# Data Mining Application in Predicting Bank Loan Defaulters

**Ashenafi Wubshet Desta, J. Sebastian Nixon**

***Abstract - Data mining is the key tools for discoveries of knowledge from large data set. Nowadays, most of the organizations using this technology to maintain their data. This paper focuses on the Bank sector in Risk management specifically, detecting Bank loan defaulters through the data mining application to examine the patterns of different attribute which would contribute for detecting and predicting defaulters thus preventing wrong loans. This process can be done without change the current systems and the data. Then it helps to distinguish borrowers who repay loans promptly from those who don't and avoid wrong loan allotment. In order to show the results of the study Classification model is implemented in order to find interesting patterns among attributes of customer. A total of 20461 sample data were taken by data base admin randomly from 3 consecutive years from the Bank database to build and test the model. In this research we used Classification model of decision tree and Naïve Bayes in Weka 3.7 tool for experiments. Modeling methodology applied to this paper was CIRSP-DM (Cross Industry Standard for Data Mining), which involves business understanding, data understanding, data preparation, model building, evaluation and deployment. Decision tree classifications with J48 implementation with 8 experiments were performed. Two experiments with different parameters were made for Naïve Bayes. Finally, evaluation and analysis of the models were performed then given a best solution to predict the defaulters.***

***Keywords : CIRSP-DM, DSS, Naïve Bayes, J48, Weka..***

## I. INTRODUCTION

### 1.1 Back ground of the study

As banking competition becomes more and more global and intense, banks have to fight more creatively and proactively to gain or even maintain market shares. Banking sector provides different kinds of services for their customer. This will make the sector to be dependent on customer. It is true, without these customers the bank sector will not exist, so this will force them to focus on customer satisfaction. In order to know their customer, the sector takes different kinds of measurement. The sector in the future will use one asset, knowledge and not financial resources for survival and excellence. Using information systems as decision support systems (DSS) helps management to make effective decisions in various ways.

**Revised Manuscript Received on February 28, 2020.**

***** Correspondence Author

**J. Sebastian Nixon*,** School of Informatics, Wolaita Sodo University, Ethiopia. E-mail:dr.nixon14@gmail.com

**Ashenafi Wubshet Desta,** School of Informatics, Wolaita Sodo University, Ethiopia. ashumarley@gmail.com

Data mining techniques are playing a major role in knowledge discovery from databases. Nowadays, the business intelligence systems are more useful to the organizations to adjust their business goals. In most cases, these insights are driven by analyses of historical data.

Data mining can also contribute a lot in solving business problems in finance and banking sectors. This helps the administrators or managers of the particular sectors to take further steps based on the customer's transaction.

Business Intelligence and data mining techniques can also help them in identifying various characteristics of customers. Even though a lot of data mining application is available for banking sector the author of this paper is going to study the profiles of high and low risk borrowers in the banking sector.

### 1.2 Statement of the problem

The customer and the bank~~s~~ while dealing with each other will always try to cover the risk factor. To identify, quantify and control the risk factor is always an area of concern for every business organization.

In commercial lending, risk assessment is usually an attempt to quantify the risk of loss to the lender while making a particular lending decision [1].

The bank doesn't know the nature of the defaulters and it will lead the bank in to crisis. It also put the customer or borrower at defaults which is finally results as a bad loan.

### 1.3 Data mining modeling & Data analysis

The data modeling will use the CRISP-DM (Cross-Industry Standard Process for Data Mining) it is a model that data miners use to understand the research area.

As shown in Figure 1.1, in CRISP–DM any given data mining project has 6 phases.

The next phase in the sequence usually depends on the outcomes associated with the preceding phase. The most significant dependencies between phases are indicated by the arrows [2].

**Business Understanding**: it includes Project objectives and requirements understanding, Data mining problem definition. In this point the author contacts the bank staff for further understanding the business area and read a lot of literature regarding with banking sector.

The IT manager of the bank paves the way for the author in order to get the necessary data and understand the business area.

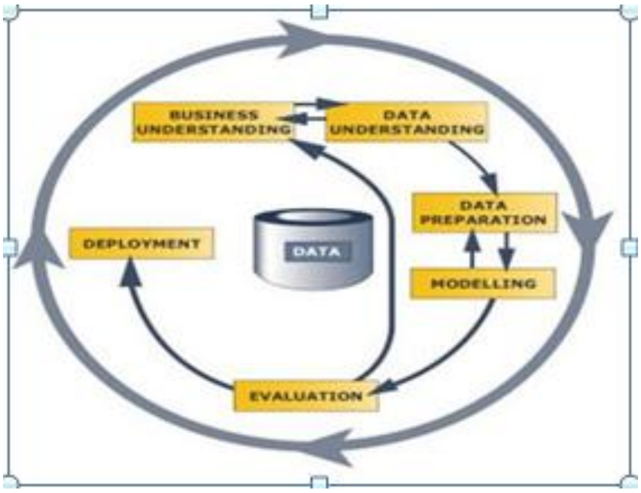


**Fig 1.1: CRISP-DM phase**

**Data Understanding:** Initial data collection and familiarization, Data quality problems identification. As though the bank system is totally changed to OMNI Enterprise core banking solutions it is difficult to identify the exact tables and attributes from the whole data base. The bank uses the OMNI Enterprise core banking solution relational database by having 106 tables in its data base. The bank uses the core banking starting from 2006. Thus for this research the data set will selected from the loan master table of the bank with 11 attributes for the 3 consecutive years randomly selected by the data administrator. These attributes are explained in chapter 3 data preprocessing like: 'AGE', 'SEX', 'REGION', 'CAR', 'MARRIED', 'SAVE_ACC', "LOANSTATUS, 'CURRENT_ACC', 'TYPESOFLOAN', 'STARTDATE', 'ENDDATE', and the derived attribute of 'LOAN DURATION' which was made by the researchers.

**Data Preparation:** Table, record and attribute selection, Data transformation and cleaning. After the author analyze the real-world services of the loan system in the bank it is easy to identify the exact table, records and attributes which is suitable for this research. In this phase the statistical summary of the data is performed which includes all the preprocessing activities.

**Modeling:** In this phase, various modeling techniques are selected and applied. But the researchers decided to use classification modeling which seems the model is appropriate for the problems. Classification modeling has a number of different algorithms appropriate for different kinds of problem. Out of them the researchers selects the Decision tree with j48 algorithm and Bayesian network with Naïve Bayes algorithm. Decision tree classifier (J48) in Weka for its visualization power of the tree structure and simplicity of understanding is used to model the preprocessed dataset. The researchers split the data set in to 70% for training and 30% for testing. By analyzing the preprocessed data which is discussed later. The experimentation and analysis is done based on the selected classification model which holds 8 experiment for decision tree and 2 experiments for Bayesian network.

**Evaluation:** It is all about visualizing the results of the data mining process. This is done by analyzing the output or results of the two-classification model. In decision tree and Bayesian network the result is compared each other by their correctly classified instances which is called as performance, the results obtain from the confusion matrix, no of leaves, size of the tree, ROC are and the like. This shows clearly the classification model result and paves the way for evaluation.

**Deployment**: Result model deployment, Repeatable data mining process implementation

### 1.4. Measurement tools:

Different kinds of tools and measurement is used for this study the Weka tools for data analysis, Microsoft Excel and Omni enterprise database used by the organization to save the customer profiles, Microsoft word for word processing of this paper and the overall word documentation.

## II. LITERATURE REVIEW

In order to control risk and maximize profits, commercial banks around the world have made great efforts to develop various analytic models to identify potential default loan applicants.

As [3], published the paper on data mining application in the analysis of default loan applicants using a real dataset consisting of 641,988 observations obtained from a Chinese commercial bank. The researchers are motivated to do this study based on:

The non-performing individual loans of Chinese commercial banks have been fast growing because of the myopic business attitude and the overheated economic growth.

To prevent the US sub-prime kind of crisis in China, An exploratory study of the dataset led to a number of interesting statistic figures that may characterize the applicants in the western region of China. In the analytic study, the researchers use SAS® Enterprise Miner to model decision tree. The models have revealed a number of useful findings that meet their expectation.

The following are some interesting findings from the exploratory study of the dataset, which may characterize the features of loan applicants in the western region of China:

1) Majority (about 90%) of loan applicants are male.

2) More than 40% loan applicants only studied at the senior high school.

3) Almost 80% loan applicants are peasants from the countryside. This characterizes the customer population in this specific bank.

4) Apart from more than 45% missing value, about 40% loan applicants fall into the group with the lowest income.

5) More than 65% of loan applicants have good health condition. Considering 20% applicants did not provide the data, the percentage of applicants with lower health condition is very low.

6) About 65% of the loan applicants have lived in their house for more than 10 years. More than three quarters of them already have their own house.

7) Less than 20% of the loan applicants claimed they have more than one dependent. Taking into account that the majority of these applicants are peasants from the countryside, we can realize the effect of one-family-one-child policy in China.

8) More than one third of loan applicants are at the age of 27 to 32.

These features can also be used to design specific financial products and services for the existing and future customers of the bank.

