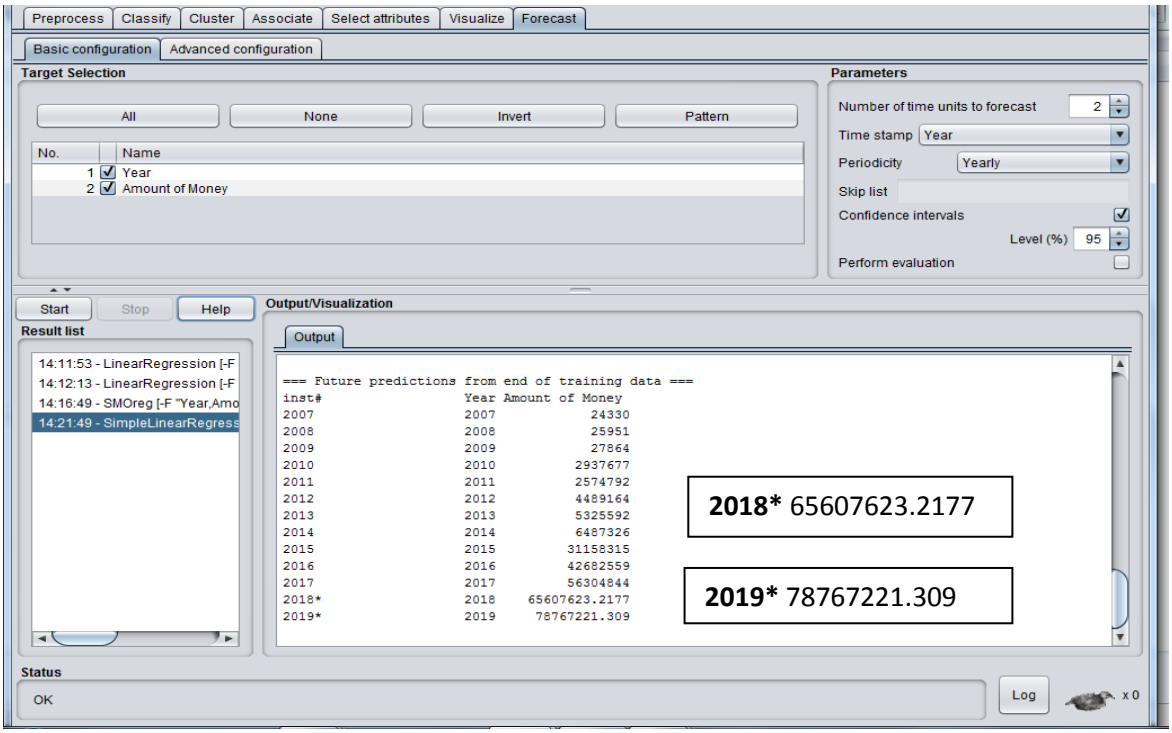



Figure 3.17 predicting amount of money using simple linear regression

Three algorithms of time series functions were used to predict the number of borrowers for the years 2018-2019 based on data from previous years for the period (2010-2017) which contains two attribute: years and the number of borrowers. Firstly, SMO regression matrix algorithm as explained in Figure 3.18 and Figure 3.19.

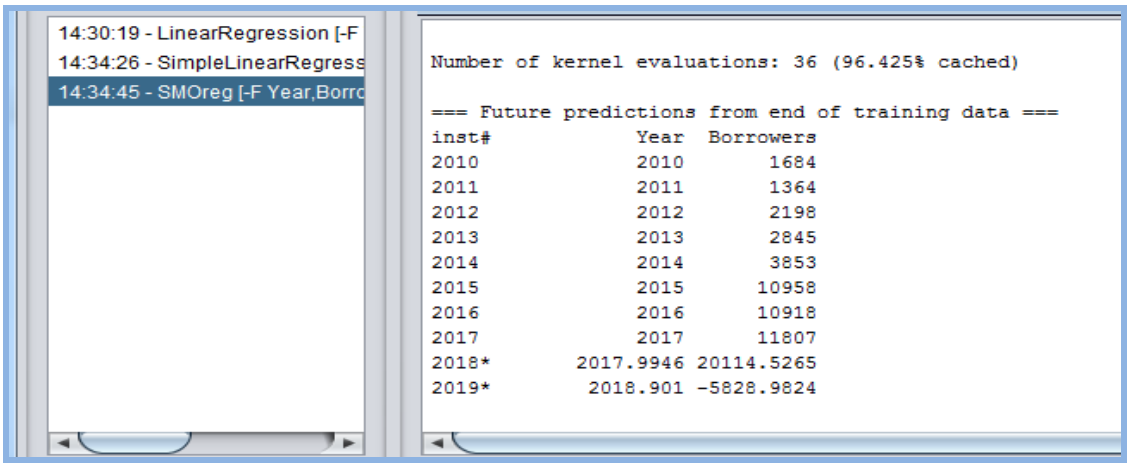



Figure 3.18 to predict number of borrowers by using SMO regression

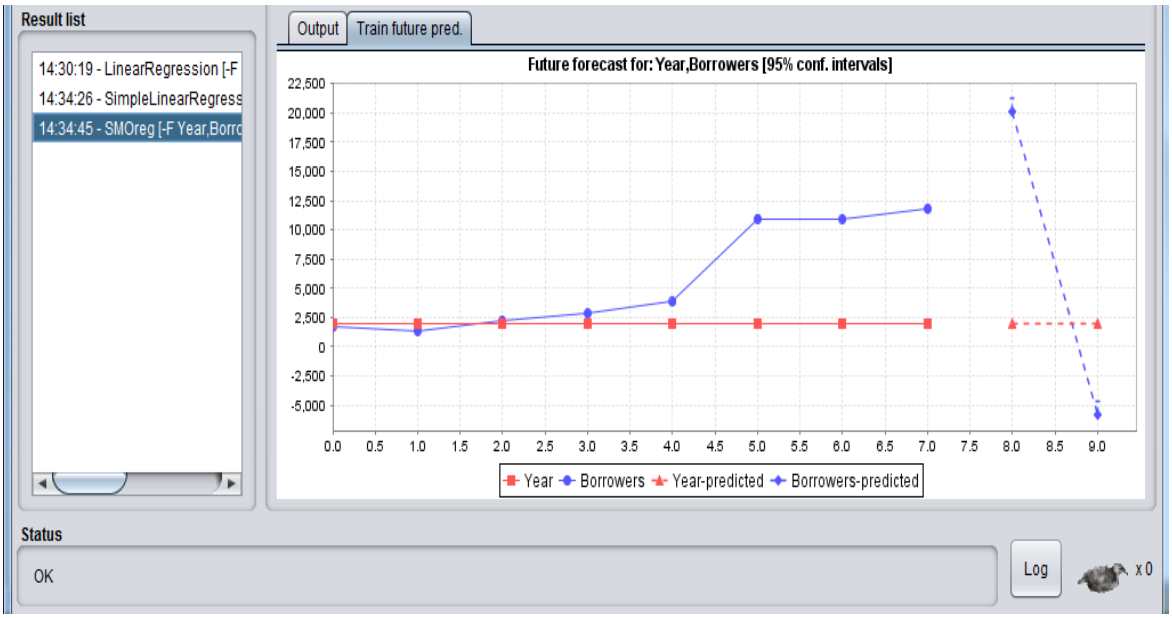



Figure 3.19 forecasting borrowers using SMO regression

Secondly, Simple Linear Regression function was used, as explained in figure 3.20 and 3.21respectively.