But as the researcher of this paper there are drawbacks in their works. They have to model the data more than one modeling in order to get best or accurate prediction. As the researcher of this paper believes two models building is better than one.

The other related works is performed by [4] is stated that application of data mining techniques for predicting loan default assumes paramount importance in Banking and financial services. The analytics involved in this context pave the way for evolving robust credit scoring models and automation of the lending process.

They also help discern the pattern of relationship between the input [borrower characteristics] and the output [loan default status]. In this paper they demonstrate that it is possible to do so using the data set of a finance company lending small loans in rural areas. Factor analysis is used to generate inputs for the application of the neural network algorithm to predict loan default and the result seen is substantial improvement in accuracy.

NN (Neural Network) is a powerful technique for predicting the behavior of output based on a given set of input values. It has a very good potential for application in loan default prediction and lending automation through credit scoring models. This is where, a conventional algorithm-based technique like factor analysis gives a lending hand. It is able to reduce the complexities to a great extent by reducing the number of variables to a few numbers of dimensions.

When these factor values are given as input values to NN, it is able to perform dramatically well. Factor analysis alone is not able to achieve 98 percent prediction accuracy. But, factor analysis, coupled with NN, is able to deliver 98 percent prediction accuracy. This has been illustrated in this paper with the help of live data from a financial services company. The future holds for a combinatorial approach to prediction with help of linear and non-linear based techniques.

The strong side of the above study is the selection of algorithm. It is true to use a neural network in Predicting bank loan defaulters is advisable because of its performance. The neural network classification in Weka is performed in functions of multilayer perception. The advantage of the usage of neural networks for prediction is that they are able to learn from examples only and that after their learning is finished, they are able to catch hidden and strongly non-linear dependencies, even when there is a significant noise in the training set.

## III. DATA PREPROCESSING

### 3.1 Introduction

The bank data is secured and it must have to be as possible as consistent. Incomplete means lacking attribute values or certain attributes of interest, or containing only aggregate data, Noisy is containing errors, or outlier values that deviate from the expected, Inconsistent (is containing discrepancies) and Irrelevant data are the characteristics of large data base system in real world. These may happen in different way like data collection instruments leads to fault, human or computer errors occurring at data entry, errors during data transmission, inconsistent formats for input fields, such as date and duplicate tuples also another cause. These errors will make the data to be a low-quality data. And a low-quality will lead to a low-quality data mining results. Therefore, the data preparation and preprocessing to clean noisy data to handle missing values and consistent attribute must have to be taken place [5].

Before the data preparation and preprocessing this chapter begins by describing about the data source.

Source of the Data: the bank uses OminiEnterprise core banking system by incorporating all customers' data to the system. The data base administrator detaches the allowed attribute with their instances from the year of 2009, 2010 and 2011only from the OminiEnterprise data base (which is allowed by the Bank's IT department manager for the researcher).

As shown in the above table 3.1 the data is filtered by the year. Each year have the same attribute with different frequency out of total records. The bank records contain different attributes like

**Table 3.1 General information of the data**

| Year | Frequency | Percent (%) |
|---|---|---|
| 2009 | 7531 | 37 |
| 2010 | 6520 | 32 |
| 2011 | 6410 | 31 |
| Total | 20461 | 100 |

- Nominal or Categorical which means no inherent orders among values and important special case like Boolean (True/False).
- Continuous or Numeric such attributes are values corresponds with numbers with interval quantities (like integer) and ratio quantities (real).

### 3.2 Descriptive summarizations of attributes:

For data preprocessing to be successful, it is essential to have an overall picture of the data. Descriptive data summarization techniques can be used to identify the typical properties of the data and highlight which data values should be treated as noise or outliers.

Thus, the researcher wants to introduce the basic concepts of descriptive data summarization before getting into the concrete workings of data preprocessing techniques. For many data preprocessing tasks, users would like to learn about data characteristics regarding both central tendency and dispersion of the data. Measures of central tendency include mean, median, mode, and midrange, while measures of data dispersion (The degree to which numerical data tend to spread) include quartiles, interquartile range (IQR), and variance.

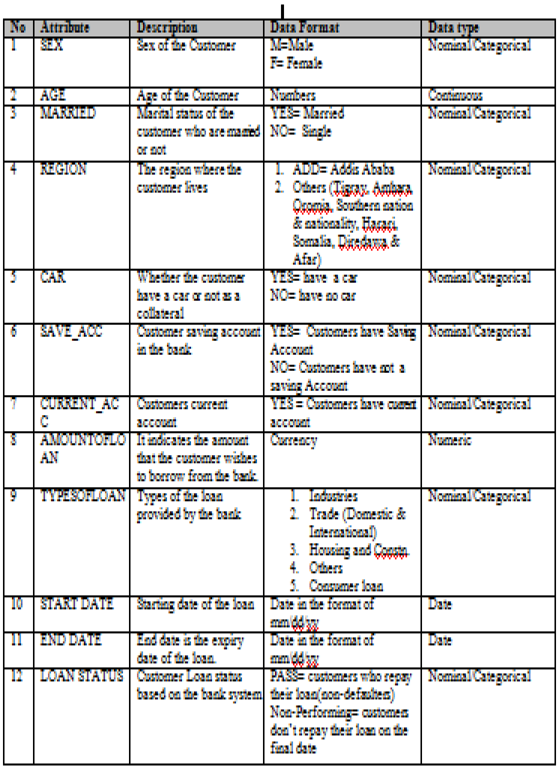
Descriptive data summarization helps us study the general characteristics of the data and identify the presence of noise or outliers, which is useful for successful data cleaning and data integration [6]. These descriptive statistics are of great help in understanding the distribution of the data. Such measures have been studied extensively in the statistical literature. From the data mining point of view, it is important to examine how they can be computed efficiently in large databases. In particular, it is necessary to introduce the notions of distributive measure, algebraic measure, and holistic measure.

Knowing what kind of measure we are dealing with can help us choose an efficient implementation for it.

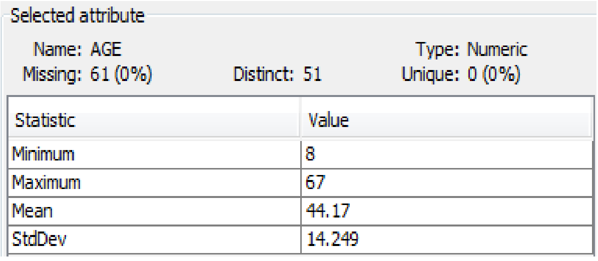
### 3.2.1 AGE Attribute

This attribute is numeric type and holds numeric age values of the bank customer. Its minimum value and Maximum value is 8 and 67 respectively. Its IQR (Inter Quartile range) is 25 which are calculated by IQR3-IQR1. The most frequent age is 64 years. The number of missing values for the attribute is 61. To handle the problem of missing values for nominal variables, replacing with modal value and for the numeric type attribute the missing values were replaced by the Mean value which is recommended by many scholars and it is also cited in Two Crows Corporation (1999) [7]. So, the missing values of this papers all attribute is performed accordingly.
The missing value of age attribute is replaced by mean age, which is 44 years, which is performed in the data set by Weka tool called *Replace Missing Value (weka. filters. unsupervised. attribute. Replace Missing Values).* The number of values which were found as outlier is 53. The researcher's decision on the outlier values to remove and replaced by the mean value.

**Table 3.2 Attributes and their Description**



| No | Attribute | Description | Data Format | Data type |
|---|---|---|---|---|
| 1 | SEX | Sex of the Customer | M=Male<br>F= Female | Nominal/Categorical |
| 2 | AGE | Age of the Customer | Numbers | Continuous |
| 3 | MARRIED | Marital status of the customer who are married or not | YES= Married<br>NO= Single | Nominal/Categorical |
| 4 | REGION | The region where the customer lives | 1. ADD= Addis Ababa<br>2. Others (Tigray, Ambara, Oromia, Southern nation & nationality, Harari, Somalia, Diredawa & Afar) | Nominal/Categorical |
| 5 | CAR | Whether the customer have a car or not as a collateral | YES= have a car<br>NO= have no car | Nominal/Categorical |
| 6 | SAVE_ACC | Customer saving account in the bank | YES= Customers have Saving Account<br>NO= Customers have not a saving Account | Nominal/Categorical |
| 7 | CURRENT_ACC | Customers current account | YES = Customers have current account | Nominal/Categorical |
| 8 | AMOUNTOFLOAN | It indicates the amount that the customer wishes to borrow from the bank. | Currency | Numeric |
| 9 | TYPESOFLOAN | Types of the loan provided by the bank | 1. Industries<br>2. Trade (Domestic & International)<br>3. Housing and Constn.<br>4. Others<br>5. Consumer loan | Nominal/Categorical |
| 10 | START DATE | Starting date of the loan | Date in the format of mm/dd/yy | Date |
| 11 | END DATE | End date is the expiry date of the loan. | Date in the format of mm/dd/yy | Date |
| 12 | LOAN STATUS | Customer Loan status based on the bank system. | PASS= customers who repay their loan(non-defaulters)<br>Non-Performing= customers don't repay their loan on the final date | Nominal/Categorical |



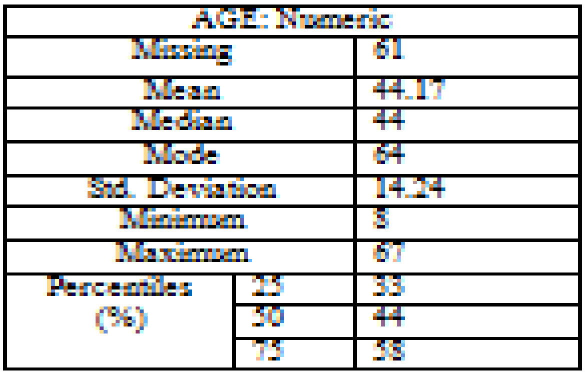
Selected attribute

Name: AGE Type: Numeric
Missing: 61 (0%) Distinct: 51 Unique: 0 (0%)

| Statistic | Value |
|---|---|
| Minimum | 8 |
| Maximum | 67 |
| Mean | 44.17 |
| StdDev | 14.249 |

**Fig. 3.1 WEKA snapshot of Age Attribute**

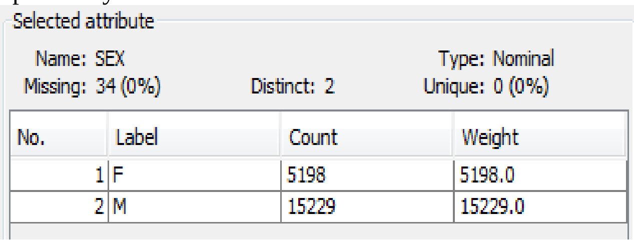
**Table 3.3: Detail Information of the AGE attribute**



| AGE: Numeric | | |
|---|---|---|
| Missing | | 61 |
| Mean | | 44.17 |
| Median | | 44 |
| Mode | | 64 |
| Std. Deviation | | 14.24 |
| Minimum | | 8 |
| Maximum | | 67 |
| Percentiles (%) | 25 | 33 |
| | 50 | 44 |
| | 75 | 58 |

### 3.2.2 SEX Attribute

As it has been described above in the attribute's description table, the SEX Attribute is nominal type. This attribute has 2 valid attribute values which mean the distinct value is 2. These are M (Male) and F (Female). The frequency of the attribute M (Male) is 15229 and F (Female) is 5198 which is explained in Fig: 3.2.
As you can see from the Fig: 3.2, the modal value for this attribute is M (Male). Different literatures recommend that the missing vales for the nominal type attribute shall be replaced by the modal value.



Selected attribute

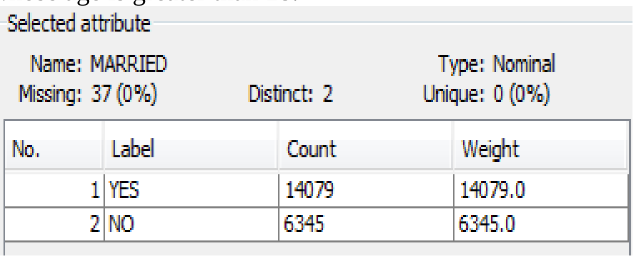
Name: SEX Type: Nominal
Missing: 34 (0%) Distinct: 2 Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | F | 5198 | 5198.0 |
| 2 | M | 15229 | 15229.0 |

**Fig: 3.2 WEKA snapshot of SEX Attribute**

### 3.2.3 MARRIED Attribute

The frequency Table of this attribute and the possible nominal values of the attribute is shown in Fig 3.3. As the Figure shows the modal value for married attribute is YES which means the customers who are "Married". The attribute has 2 distinct values according to Two Crows Corporation (1999) [7]. The missing values of the attribute are 37. The customer whose age is less than 18 years is replaced by the 'NO' value because as the norm of the nation is not practiced a person married bellow age 18 and the bank rule during loan agreement. And the modal value 'YES' is also for those whose age is greater than 18.



Selected attribute

Name: MARRIED Type: Nominal
Missing: 37 (0%) Distinct: 2 Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | YES | 14079 | 14079.0 |
| 2 | NO | 6345 | 6345.0 |

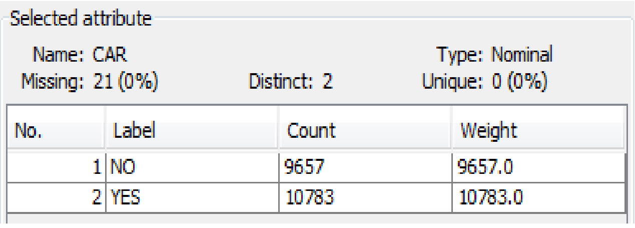
**Fig: 3.3 WEKA snapshot of MARRIED Attribute**

Retrieval Number: D2037029420/2020©BEIESP
DOI: 10.35940/ijitee.D2037.029420
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

### 3.2.4 CAR: Attribute

This attribute refers to if the customer provides a car or not for collateral purpose. It is a Nominal attribute whose distinct values is 2 “YES” to mean the customer provides a car as collateral for his loan agreement and “NO” means customer doesn’t provide a car for collateral he may provide another asset which the researcher doesn’t concern about. As depicted in the bellow table the most frequent value is “NO” and the missing value is 21. So the missing value is replaced by the modal value “NO”.

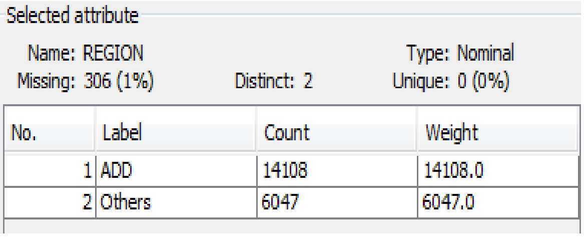

Selected attribute
Name: CAR
Missing: 21 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | NO | 9657 | 9657.0 |
| 2 | YES | 10783 | 10783.0 |

**Fig: 3.4 WEKA snapshot of CAR Attribute**

### 3.2.5 REGION: Attribute

It refers to the customer resident areas who take loan services from the bank. It has two distinct values which are ‘ADD’ means Addis Abeba and ‘Others’ which holds the rest of the region in Ethiopia namely Tigray, Amhara, Oromia, Southern Nation & Nationality, Harari, Somalia, Dredawa and Afar. The most frequent region is Addis Abeba and the missing values for the attribute are 306 (1%) from the total value. The region attribute is nominal so the researcher decided to replace the missing value by ADD because it is the one whose frequency is highest.

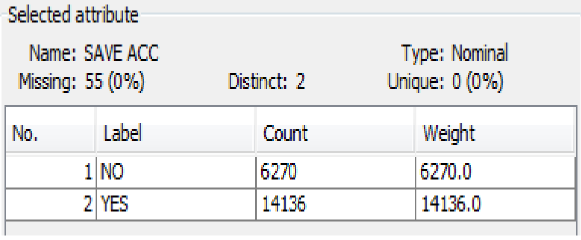

Selected attribute
Name: REGION
Missing: 306 (1%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | ADD | 14108 | 14108.0 |
| 2 | Others | 6047 | 6047.0 |

**Fig: 3.5 WEKA snapshot of REGION Attribute**

### 3.2.6 SAVE ACC: Attribute

This attribute refers if the customer has a saving account in the bank or not. It is a Nominal attribute whose distinct values is 2 “YES” and ”NO”. “YES” means the customer has saving account and “NO” means customer doesn’t have saving account at the time of loan agreement. As Fig 3.6 describes the attribute in detail the most frequent value is “YES” and the missing value is 55. So the missing value is replaced by the modal value called “YES”.


Selected attribute
Name: SAVE ACC
Missing: 55 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | NO | 6270 | 6270.0 |
| 2 | YES | 14136 | 14136.0 |

**Fig: 3.5 WEKA snapshot of SAVEACC Attribute**

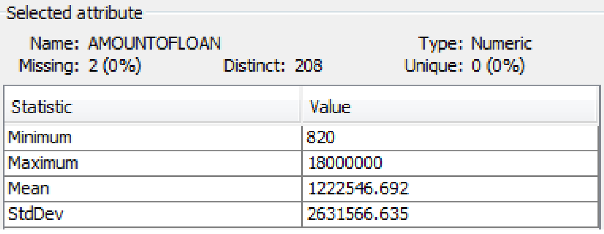
### 3.2.7 CURRENT_ACCT: Attribute

This attribute refers if the customer has a Current account in the bank or not. It is a Nominal attribute whose distinct value is 1 “YES”. It means the customer have current account at the time of loan agreement. The researcher decided to remove this attribute from the data set because every customer must have current account at the time of loan agreement so the importance of the attribute is meaningless.

### 3.2.8 AMOUNT OF LOAN: Attribute

It is a numeric attribute which refers to the amount of money which the bank approved for customer. It has 208 distinct values. The most frequent Value is 2,000,000.00. The missing values are only 2 in number and it is going to be replaced by the mean value. The attribute needs discretization as its distinct values are too much.