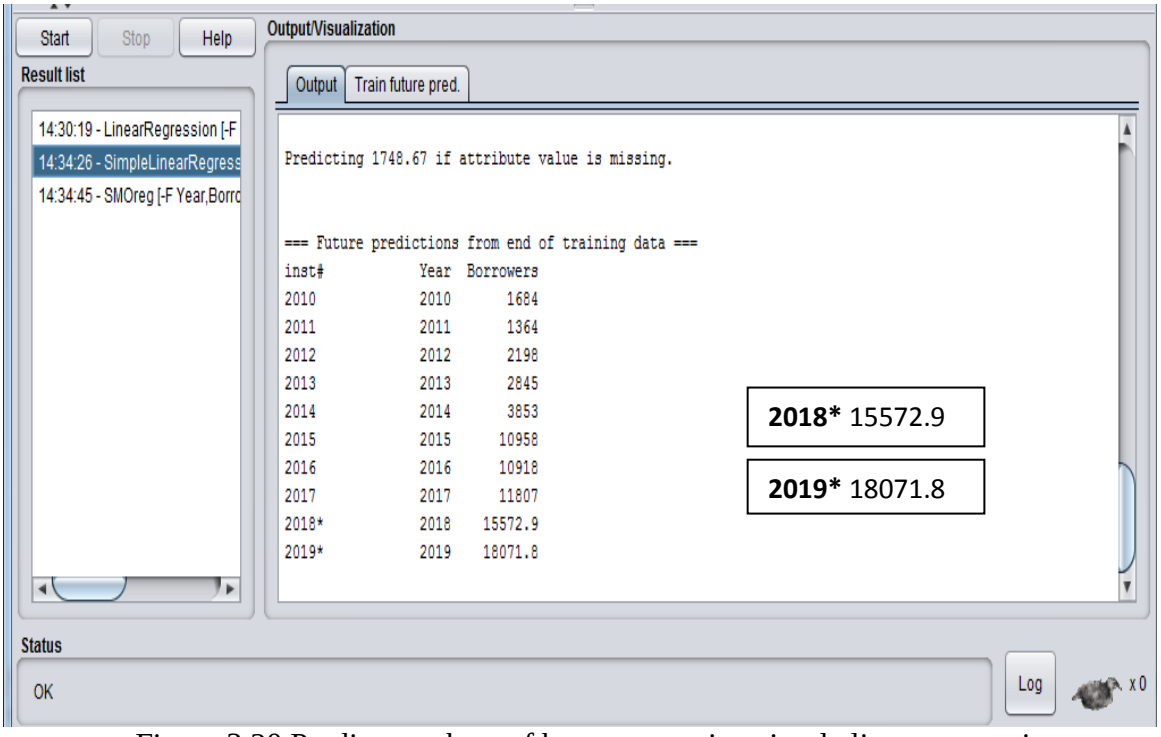



Figure 3.20 Predict numbers of borrowers using simple linear regression

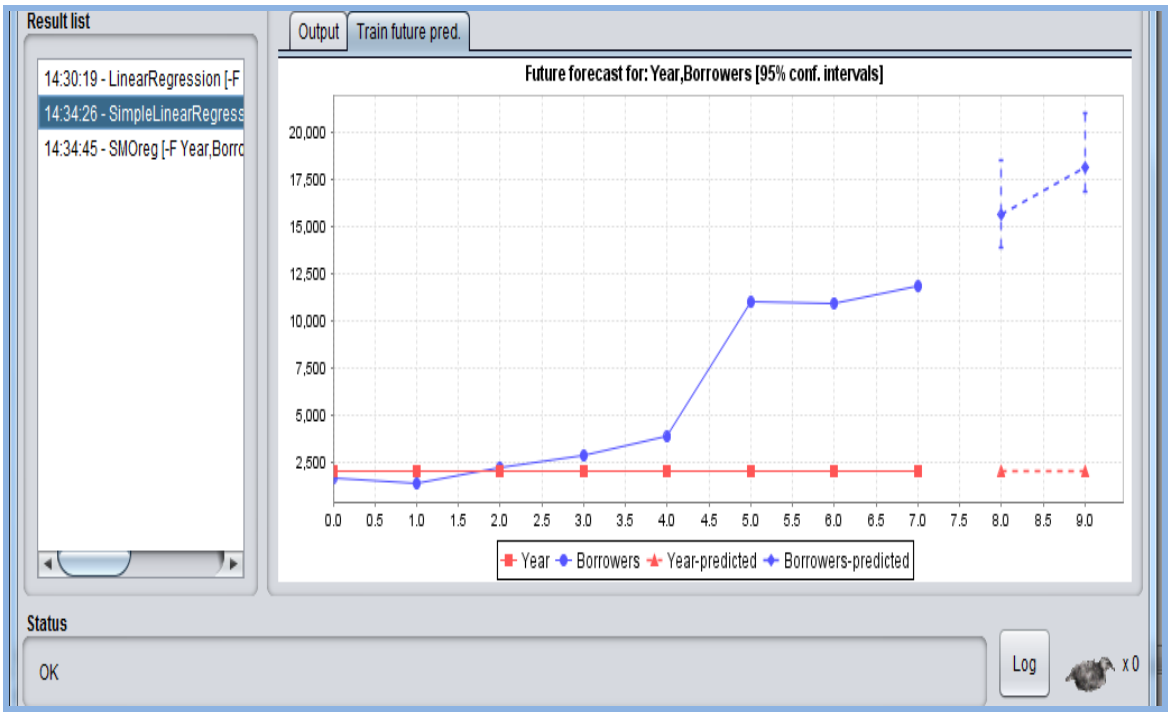



Figure 3.21 Future forecasting for borrowers using Simple linear regression

Thirdly, Linear Regression, explained was performed as shown in figures 3.22 and figure 3.23.