We can see in fig 3.6 for important statistical summary. The researcher omits the frequency table since it has too large distinct value.


Selected attribute
Name: AMOUNTOFLOAN
Missing: 2 (0%)
Distinct: 208
Type: Numeric
Unique: 0 (0%)

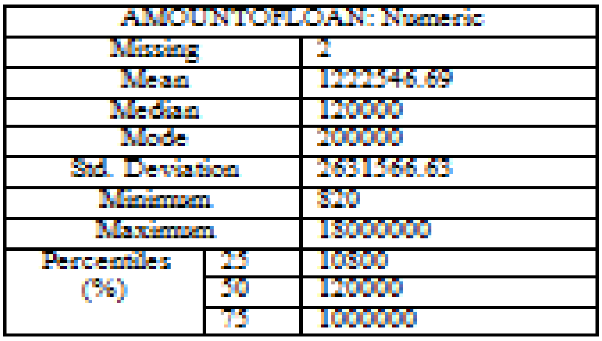
| Statistic | Value |
|---|---|
| Minimum | 820 |
| Maximum | 18000000 |
| Mean | 1222546.692 |
| StdDev | 2631566.635 |

**Fig: 3.6 WEKA snapshot of AMOUNTOFLOAN Attribute**

### 3.2.9 TYPES OF LOAN: Attribute

It is a type of loan the bank will provide for its customer. This attribute is a nominal type having seven distinct vales Industries, Trade, Housing and Constn, Consumer loan and Others. As the bellow figure shows that the modal value from the given distinct value is ‘Consumer loan’. In this Attribute there is no missing value all the tuples fill in the correct way.

**Table 3.4: Detail information of AMOUNTOFLOAN Attribute**



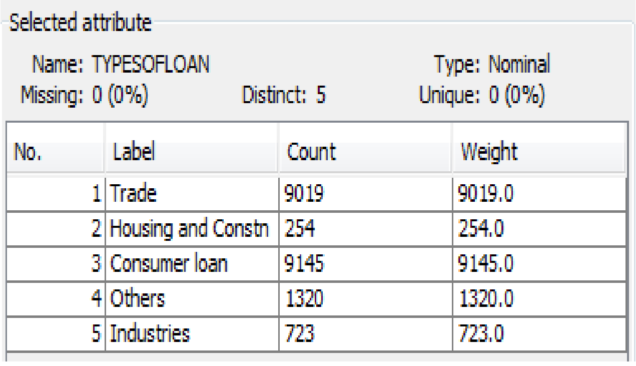
| AMOUNTOFLOAN: Numeric | | |
|---|---|---|
| Missing | | 2 |
| Mean | | 1222546.69 |
| Median | | 1200000 |
| Mode | | 2000000 |
| Std. Deviation | | 2631566.63 |
| Minimum | | 820 |
| Maximum | | 18000000 |
| Percentiles (%) | 25 | 108000 |
| | 50 | 1200000 |
| | 75 | 1000000 |

### 3.2.10 LOANDURATION: Attribute

The “LOAN_DURATION” attribute is derived attribute from the “STARTDATE” and ”ENDDATE”. It is derived as the difference between the two date attribute so that more general information will be obtained related to the duration information and it will ease the interpretation of the result.

The “STARTDATE” has 3values which are 2009, 2010 and 2011all the year is in Gregorian calendar format but the “ENDDATE” has 23 distinct values which is large to handle from the minimum year of 2009 to the maximum of 2041all are with different frequencies. So it necessary to have a derived attribute for the sake of clarity and good analysis purpose.

The new attribute is generated as a duration which means in how many years the loan will be repaid.

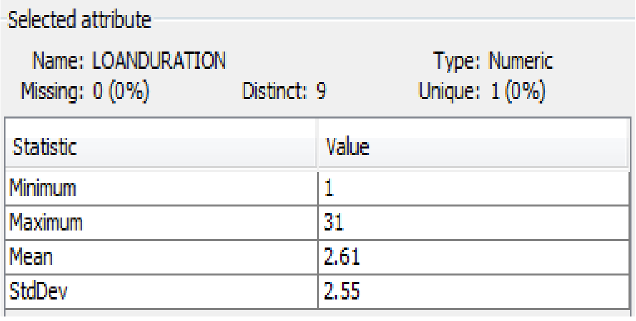


Selected attribute

Name: TYPESOFLOAN | Type: Nominal
Missing: 0 (0%) | Distinct: 5 | Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | Trade | 9019 | 9019.0 |
| 2 | Housing and Constn | 254 | 254.0 |
| 3 | Consumer loan | 9145 | 9145.0 |
| 4 | Others | 1320 | 1320.0 |
| 5 | Industries | 723 | 723.0 |

**Fig 3.7: Weka Snapshot of TYPESOFLOAN Attribute**



Selected attribute

Name: LOANDURATION | Type: Numeric
Missing: 0 (0%) | Distinct: 9 | Unique: 1 (0%)

| Statistic | Value |
|---|---|
| Minimum | 1 |
| Maximum | 31 |
| Mean | 2.61 |
| StdDev | 2.55 |

**Fig 3.8: Weka Snapshot for LOANDURATION Attribute**

As we can see from the above figure the minimum value and the maximum values are 1 and 31 respectively there is no missing value in this attribute.

### 3.2.11 STARTDATE: Attribute

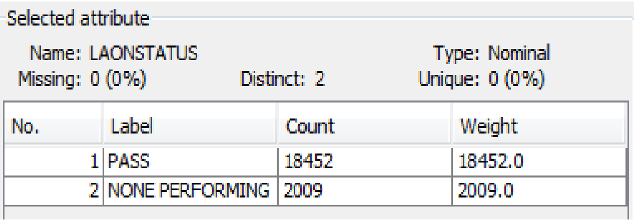
In this attribute the date format is changed in to year. Since the “STARTDATE” in the bank database is kept in the form of ‘dd/mm/yy’ which means it has date/month/year this makes the analysis difficult. So the researcher takes only the year from the format. The year is in European calendar. The beginning date for”STARTDATE” attribute is 2009 and the last year is 2011so it has three distinct values. The mode is 2009 the year in which many customers received and start their loan service from the bank.

### 3.2.12 ENDDATE: Attribute

In this attribute the date format is changed in to year the same to “STARTDATE” attribute. The year is in European calendar. The beginning date for ”ENDDATE” attribute is 2009 and the last year is 2041. The mode is 2009 the year in which many loan service end date for the customer.

### 3.2.13 LOAN STATUS: Attribute

The loan status is the status in which a customer categorized as PASS (a customer who can pay a loan) and NONE PERFORMING ( a customer who cannot pay his loan at all). So accordingly the data set has almost 2009 records out of the total 20461 record.



Selected attribute

Name: LAONSTATUS | Type: Nominal
Missing: 0 (0%) | Distinct: 2 | Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | PASS | 18452 | 18452.0 |
| 2 | NONE PERFORMING | 2009 | 2009.0 |

**Fig 3.9: Weka Snapshot for LOAN STATUS Attribute**

### 3.3 Data Cleaning

At this stage the data will be cleaned in order to fill the missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.

Therefore, a useful preprocessing step is to run your data through some data cleaning routines. In this subsection the researcher discusses methods for cleaning up data.

### 3.3.1 Handling Missing Values

Therefore, in the dataset collected for this research work, both nominal and numeric attributes missing values were handled. As it has been shown above in the statistical summary of the attributes in the bank loan dataset, there are missing values in all except ‘CURRENTACC’, ‘TYPESOFLOAN’, ‘STARTDATE’, ‘ENDDATE’ and the derived of ‘LOANDURATION’ attributes. So, excluding the above attribute the researcher tries to handle the missing value appropriately.

The attributes are: ‘SEX’, ‘REGION’, ‘CAR’, ‘SAVE ACC’ and “LOANSTATUS”. Since all the above fields are nominal variables, for any missing values in those fields, the modal (most frequent) value was used.

For ‘AGE’ attribute as the nature of the attribute is Numeric so the missing value has to be replaced by the mean value the researcher shows in detail in section 3.2.1

For ‘MARRIED’ attribute, special consideration was made according to the customers’ age and the bank rule for loan agreement. As the researcher get information from the data base administrator a person bellow age 18 is not allowed for loan agreement. For those customers whose age is below 18 years and having missing values in the ‘MARRIED’ were replaced by ‘NO’ and for those customers whose age is above 18 years, the missing values for the attribute is replaced by the most frequent value of ‘MARRIED’ attribute.

### 3.3.2 Handling Outliers Values

The degree to which numerical data tend to spread is called the dispersion, or variance of the data. The most common measures of data dispersion are range, the five-number summary (based on quartiles), the interquartile range, and the standard deviation.

By having the IQR then we calculate the rest value in order to get the upper and lower limit of possible outliers. Let’s take a look in the AGE attribute how to calculate the outliers.

As shown in the table 3.3 AGE Attribute, the third Quartile (Q3) for the AGE attribute is 58, the first quartile is 33.

The IQR is 25. Then 1.5 * IQR is 37.5years. So, by the formula of Q3+ (1.5*IQR) we calculate the upper limit as 58+37 is 95 years is the upper limit for outliers i.e. age values beyond 95 years are outliers. For the lower limit is Q1-(1.5*IQR) which is 33-37 = -4. Age values bellow -4 years can be considered outlier in this dataset. But on the 'AGE' attribute, special consideration was made according to the customers' age of the bank rule for loan agreement. As the researcher get information from the data base administrator a person bellow age 18 is not allowed for loan agreement. So, a customer whose age is below 18 is detected as outlier

### 3.3.3 Handling Noisy Values

Noise is a random error or variance in a measured variable. It may happen in the attribute in different ways. The most known and familiar noisy values are those unknown encoding will be found in the attribute like 'Sex= x', out of range values may also store in the dataset like 'Temperature =2220' and 'Age=367' and different in compatible formats will make the measured variable noisy.

**Table 3.5 attributes with their Modal value**

| **Attribute** | **Modal Value (the most frequent value)** |
|---|---|
| SEX | M |
| MARRIED | YES |
| REGION | ADD |
| CAR | NO |
| SAVEACC | YES |
| LOANSTATUS | PASS |

**Table 3.6 attributes with their Mean value**

| **Attribute** | **Mean Value** |
|---|---|
| **AGE** | **44.17** |
| **AMOUNTOFLOAN** | **1222546.69** |

In this paper the noisy values will be detected in the attribute of "AMOUNTOFLOAN", "AGE" and" REGION" and the researcher will be decided to handle such noisy values by the most recommended way.

On the "STARTDATE" there are also four noisy values like '11/30/9999' which don't support any date format. The researcher recognizes such value as a missing value and replace by the most frequent values.

### 3.3.4 Data Reduction

The researcher decides to add or construct which is a derived attribute from the 'STARTDATE' and 'ENDDATE' called 'LOANDURATION' which is explained before in Descriptive summarization of attributes in 3.2.10 LOANDURATION attribute and removes some insignificant attributes by considering their advantages in the process of finding some interesting patterns.

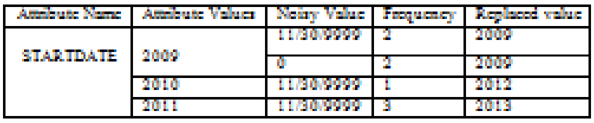
Therefore, the "APPROVAL DATE" is removed from the dataset. Since it have the same value with the "START DATE". The "CURRENT_ACCT" is also removed from the given data set because it is a must to have current account whenever one wants to apply a loan this means all the values in the attribute is "YES".

### 3.3.5 Data Integration and Transformation

In this dataset the 'AGE', 'LOANDURATION' and 'AMOUNTOFLOAN' attributes were discretized (binned) both to reduce the distinct values of the attributes in order to get the best performance out of the mining tool.

The researcher decided to carry out the discretization process on the numeric attribute of 'AGE'. On the 'AMOUNTOFLOAN' attribute simply the binning method will be applied in order to minimize the number of distinct values without the conceptual hierarchy.

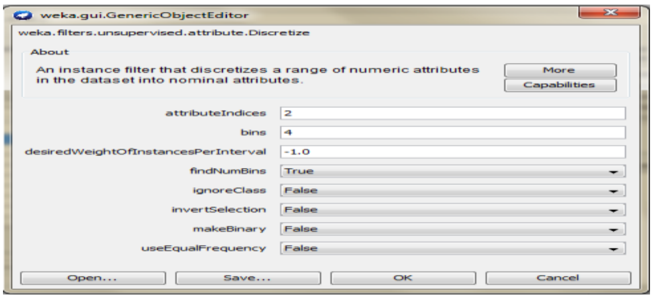
**Table 3.7 STARTDATE attributes with its Noisy value**



| Attribute Name | Attribute Values | Noisy Value | Frequency | Replaced value |
|---|---|---|---|---|
| STARTDATE | 2009 | 11/30/9999 | 2 | 2009 |
| | | 0 | 2 | 2009 |
| | 2010 | 11/30/9999 | 1 | 2012 |
| | 2011 | 11/30/9999 | 3 | 2013 |

Binning is a top-down splitting technique based on a specified number of bins. If A and B are the lowest and highest values of the attribute respectively, the width of intervals will be calculated as: W = (B-A)/N

There for the researcher will rely on WEKA software to perform discretization on the 'AGE' and 'LOANDURATION' attribute. The 'AGE' attribute divided in to 4 bins (intervals) and on the property box by turning findNumBins(Optimize number of equal-width bins using leave-one-out. Doesn't work for equal-frequency binning) to true we get summarized discretize attribute as detail information shown infigure 3.8 and 3.9 respectively.



weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Discretize

About

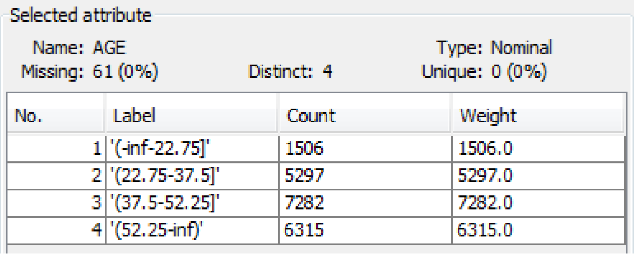
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.

attributeIndices: 2
bins: 4
desiredWeightOfInstancesPerInterval: -1.0
findNumBins: True
ignoreClass: False
invertSelection: False
makeBinary: False
useEqualFrequency: False

**Fig 3.10: Weka Snapshot for AGE attribute discretization Property box**

As we can see from the figure 3.1 of AGE Attribute, the distinct values are reported as 51 in the Descriptive summary section. But now the attribute is reduced into four labels.

For the "LOANDURATION" attribute as it has nine distinct values as stated in 3.2.10 section in Descriptive summarization section but after discretization by WEKA tool



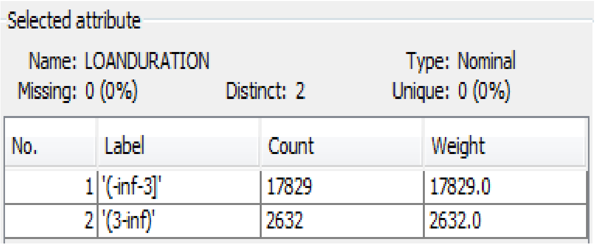
Selected attribute

Name: AGE Type: Nominal
Missing: 61 (0%) Distinct: 4 Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | '(-inf-22.75]' | 1506 | 1506.0 |
| 2 | '(22.75-37.5]' | 5297 | 5297.0 |
| 3 | '(37.5-52.25]' | 7282 | 7282.0 |
| 4 | '(52.25-inf)' | 6315 | 6315.0 |

**Fig 3.11: Weka Snapshot for AGE discretization Attribute**

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.



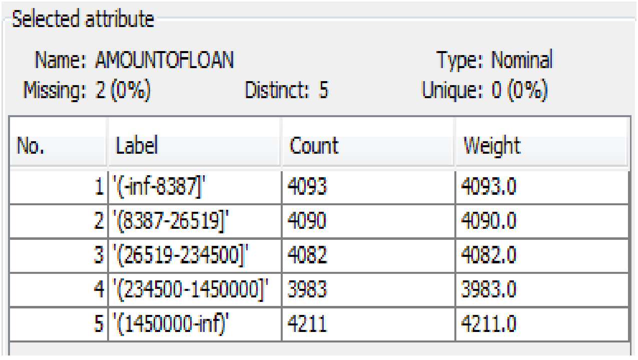
Selected attribute

Name: LOANDURATION Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | '(-inf-3]' | 17829 | 17829.0 |
| 2 | '(3-inf)' | 2632 | 2632.0 |

**Fig: 3.12 WEKA snapshot of LOANDURATION Attribute after Discretization**

As the above figure 3.12 shows it will be reduced to 2 distinct values. The result of the discretization on the "AMOUNTOFLOAN" attribute looks like the following in figure 3.12. As you can see from the Descriptive summarization section, this attribute had 208 distinct values. The Weka tool discretize in to 5 numbers of bins by setting *'useEqualFrequency'* to true which means if it is set to true, equal-frequency binning will be used instead of equal-width binning. This is done for the sake of clarity and analysis purpose.



Selected attribute

Name: AMOUNTOFLOAN Type: Nominal
Missing: 2 (0%) Distinct: 5 Unique: 0 (0%)

| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | '(-inf-8387]' | 4093 | 4093.0 |
| 2 | '(8387-26519]' | 4090 | 4090.0 |
| 3 | '(26519-234500]' | 4082 | 4082.0 |
| 4 | '(234500-1450000]' | 3983 | 3983.0 |
| 5 | '(1450000-inf)' | 4211 | 4211.0 |

**Fig: 3.13 WEKA snapshot of AMOUNTOFLOAN Attribute after Discretization**

A total of 12 attributes were selected for the research based on their relevance and preprocessing activities of the problem. There were two attributes which were excluded in the preliminary data observation. Which is 'CURRENT_ACCT', and 'APPROVALDATE', since they are no more important for the mining purpose and it is also stated in the descriptive statistical section in detail. 'LOANDURATION' attribute was newly constructed attribute from 'STARTDATE' and 'ENDDATE'.