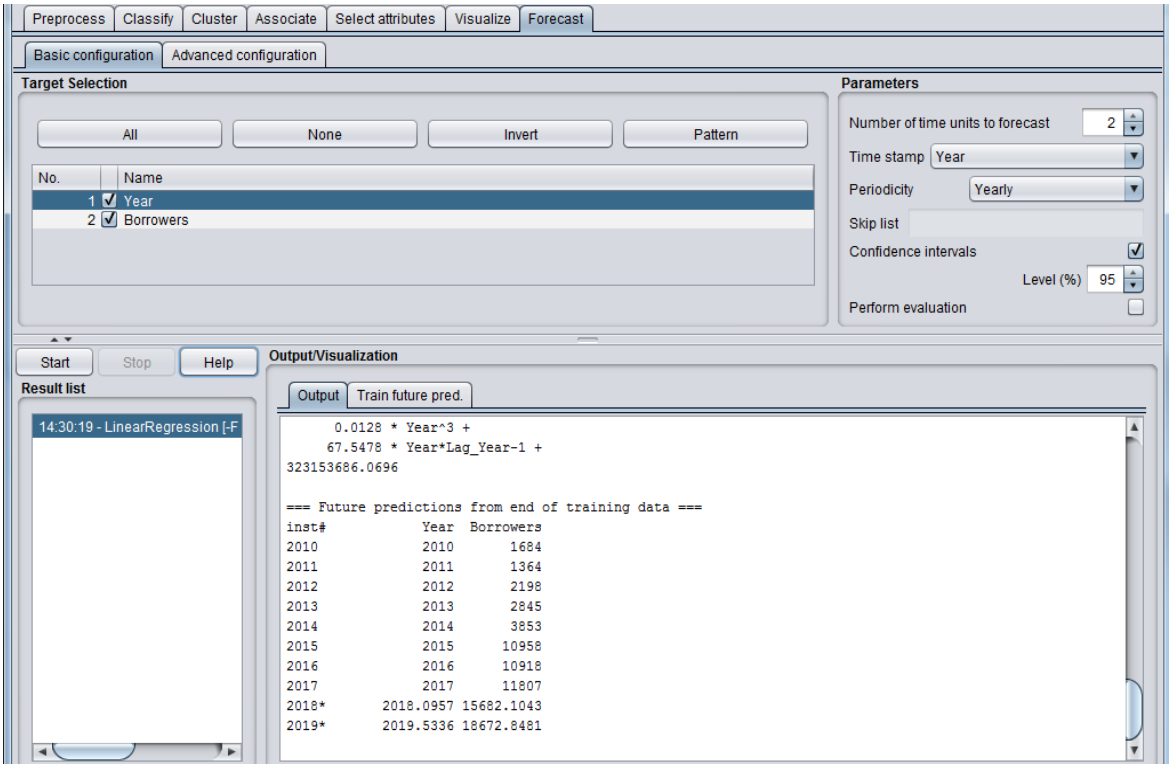



Figure 3.22 Predict numbers of borrowers using linear regression

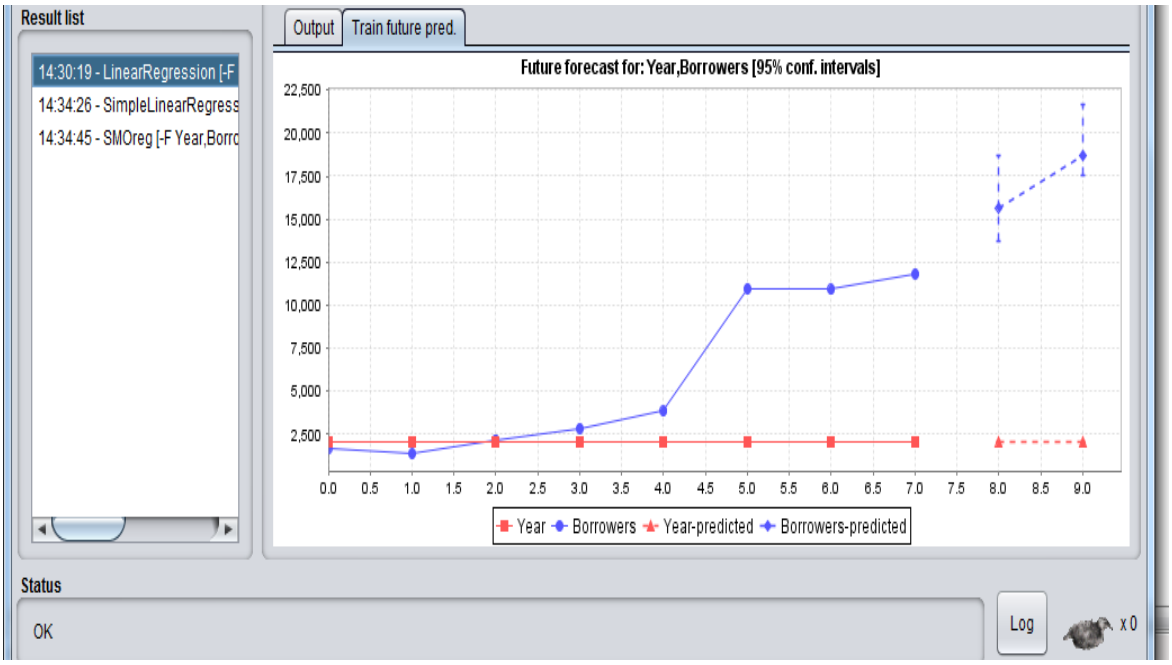



Figure 3.23 Future forecasting for borrowers using linear regression

The amount of money and number of borrowers for year 2017 is given, but foreseen, forecasting year 2017to ensure that the previous expected results for years2018-2019 are correct with high accuracy.

Firstly, predict Amount of Money by using linear regression, with parameters and confidence interval95%, explained in figure 3.24.

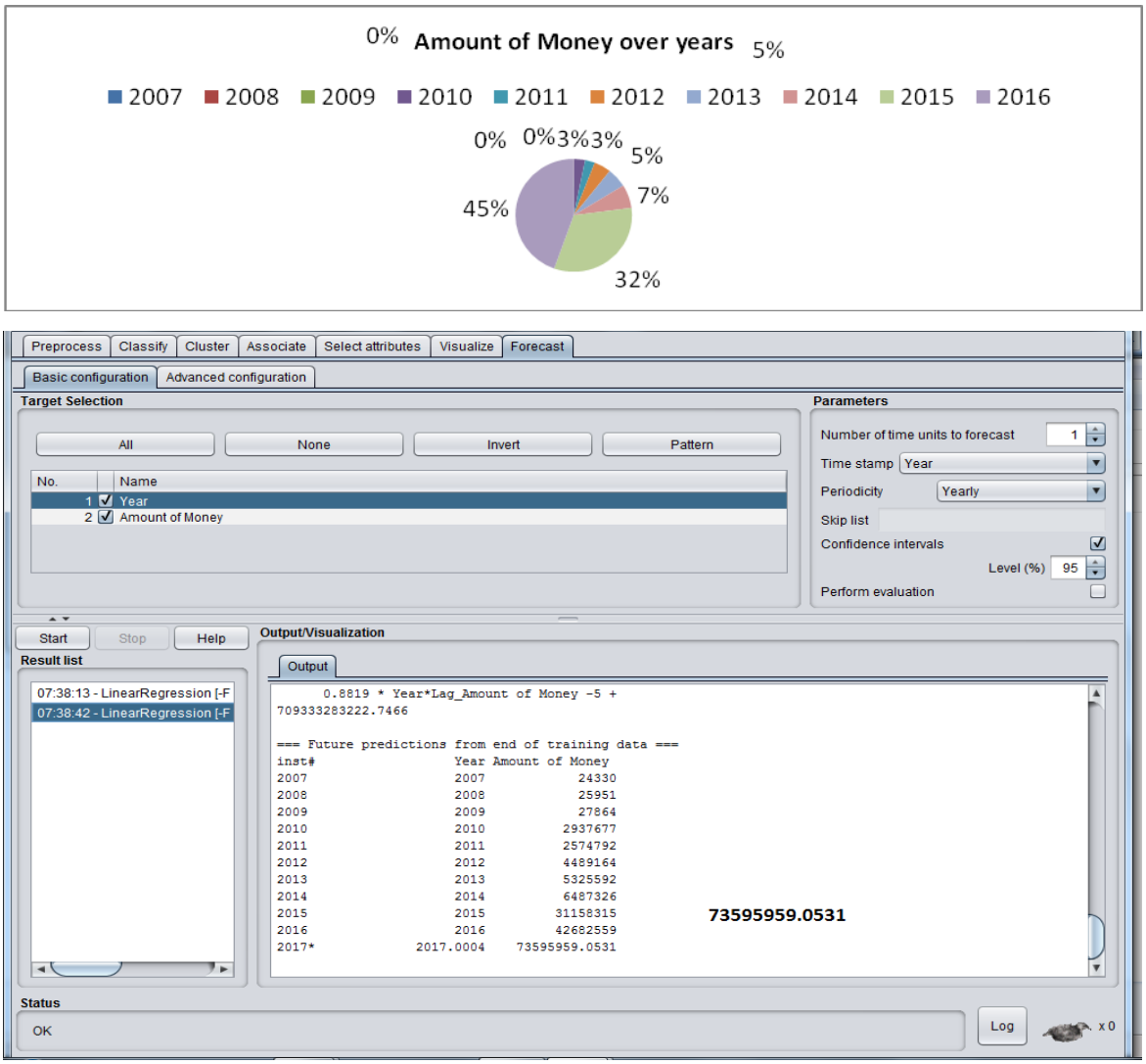



Figure 3.24 Predict amount of money using linear regression

Secondly, by using SMO regression forecast the amount of money in year 2017 with parameters and confidence interval **95%**, explained in Figure 3.25.