Hence the final attributes used for model building are 10 in number which are; 'AGE', 'SEX', 'MARRIED', 'REGION', 'CAR', 'SAVE ACC', 'AMOUNTOFLOAN', 'TYPESOFLOAN', 'LOANDURATION', 'LOANSTATUS'.

## IV. METHODOLOGY

### 4. 1 Introduction

We decided to deal with classification rule problems. For the classification technique we interested to experiment both binary decision tree and generalized decision tree with and without pruning and Simple Bayesian network with and without balanced data set. So it is necessary to explain in detail the two selected approaches. Hence, it is important to explain the classification rule problems implementations for model building and experiments to be carried out in the data mining process, which also involve data mining tool selection and algorithms used for modeling. The classification model to be built is decision tree and Bayesian Networks. The researcher is interested to experiment both binary decision tree and generalized decision tree with and without pruning. That means there are four scenarios to be experimented for classification purpose and Simple Bayesian Networks by using the Naïve Bayes algorithm which are explained latter.

### 4.2 Decision Tree:

A decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The top most nodes in a tree are the root node.

The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy[8].

Algorithms for Generate a decision tree from the training tuples of data partition D.

Data partition, D: This is a set of training tuples and their associated class labels.

Attribute list: the set of candidate attributes.

Attribute selection method: a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a splitting attribute and possibly, either a split point or splitting subset. This procedure employs an attribute selection measure, such as information gain or the gain index. Whether the tree is strictly binary is generally driven by the attribute selection measure. Some attribute selection measures, such as the gini index, enforce the resulting tree to be binary. Others, like information gain, do not, there in allowing multi-way splits (i.e., two or more branches to be grown from a node) [8].

### 4.2.1 Attribute Selection

This chapter starts to discuss on the model building of the paper in detail. There are different kinds attribute selection algorithm in data mining of which the most popular attribute selections measures are Information Gain, Gain Ratio, and Gini Index. Attribute selection sometimes called as Feature selection is one of the major tasks in decision tree because it will enhance the accuracy and it will help us for selecting a subset of relevant features for building robust learning models. The researcher decides to use the Information gain measures. Then first explained the selected attribute selection algorithm then proceed to the weka result by using the selected attribute.

### 4.2.2 The J48 Decision Tree Algorithm

For classification purpose with implementation J48 classifier, parameters are changed from the weka Generic ObjectEditor window as indicated on figure 3.13 for building different decision trees. The above figure 3.13 windows give a way for selecting whether to build binary or generalized decision tree. The parameter used for this purpose is ‘*binarySplit*’; which explained in the above option description at number 1. Another important parameter on this window is relevant to this research is ‘*unpruned*’, again it also explained on the above option description at number 11.

So, implementing decision tree classification in this research is binary and generalized decision tree. Binary trees split internal node branches to exactly two sub trees or we will have two branches at each node. A binary tree is a tree with one more restriction: no node may have more than 2 children as compared to that the generalized tree one when a node can carry a minimum of two and maximum of many children [9].

We implemented binary decision tree with the following scenarios:

1. Binary decision tree with all attributes without pruning
2. Binary decision tree with all attributes with pruning
3. Binary decision tree with some selected attributes without pruning
4. Binary decision tree with some selected attributes with pruning.

When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of overfitting the data. Such methods typically use statistical measures to remove the least reliable branches.

A generalized decision tree model is built by setting the ‘binarySplit’ to ‘False’ so that a single node can be splited into more than two sub trees. But those tree branches may be of less importance to reveal important and concise knowledge as discussed above in decision tree building section.

Like the binary decision tree, the researcher interested to build generalized decision tree to compare its efficiency as compared to other scenarios experimented in building the decision tree classification model. Therefore, there are also four scenarios for generalized decision tree experimented in this research these are:

1. Generalized decision tree with all attributes with pruning
2. Generalized decision tree with all attributes without pruning
3. Generalized decision tree with some of the attributes with pruning and
4. Generalized decision tree with some of the attributes without pruning.

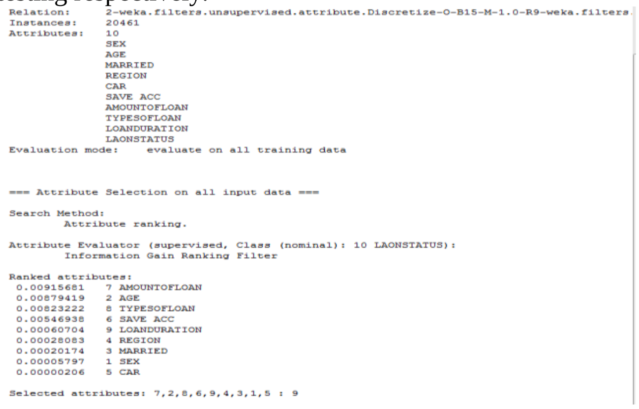
Hence binary and generalized decision tree with and without pruning using all and selected attribute is implemented by using J48 classifier in Weka. The J48 is explained before but the experimentation result will be analyzed in the next chapter.

As we can see from the result of attribute selection using entropy-based information gain ranking filter method of Weka. The determining attributes of the dataset for predicting for loan defaults by the customer loan status are ranked accordingly, ‘AMOUNTOFLOAN’, ‘AGE’, ‘TYPESOFLOAN’, ‘SAVE ACC’, ‘LOANDURATION’, ‘REGION’, ‘MARRIED’, ‘SEX’, ‘CAR’ by their ascending order. The class label attribute is ‘LOANSTATUS’ which is out of the ranked attributes. This process is useful because it will help us for latter experimentation by excluding the least relevant attributes.

### 4.2.3 Selection of Validation Method for Decision Tree Models

Now consider what to do when the amount of data for training and testing is limited. The holdout method reserves a certain amount for testing and uses the remainder for training. In practical terms, it is common to hold out one-third of the data for testing and use the remaining two-thirds for training.

The researcher tries to use the WEKA tool in order to select the validation method for the decision tree which is completely done by the WEKA tool performing on the total dataset. The first step to use is selecting the split percentage in which it starts by making the percent in to 10 to 90. When it makes to 10% the result we get from the machine is 99% Performance which is to mean that it is a Correctly Classified Instances. So, the researcher applies all the percentage in the bellow table with 10 value differences from 0 to 100 and get the same performance or correctly classified instances. Finally, the researcher decides to use the well recommended way of selection validation which is 70:30 for training and testing respectively.



```
Relation:     2-weka.filters.unsupervised.attribute.Discretize-O-B15-M-1.0-R9-weka.filters.
Instances:    20461
Attributes:   10
              SEX
              AGE
              MARRIED
              REGION
              CAR
              SAVE ACC
              AMOUNTOFLOAN
              TYPESOFLOAN
              LOANDURATION
              LAONSTATUS
Evaluation mode:    evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 10 LAONSTATUS):
        Information Gain Ranking Filter

Ranked attributes:
 0.00915681   7 AMOUNTOFLOAN
 0.00879419   2 AGE
 0.00823222   8 TYPESOFLOAN
 0.00546938   6 SAVE ACC
 0.00060704   9 LOANDURATION
 0.00028083   4 REGION
 0.00020174   3 MARRIED
 0.00005797   1 SEX
 0.00000206   5 CAR

Selected attributes: 7,2,8,6,9,4,3,1,5 : 9
```

**Fig 4.3 Weka snapshot of Attribute selection**

In the above figure 4.4 shows the learning curve which shows what the performance and sample data interact each other. Unfortunately, this is only really plausible with a 50: 50 split between training and test data, which is generally not ideal. It is better to use more than half the data for training and the rest for test data. So the researcher decided to use 70% for the training and 30% for the test data.

**Table 4.1: Sample Data and performance table**

| Sample | Performance (%) |
|---:|---:|
| 10% | 99 |
| 20% | 99 |
| 30% | 99 |
| 40% | 99 |
| 50% | 99 |
| 60% | 99 |
| 70% | 99 |
| 80% | 99 |
| 90% | 99 |

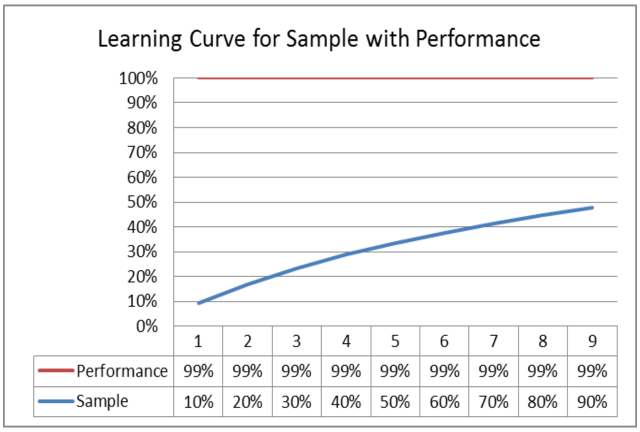


**Fig 4.4: Learning curve for sample data with Performance**

### 4.3 The Bayesian Network

The Bayesian network has different kinds of algorithms for classification purpose. But the most familiar algorithms for prediction are the NaïveBayes and BayesNet.