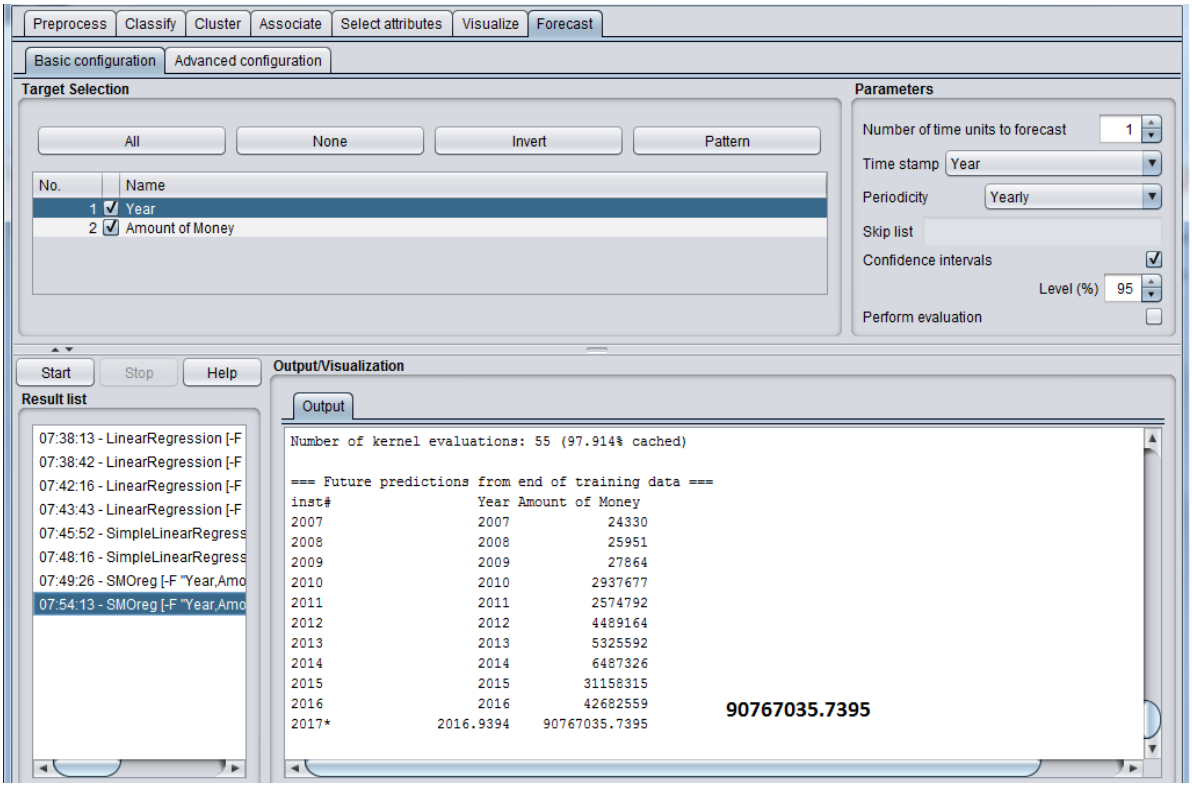



Figure 3.25 Predict amount of money using SMO regression

Thirdly, by using simple linear regression to forecasting the amount of money in year 2017, was explained in figure 3.26.

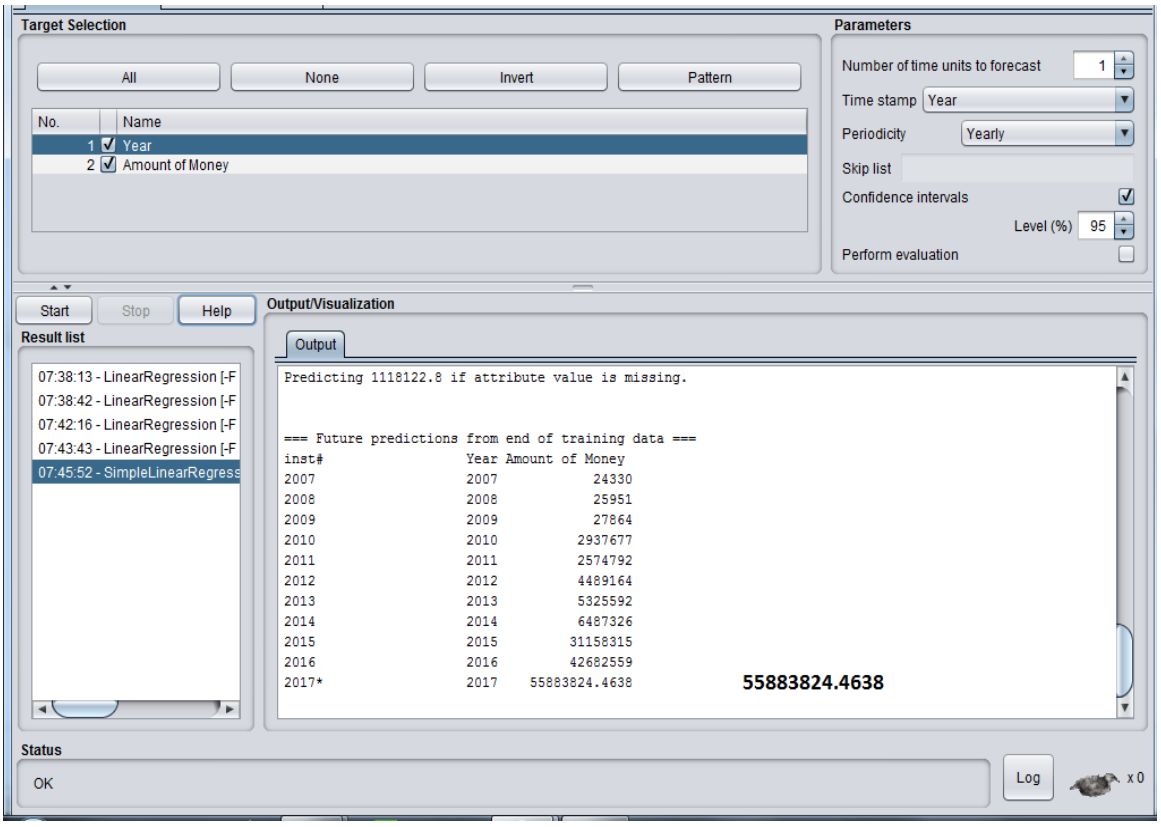



Figure 3.26 Predict amount of money using Simple linear regression

Then used classifiers functions for forecasting number of borrowers the data contain two attribute years & number of borrowers, all these was explained in coming figures .firstly, forecasting number of borrowers used linear regression, as explained in figure 3.27 andfigure 3.28.

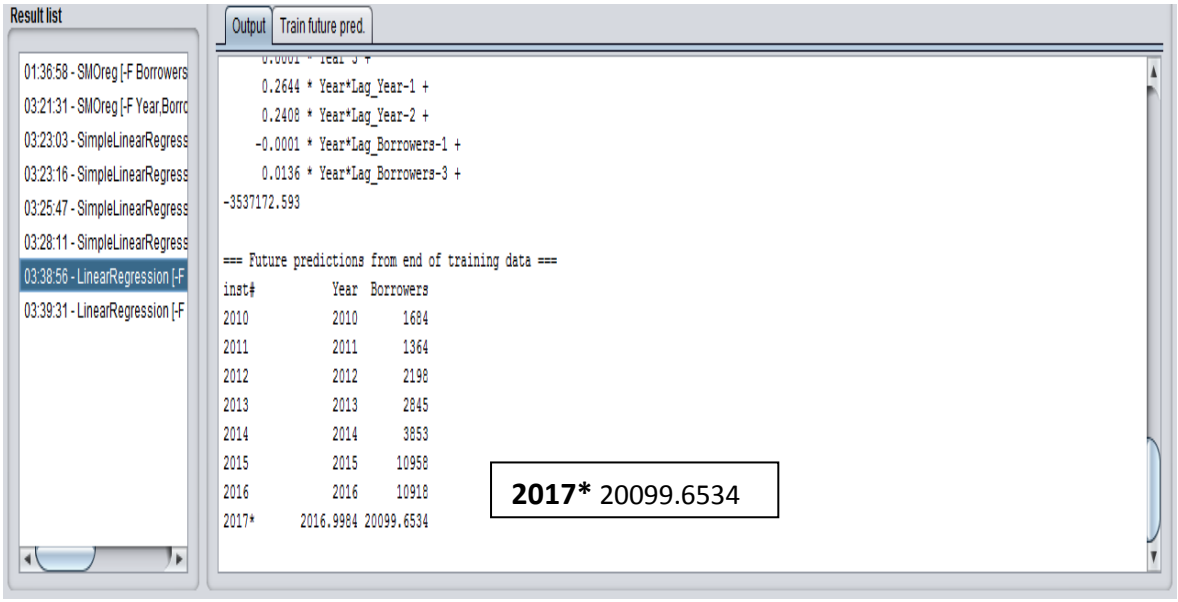



Figure 3.27 Predict number of borrowers for year 2017 using linear regression


Figure 3.28 Future forecasting for borrowers for year 2017(linear regression)

Then using SMO regression explained in figure 3.29 and 3.30 respectively.

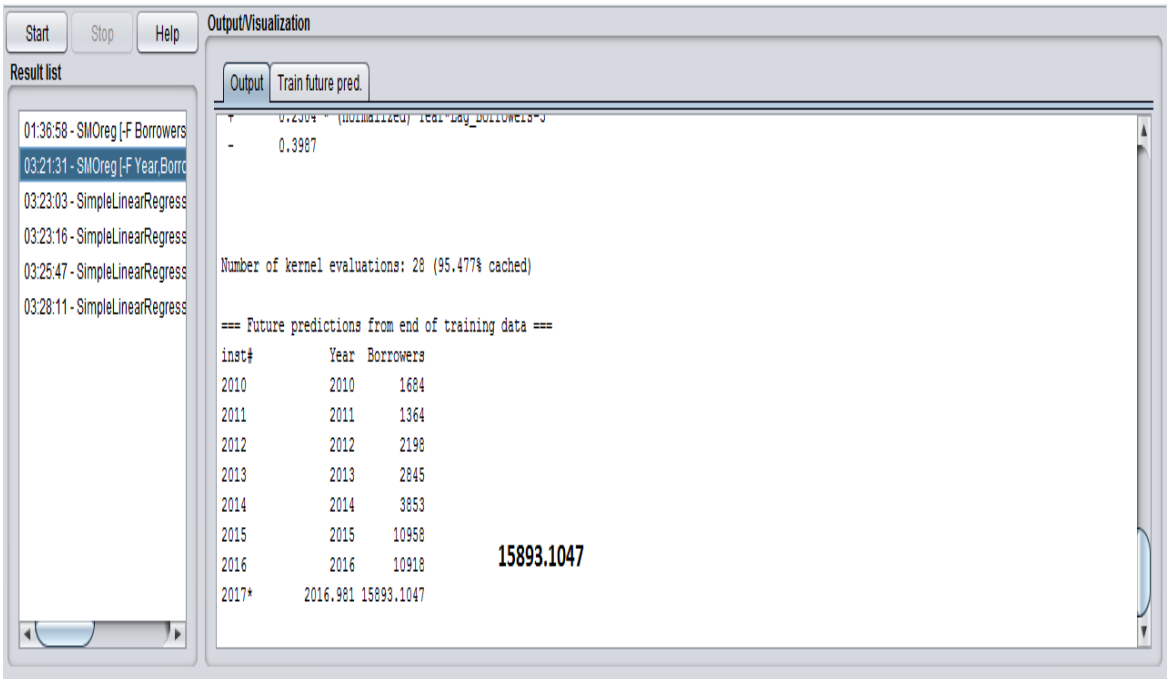



Figure 3.29 Predict numbers of borrowers using SMO regression

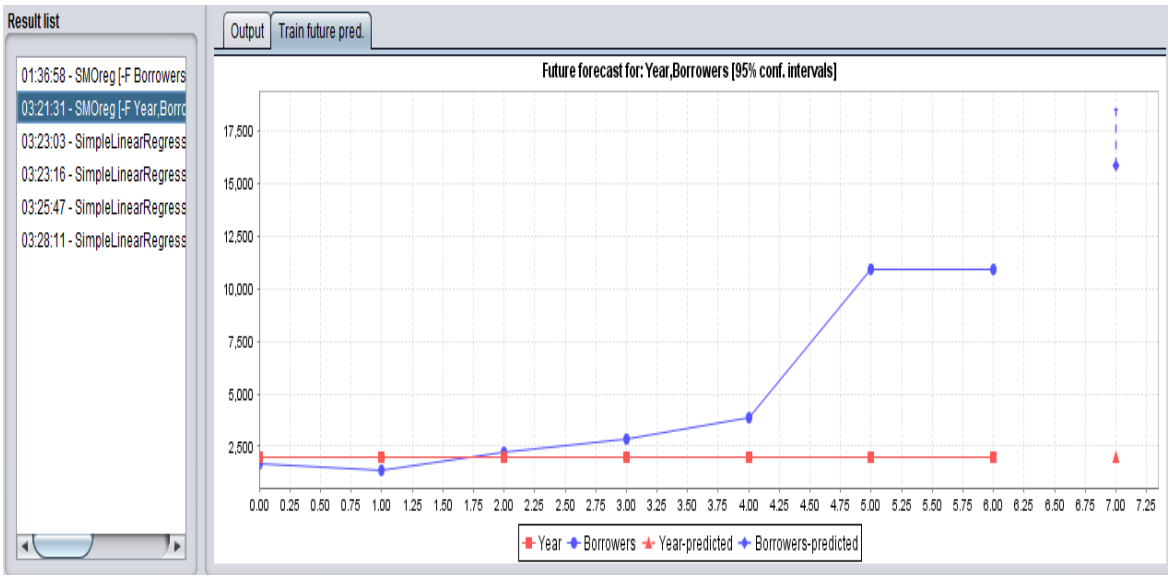



Figure 3.30 Future forecasting for borrowers using SMO regression

After that used simple linear regression to forecasting the number of borrowers in year 2017, as explained in figure 3.31and figure 3.32.