### 4.3.1 Balanced and unbalanced data set

The data set is highly unbalanced with 18452 records in 'PASS' class and 2009 records in 'NONE PERFORMING' class. This forces the researcher in order to make balanced data set. Because the minimum class value is 2009 so by taking 2009 records from each of the classes and finally 4018 total records will be ready for the Naïve Bayes experimentation. This is done by Weka *SpreadSubsample* technique. On the property box increase the *maxCount* to 2009. This is the value of the minimum class value. *weka.filters.supervised.instance.SpreadSubsample.*

## V. RESULT AND DISCUSSION

### 5. 1 Experiment and Analysis of Decision tree

Analysis of the decision tree models are made in terms of detailed accuracy of the classifier on the training dataset as tested on the test data based on a confusion matrix of each model result. The confusion matrix is a useful tool for analyzing how well our classifier can recognize tuples of different classes. The class level name is 'LOANSTATUS' by having values called 'PASS' and 'NONE PERFORMING'. Which is 'PASS' means a customer who pays a loan and 'NONE PERFORMING' means a customer who doesn't pay a loan.

There are eight experiments to be analyzed for decision tree classification. As stated on the previous chapter these Processes are going to be experimented and analyzed to compare them to each other in terms of different

1. Performance matrices values
2. Accuracies
3. Number of leaves, and size of tree generated
4. ROC curves and execution time.

The process for decision tree classification experimented in this research areas are listed below:

For Binary decision tree process

i. Binary Decision Tree without pruning with All Attribute
ii. Binary Decision Tree without pruning with reduced Attribute
iii. Binary Decision Tree with pruning with All Attribute
iv. Binary Decision Tree with pruning with reduced Attribute

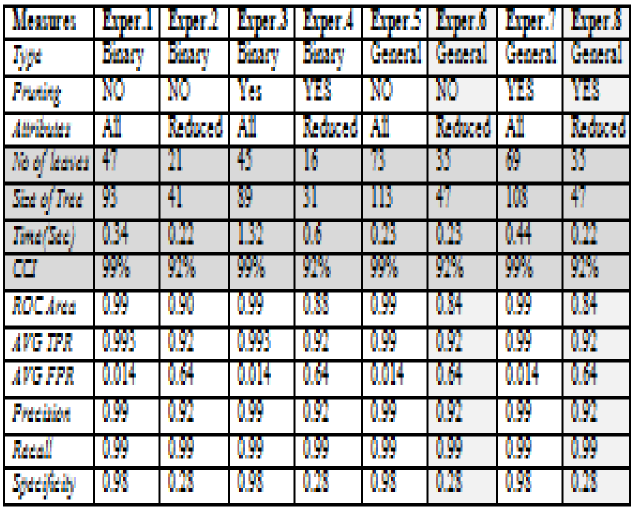
For Generalized decision tree process

i. General Decision Tree without pruning with all Attribute
ii. General Decision Tree without pruning with reduced Attribute
iii. General Decision Tree with pruning with all Attribute
iv. General Decision Tree with pruning with reduced Attribute

The researcher is going to see the models results and the analysis of each result and compare the result of one model to the other and finally find out the outperforming model based on the criteria of evaluation.

As it is selected in the previous chapter of the method validation it is good to use 70% of the training set splitting and 30% percent for test dataset

**Table 5.1 Detail information of all experiment**



| Measures | Exper.1 | Exper.2 | Exper.3 | Exper.4 | Exper.5 | Exper.6 | Exper.7 | Exper.8 |
|---|---|---|---|---|---|---|---|---|
| Type | Binary | Binary | Binary | Binary | General | General | General | General |
| Pruning | NO | NO | Yes | YES | NO | NO | YES | YES |
| Attributes | All | Reduced | All | Reduced | All | Reduced | All | Reduced |
| No of leaves | 47 | 21 | 45 | 16 | 73 | 35 | 69 | 35 |
| Size of Tree | 93 | 41 | 89 | 31 | 113 | 47 | 108 | 47 |
| Time(Sec) | 0.34 | 0.22 | 1.32 | 0.6 | 0.23 | 0.23 | 0.44 | 0.22 |
| CCI | 99% | 92% | 99% | 92% | 99% | 92% | 99% | 92% |
| ROC Area | 0.99 | 0.90 | 0.99 | 0.88 | 0.99 | 0.84 | 0.99 | 0.84 |
| AVG TPR | 0.993 | 0.92 | 0.993 | 0.92 | 0.99 | 0.92 | 0.99 | 0.92 |
| AVG FPR | 0.014 | 0.64 | 0.014 | 0.64 | 0.014 | 0.64 | 0.014 | 0.64 |
| Precision | 0.99 | 0.92 | 0.99 | 0.92 | 0.99 | 0.92 | 0.99 | 0.92 |
| Recall | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| Specificity | 0.98 | 0.28 | 0.98 | 0.28 | 0.98 | 0.28 | 0.98 | 0.28 |

### 5.1.1 Binary Decision Tree Experiment

All experiments compared in one table as shown in the bellow table.

Based on the above total experiment table let's take a look in the bellow tree size and no of leaves and execution time graphs the it is possible to select the best easily interpretable decision tree.

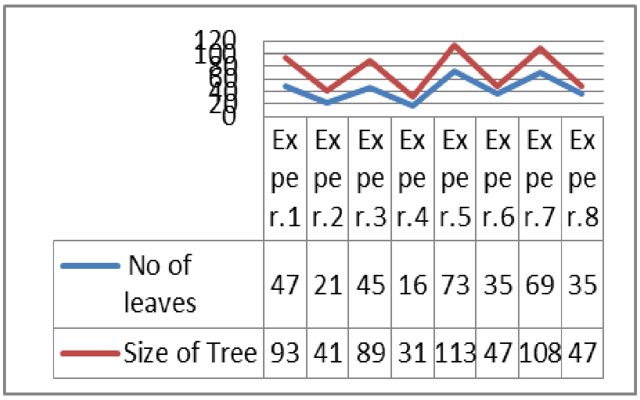


**Fig 5.1: Graphs for tree complexity for all experiment**

Experiment 1,3,5 and 7 has high performance better than the rest of experiment and their size of the tree and no of leaves is again high with the rest of experiment this makes the tree not easily interpretable when we compare with the rest. Even if it has large no of leaves and size of the tree the researcher decided to select one from the experiment with the highest performance by comparing their execution time, no of leaves and size of the tree eah other from experiment 1, 3, 5 and 7

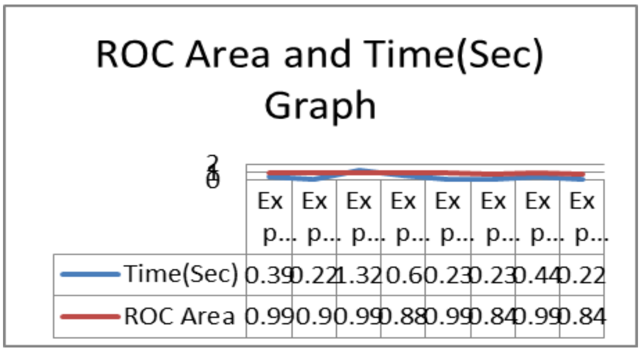


**Fig 5.2: Graphs for execution time and ROC Area for all experiment**

It is possible to say Binary decision tree without pruning with all attribute is easy and interpretable tree structure with high performance. So the knowledge obtained from the pruned Binary decision tree with all attributes model has got meaningful contributions for default prediction which is stated in the bellow paragraph.

From the decision tree developed in the aforementioned experiments, it is possible to find out a set of rules simply by traversing the decision tree and generating a rule for each leaf and making a combination of all the tests found on the path from the root to the leaf node[10]. This produces rules that are unambiguous in that it doesn't matter in what order they are executed. The following are some of the rules extracted from the decision tree.

- If a person whose AGE is between 22-37 in which he/she applied for housing and construction types of Loan is categorized as NONE PERFORMING (Defaulters) at all. But the same person whose AGE is not in between 22-37 is categorized as PASS (not defaulters) which means they will repay the loan.
- If a person who is applied for a loan to repay in less than 3 years but the types of loan is different from Housing and construction, Industries and Others in which he bring a CAR as collateral and he is married and male borrowing a loan with the amount of less than 8387 birr where he lives in Addis Ababa is more likely repay his loan so the classifier categorized as PASS but the same person when he lives out of Addis Ababa (In different regions of Ethiopia) has 84% probability to be a defaulters (NONE PERFORMING).
- If the type of Loan is different from Housing and Const. and amount of loan is less than 1,450,000 birr and if it has a saving account whose age is different from 37-52 and loan duration is greater than 3 years will repay their loan based on their agreement.
- If the type of Loan is different from Housing and Const. and amount of loan is less than 8387 birr and their Age is less than 22 years old. Will repay their loan which means they have a chance to become a defaulter.
- If the type of Loan is different from Housing and Const. and amount of loan is less than 8387 birr and if the age is between 22-37 years old and if they doesn't have a car as collateral they will probably repay their loan.
- If the type of Loan is different from Housing and Const. and amount of loan is less than 8387 birr and if the age is between 22-37 years old and if they have a car as collateral. And if they are married and their loan duration to repay is less than 3 years is more likely a defaulter. But when the loan duration is greater than 3 years they are not defaulters.