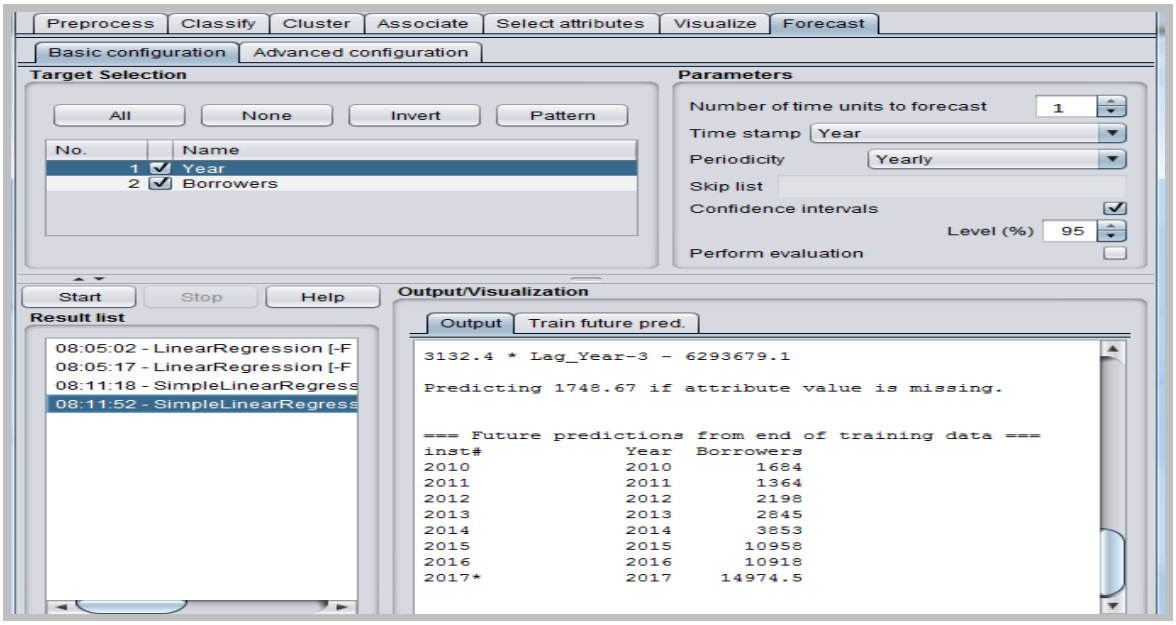



Figure 3.31 Predict number of borrowers using simple linear regression

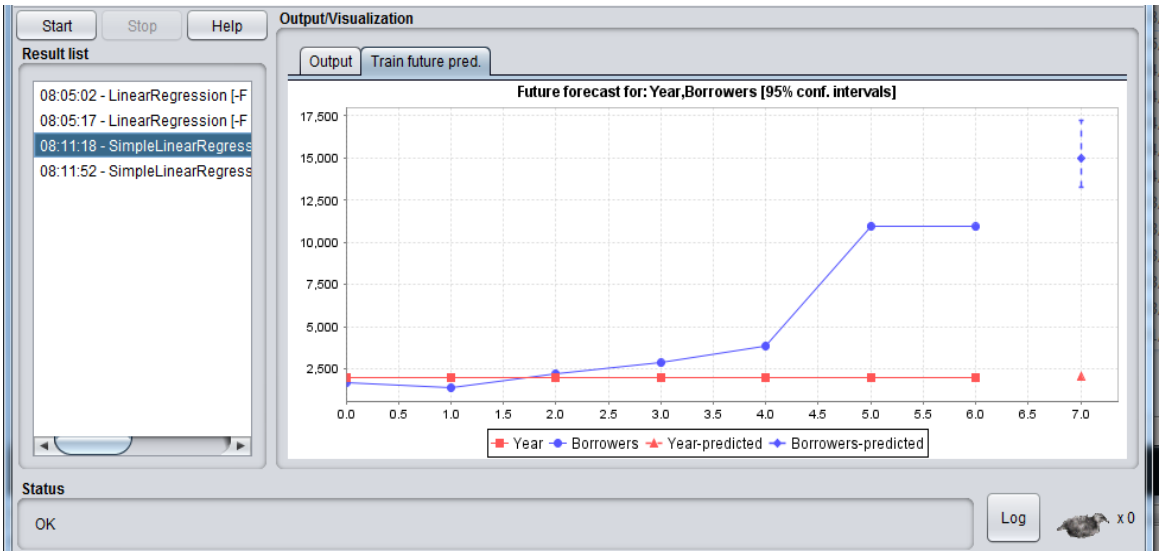



Figure 3.32 Future forecasting for borrowers using simple linear regression

## 3.6 Summary:

The methodology adopted by the researcher was described and followed step by step in this chapter. A classification model was investigated. Three regression techniques were tested on the data. The results of the model and the forecasting will be addresed and discussed in the next chapter.

# CHAPTER FOUR

# Results and Discussions

## 4.1 Introduction:

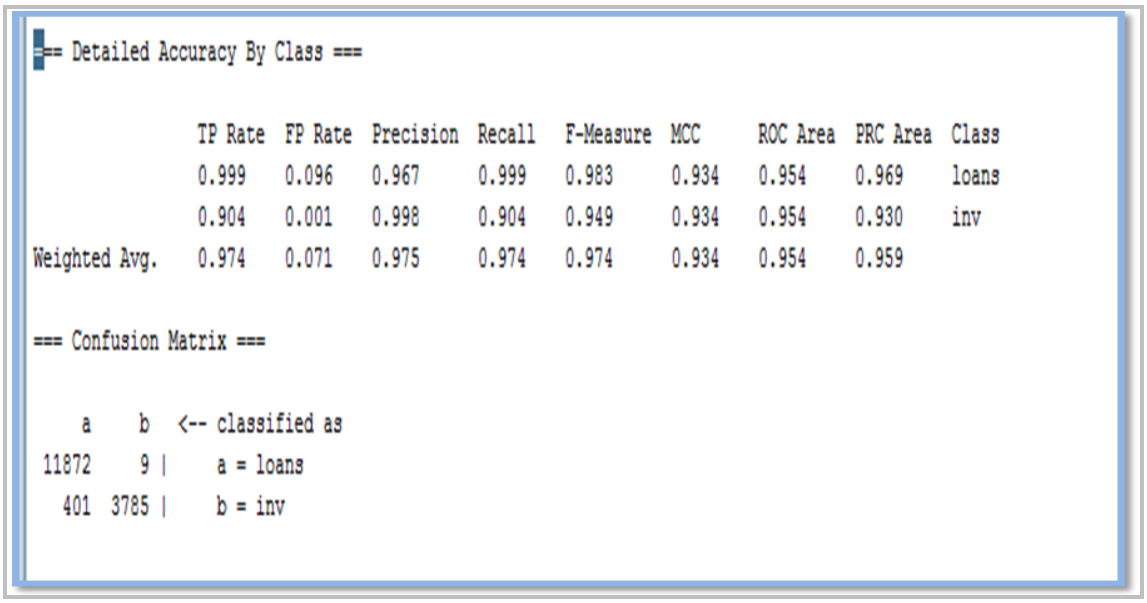
This chapter includes the results that the researcher reached after following the steps in the previous chapter and discussion the results.

## 4. 2 Results of classifying techniques:

Classifying techniques using TreesJ48 (Class for generating a pruned or unpruned C4), the results showed that the correlation coefficient is high indicating.

Correctly classified instances are equal to 97.4482%, The time taken to test the model on test split: 0.2 seconds,as explained in figure 4.1on detailed accuracy by class and confusion matrix.

Also results shown the decision tree explained that the attributes type of loan and loan amount are the most influential when the loan type is greater than or equal to 1 and the amount of the loan is greater than or equal to 3500, visualized tree in figure 4.2.



```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.999    0.096    0.967      0.999    0.983      0.934    0.954     0.969     loans
               0.904    0.001    0.998      0.904    0.949      0.934    0.954     0.930     inv
Weighted Avg.  0.974    0.071    0.975      0.974    0.974      0.934    0.954     0.959

=== Confusion Matrix ===

     a     b   <-- classified as
 11872     9 |     a = loans
   401  3785 |     b = inv
```

Figure 4.1: Results of classify techniques using TreesJ48

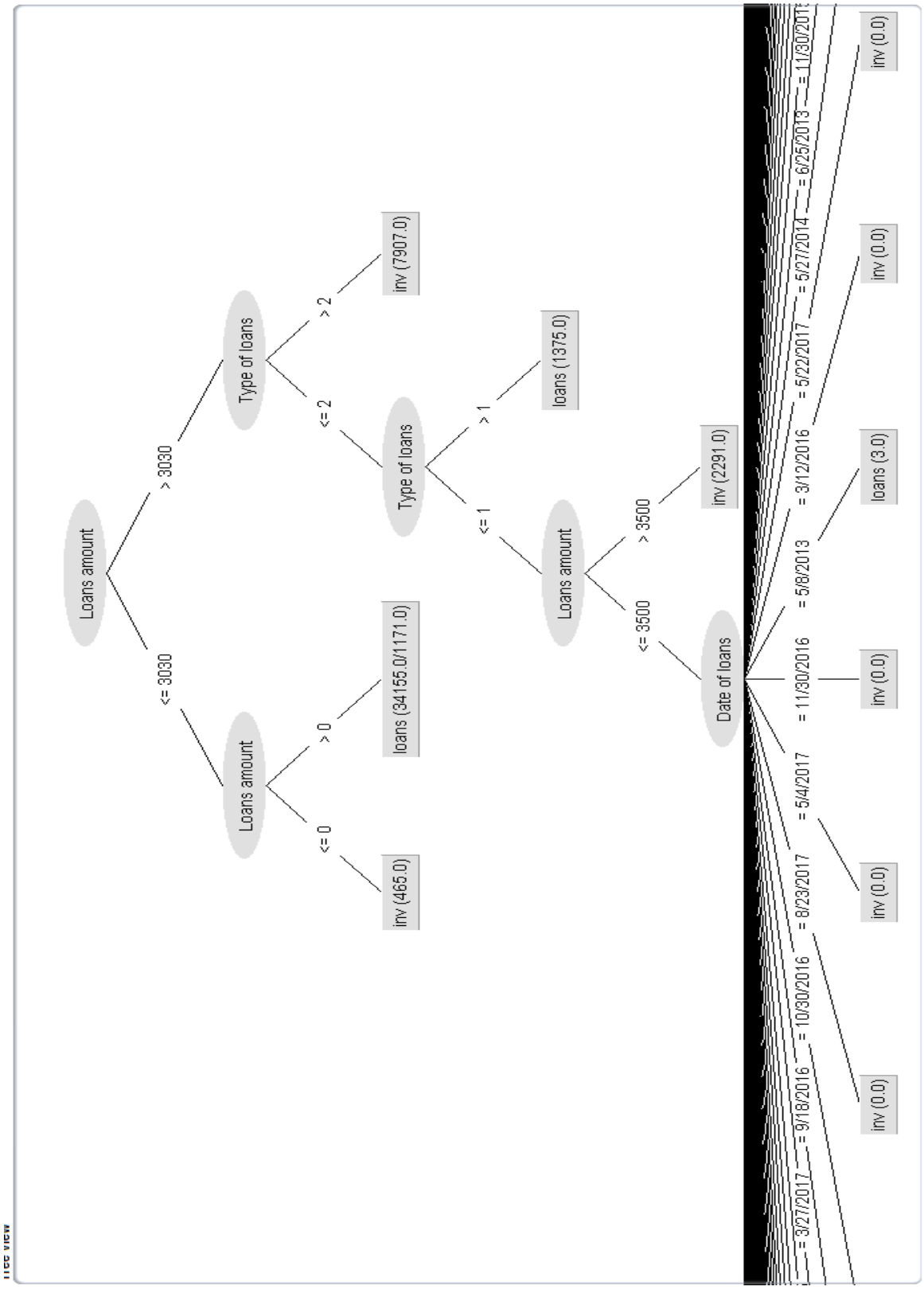



Figure 4.2 Visualized trees using TreesJ48

## 4.3 Results of classifiers function:

There are two results of classifier function one for Amount of Money and other for borrowers.

### 4.3.1 Results of classifiers function for amount of money for years 2018-2019:

The results of forecasting Amount of Money for years (2018-2019) was explained in figure 3.15 by using SMO regression are asfollows:

- Amount of Money2018 =52495089.9779 SDG
- Amount of Money2019 =43278774.3232SDG

The results of used linear regression to forecast Amount of money for years 2018-2019 from years (2007 to 2017) as explained in figure 3.16.by using linear regression with confidence 95%as follows:

- Amount of Money in Year 2018=72561386.2627SDG
- Amount of Money in Year 2019=94025631.6948 SDG

The results of forecasting Amount of Money for years 2018-2019 was explained in figure 3.17.by using Simple linear regression as follows:

- Amount of Money in Year 2018 =65607623.2177 SDG
- Amount of Money in Year 2019 =78767221.309SDG

### 4.3.2 Results of classifiers function for borrowers for years 2018-2019:

The results of used SMO regression classifiers functions to forecast number of borrowers for years (2018-2019) from years (2010 - 2017) was explained in Figure 3.18 and Figure 3.19 when the data contain two attribute years & number of borrowers as follows:

- Year 2018= 20114.5265 borrowers.
- Year 2019= -5828.9824 borrowers.

The results of forecasting number of borrowers for years (2018-2019) as explained in figure 3.20 and figure 3.21by using Simple Linear Regression as follows:

- Year 2018 = 15572.9 borrowers.
- Year 2019 = 18071.8 borrowers.

The results of forecasting number of borrowers for a year 2018-2019 as explained in figures 3.22 and figure 3.23 by using Linear Regression as follows:

- Year 2018 =15682.1043 borrowers.
- Year 2019 = 18672.8481 borrowers.

### 4.3.3 Predicting results for amount of money for year2017:

The results showed the highest accuracy with few error is when used Simple Linear regression to expected amount of money for loans comparing with SMO regression and linear regression.

The result of forecasting Amount of Money for a year 2017( as explained in figure 3.24)by using linear regression is:

- Year 2017 =7359595.0531SDG

The result of forecasting Amount of Money for a year2017( as explained in figure 3.25)by using SMO regression as follows:

- Year 2017 = 90767035.7395 SDG

The result of forecasting Amount of Money for a year 2017( as explained in figure 3.26)by using simple linear regression as follows:

- Year 2017 predict amount of money= 55883824.4638SDG

Amount of Money predicted in years 2017-2018-2019 was explained in this table.

Table 4.1 Amount of Money predicted in years2017-2018-2019

| Years | Linear Regression | SMO Regression | Simple Linear Regression |
|---|---|---|---|
| **2017** | 7359595.0531 | 90767035.7395 | 55883824.4638 |
| **2018** | 72561386.2627 | 5249089.9779 | 65607623.2177 |
| **2019** | 94025631.6948 | 43278774.3232 | 78767221,309 |

### 4.3.4 Predicting results for amount of money for year2017:

The results of used linear Regression (explained in figure 3.27 and figure 3.28) to forecasting the borrowers in the year 2017as follows:

- Borrowers in Year 2017 = 20099.6534 borrowers

The results of used SMO regression explained in figure 3.29 and figure 3.30 to forecasting number borrowers in year 2017is:

- Borrowers in Year 2017 =15893.1047 borrowers

The results of used simple linear regression explained in figure 3.31 and figure 3.32to forecasting number borrowers in years 2017 is:

- Borrowers in Year 2017= 14974.5 borrowers

* Numbers of borrowers predicted in years2017-2018-2019 were explained in table.4.2.

Table 4.2 Numbers of borrowers predicted in years2017-2018-2019

| Years | Linear regression | SMO Regression | Simple Linear Regression |
|---|---|---|---|
| **2017** | 20099.6534 | 15893.1047 | 14974.5 |
| **2018** | 15682.1043 | 20114.5265 | 15572.9 |
| **2019** | 18672.8481 | -5828.9824 | 18071.8 |

## 4.4 Comparison of results:

Comparing the expected results with real for the year 2017

### 4.4.1 Comparing the real amount of money with predicting one :

When comparing the expected results with the amount of money spent for the year 2017, the closest result of the real and little errors was upon using simple linear regression

- Real amount of money in 2017 =56304844SDG
- Predictedamount of money in 2017 = 55883824.4638 SDG
- Prediction error = 421019.5362
- Predicted amount of money in 2017 by using SMO regression = 90767035.7395 SDG
- Prediction error =34462191.7395
- Predicted amount of money in 2017 by using linear regression = 48945248.9469 SDG

### 4.4.2 Comparing the real number of borrowers with predicting one:

When comparing the results obtained with the number of real borrowers in 2017 the researcher found that the best way to predict number of the borrowers is simple linear regression.

- Real number of borrowers in 2017 =11807borrowers
- Predicted number of borrowers in2017 by using simple linear regression= 14974.5 borrowers
- Prediction error = 3167.5
- Predicted number of borrowers in 2017by using SMO regression = 15893.1047 borrowers
- Prediction error = 4086.1047

- Predicted number of borrowers in 2017by using linear regression = 120099.6534 borrowers
- Prediction error = 108292.6534

The observations of this result are low error percentage with simple linear regression, so, the researcher proposed using simple linear regression to predict number of borrowers in the coming years.