The above sample interpretation of the tree is derived from Experiment 1 tree structure which is Pruned binary decision tree with all attribute and it is depicted in Appendix 1.

### 5.2 Experiment and Analysis of Bayesian Network

There are two experiments to be analyzed for Bayesian network based on their class of 'PASS' and 'NONE PERFORMING'. As stated on the previous chapter these Processes are going to be experimented and analyzed to compare them to each other.

The experimentations are performed on the balanced and unbalanced data set which is explained in chapter four. The algorithms that the researcher uses is a Naïve Bayes classification or sometimes called as simple Bayesian network is a special form of Bayesian network, which has been widely used for data classification.

The researcher is going to see the models results and the analysis of each result and compare the result of balanced and unbalanced data set and finally find out the best performance model based on the criteria of evaluation.

As it is selected for the decision tree classification in the previous section of the method validation it is also good to use 70% of the training set splitting and 30% percent for test dataset in this Naïve Bayes classification.

By comparing the two experiments based on their performance, execution time, recall precision, ROC Area and Recall it is possible to make the correct comparison.

**Table 5.2 Detail information for the two experiments**

| **Measures** | **Experiment 1 Result** | **Experiment 2 Result** |
|---|---|---|
| *Type* | unbalanced Data set | Balanced Data set |
| *Time(Sec)* | 0.06 Seconds | 0.03 Seconds |
| **CCI* | 89 % | 77 % |
| *ROC Area* | 0.77 | 0.78 |
| *AVG TPR* | 0.89 | 0.77 |
| *AVG FPR* | 0.90 | 0.21 |

**CCI Correctly Classified Instance*

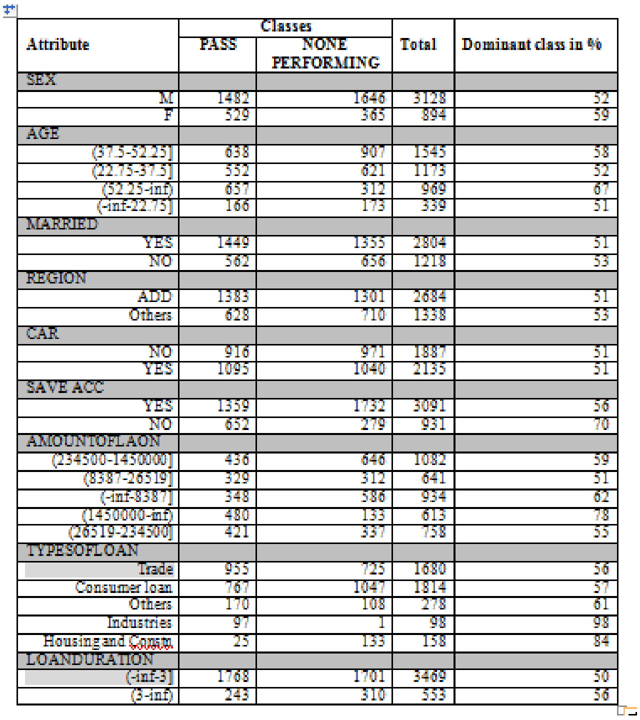
So as shown in the above table 5.2 the execution time, performance, AVG TPR and AVG FPR of the unbalanced data set is higher than the balanced data set which holds balanced data set. But the experiment 2 ROC Area is better than that of experiment 1. So, the researcher decided to use experiment 2 for classifying bank loan data set by using naïve Bayesian network. Because it is good to use the balanced data set in order to get equal distribution of tuples/record in the class level called 'LOANSTATUS'. This means it is also helpful to avoid the chance of 0.9 'PASS' class over 0.1 'NONE PERFORMING' class.

This leads the classifiers classify as 'PASS' in all cases.

The bellow table 5.3 shows that the ratio of each attribute values in each class. The other point is the dominant class which is printed in percent.

- A male Customer is become a defaulter with a probability of 52 percent but a female customer will repay their loan in 59% probability. This means females are better than male customer in repaying their loan.
- A person who takes the Industries TYPESOFLOAN is likely repaid his loan in 98% of probability. This implies a bank have no risky borrower in Industries types of loan. Since its risk is 2%.

**Table 5.3 Analysis of Naïve Bayes on balanced data**



| Attribute | Classes |  | Total | Dominant class in % |
|---|---|---|---|---|
|  | PASS | NONE PERFORMING |  |  |
| SEX |  |  |  |  |
| M | 1482 | 1646 | 3128 | 52 |
| F | 529 | 365 | 894 | 59 |
| AGE |  |  |  |  |
| (37.5-52.25] | 638 | 907 | 1545 | 58 |
| (22.75-37.5] | 552 | 621 | 1173 | 52 |
| (52.25-inf) | 657 | 312 | 969 | 67 |
| (-inf-22.75] | 166 | 173 | 339 | 51 |
| MARRIED |  |  |  |  |
| YES | 1449 | 1355 | 2804 | 51 |
| NO | 562 | 656 | 1218 | 53 |
| REGION |  |  |  |  |
| ADD | 1383 | 1301 | 2684 | 51 |
| Others | 628 | 710 | 1338 | 53 |
| CAR |  |  |  |  |
| NO | 916 | 971 | 1887 | 51 |
| YES | 1095 | 1040 | 2135 | 51 |
| SAVE ACC |  |  |  |  |
| YES | 1359 | 1732 | 3091 | 56 |
| NO | 652 | 279 | 931 | 70 |
| AMOUNTOFLAON |  |  |  |  |
| (234500-1450000] | 436 | 646 | 1082 | 59 |
| (8387-26519] | 329 | 312 | 641 | 51 |
| (-inf-8387] | 348 | 586 | 934 | 62 |
| (1450000-inf) | 480 | 133 | 613 | 78 |
| (26519-234500] | 421 | 337 | 758 | 55 |
| TYPESOFLOAN |  |  |  |  |
| Trade | 955 | 725 | 1680 | 56 |
| Consumer loan | 767 | 1047 | 1814 | 57 |
| Others | 170 | 108 | 278 | 61 |
| Industries | 97 | 1 | 98 | 98 |
| Housing and Constn | 25 | 133 | 158 | 84 |
| LOANDURATION |  |  |  |  |
| (-inf-3] | 1768 | 1701 | 3469 | 50 |
| (3-inf) | 243 | 310 | 553 | 56 |

- When a customer takes Housing and const. loan type. It is relatively a defaulter with a probability of 84% when we compare with other loan type. This means the bank should have to allow the loan agreement carefully.
- A customer whose age is greater than 52 is better age for repaid their loan than the other age since it scores 67% probability of repaying the loan.
- Less than 3 years of loan duration is preferable than more than 3 years of loan duration in order to not to be risky borrower.
- Customer who lives in Addis Ababa is most likely better than the regional customer since 51% of the customer is a PASS class. This means they repaid regularly based on their schedule.
- Since a bank has different amount of money granted for their customer, they should consider further details when the grant is less than 8387 and between 234500 and 1450000 amounts of money. Because these categories have relatively risky according to this research a customer who makes a loan agreement in this category may repay their loan in 62 and 59 percent probability respectively. It is considerably small when we compare with customers who make a loan agreement in a category greater than 1450000 amounts of loan because they have a probability of 78% percent repay their loan regularly.

## VI. CONCLUSION

The purpose of this study was to explore the applicability of data mining techniques to predict loan defaulters in the case of Wegagen bank in Ethiopia. A total of 20461 sample data were taken and tested. The results of the experiments carried out in this research using decision tree and Naïve Bayes in Weka 3.7 tool.

From the total decision tree experiment a classification which was performed in all attribute have better performance than that of the reduced attribute but in terms of number of leaves and size of the tree a classification which is performed in reduced attribute is better than on all attribute. The researchers decided to make their intention on performance. So, the Binary Decision tree without pruning with all attribute was better in performance than experiment 2, 4, 6 and 8. It was also good in understanding the decision tree with easy interpretability than experiment 3, 5 and 7.

On the other hand, the results of Naïve Bayes experiment also carried out on the same bank loan data set. There were two experiments performed, the first experiment was Naïve Bayes classification with unbalanced dataset and the second experiment is Naïve Bayes classification with balanced dataset. The 'NONE PERFORMING' class has 2009 records whereas 'PASS' class has 18452 records. This distribution of the data forces the researcher to balance the data by using Weka tools. The performances of the 2 experiments were different. Even if this performance makes the experiment 1 to be selected, but we decided to use experiment 2 means balanced dataset. This is because