## 4.5 Summary:

After implementing many prediction methods, it was evident that Simple Linear Regression ide yields the best results. It provides the best results to forecast financial amount in the coming years and number of borrowers as well to help decision makers to prepare the appropriate budget for the next years.

# CHAPTER FIVE

# Conclusions and Future Work

## 5.1 Conclusions

The nature of pension data and its large size was a challenge to be worked on and after the integration of investment and loans databases, it was found that the most appropriate & the best way to build the model is the J48.

The best way to predict the number of borrowers and the budget of investments and loans is the simple linear regression as it proved its worth with the least error and it gives more accurate result.

## 5.2 Future Work

To benefit from the huge data which is stored in the data base, it is recommended to:

- Develop self-tools for additional exploration of data which is not included in the existing databases.
- Apply a technique to understand the behavior of the pensioners through investment applications and loans.
- Deploy a new method for predicting the type of investment required by pensioners for the coming years.

# References:


Adedara, O. a. (2012). Forecasting Portfolio Investment Using Data Mining. *African Journal of Computing & ICT*, 103-108.

Bharati M. Ramageri, ". (2010). Data MINING TECHNIQUES AND APPLICATIONS". *Journal of Computer Science and Engineering*.

Chawan, A. A. (May 2013). Study of Data Mining Techniques used for Financial Data Analysis. *International Journal of Engineering Science and Innovative Technology (IJESIT)*, 503-509.

Elsalamony, H. A. (January 2014). Bank Direct Marketing Analysis of Data Mining Techniques. *International Journal of Computer Applications*, 12-22.

Jonah Mushava ∗, M. M. (19 February 2018). An experimental comparison of classification techniques in debt recoveries scoring: Evidence from South Africa's unsecured lending market. *Elsevier Ltd*, 35-50.

Joo2, J. E.-H. (2004). Environmental and Biophysical Factors Associated with Financial Risk Tolerance. *Journal of Financial Counseling and Planning*, 8.

Kalmegh, S. (February 2015). Analysis of weka data mining algorithm REPTree, Simple Cart. *IJISET - International Journal of Innovative Science, Engineering & Technology*.

Kamber/JianPei, J. H. (n.d.). *Data Mining-Concepts and Techniques* (3rd Edition ed.).

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

Moin, K. a. (2012). Use of data mining in banking. *International Journal of Engineering Research and Applications, 2(2)*, 738-742.

Sakshi Singh, H. M. (2014). Prediction of investment patterns using data mining techniques. *International Journal of Computer and Communication Engineering*, (145-148).

Sakshi Singh, H. M. (2014). Prediction of investment patterns using data mining techniques. *International Journal of Computer and Communication Engineering*, 145-148.

Study of Data Mining Techniques used for Financial Data Analysis. (n.d.).

Veld, C. a.-M. (2008). The risk perceptions of individual investors. *Journal of Economic Psychology* , 226-252.

Sudan University of Science & Technology
College of Graduate Studies
College of Computer Science and Information Technology

# Building a Classification and Forecasting Model to Support Decision Making

**Case study: The National Pensions & Social Insurance Fund (Government sector)**

**بناء نموذج للتصنيف والتنبؤ لدعم اتخاذ القرار**

**دراسة حالة:الصندوق القومي للمعاشات والتأمينات(القطاع الحكومي)**

**Thesis Submitted in Partial Fulfillment of Requirement for Master Degree in Information Technology**

## Submitted by:

**Hajer Ibrahim Hassan Idrees**

## Supervisor:

**Dr. Wafaa Faisal Mukhtar**

30.Jan.2018

أعوذ بالله من الشيطان الرجيم

بسم الله الرحمن الرحيم

(وقل ربِّ زدني علماً)

سورة طه آية(114)

(أَمَّنْ هو قانتٌ آناءَ اللَّيلِ ساجداً وقائماً يَحذَرُ الآخرةَ ويَرجو
رحمةَ ربِّهِ قلْ هل يستوي الَّذين يعلمونَ والَّذين لا يعلمونَ
إنَّما يَتَذكَّرُ أولوا الألباب)

سورةالزمر آية(9)

الحمد لله الذي بنعمته تتم الصالحات

# *DEDICATION*

***To My Mother & the Spirit of My Father for being an example of love and care taught me the power of principle. To My Doctors who leads me to successful and progress I am very thankful and grateful for them. To my sisters and brother who gave me their time. To everyone who taught me useful things in my life. To all, who by example, friends and person aimed to release the computer technology in my country .***

# *ACKNOWLEDGEMENT*

***Thanks to Mighty Allah, for giving me the power to complete this research…***

***All thanks to my supervisor: Dr: WafaaFaisal Mukhtar who guided and supported me to complete this research…***

***Many thanks to the National Pension and Insurance Funds (government sector), especially thanks for general manager Miss.Bothena Mohammed Ibrahim and the staff of Computer Management …***

***Many thanks to teacher Ali Abdallah Abker in AL-Nilain University, faculty of Information System who help me in implementing of the research.***

***To SUST and college of Computer Science and Information Technology family, and to my colleagues many thanks and respect to all.***

# ABSTRACT

Data might be one of the most valuable assets of any corporation but only if it knows how to reveal valuable knowledge hidden in raw data. Data mining allows extracting precious knowledge from historical data, and predicting the outcomes of future situations.

A large database of about thousands is stored in an Oracle database system. It contains the demographical information and more data about the history of the pensioner financial history at the National Pension and Social Insurance Fund. Managing loans for regular use or investment is unpredicted.

The main aim of this research is to predict the numbers of borrowers and the budget of investment & loans by using classification and regression techniques. The model was built by using the Trees function-RJ48 algorithm. Different regression algorithms were used such as Sequential Minimal Optimization Regression, Linear Regression, and Simple Linear Regression. Regression results were compared against the actual money spent for the years 2017-2019 to predict the number of borrowers and the budget. The simple linear regression proved to be the most accurate, with the least error ratio.

# المستخلص

تعتبر البيانات واحدة من أكثر الأصول قيمة لأي مؤسسة إذا كانت تعرف كيفية الكشف عن المعرفة القيمة المخبأة في البيانات الخام. يتيح تنقيب البيانات استخراج المعرفة من البيانات التاريخية ، والتنبؤ بنتائج المواقف المستقبلية.

هنالك قاعده بيانات ضخمة للمعاشيين مخزنة في قاعدة البيانات اوراكل. حيث تحتوى على المعلومات الأساسية وبيانات اخرى مالية تخص الصندوق الوطني للمعاشات والتأمينات الإجتماعية (القطاع الحكومي). إدارة السلفيات العادية والاستثمارات لا يمكن توقعها.

الهدف الرئيس من هذه الدراسة هو التنبؤ بأعداد المقترضين وميزانية الإستثمار والقروض وبناء نموذج تصنيف لأنواع القروض. تم بإستخدام تقنيات التصنيف واستخدام خوارزمية الأشجار (RJ48) وتطبيقها على بيانات الإستثمار والقروض المأخوذة من قاعدة بيانات الصندوق الوطني للمعاشات والتأمينات الإجتماعية (القطاع الحكومي) بناء النموذج. كما تم استخدام عدة خوارزميات لقياس الانحدار مثل الحد الأدنى المتسلسل الأمثل(**SMO**)، والانحدار الخطي ، والانحدار الخطي البسيط. تمت مقارنة النتائج المتحصل عليها مع المبالغ الفعلية المنصرفة للسنوات 2017-2019 للتنبؤ بعدد المقترضين. ثبت أن الإنحدار الخطي البسيط هو الأكثر دقة ، مع أقل نسبة للخطأ.

# Table of Contents:

# List of Tables:

# List of Figures: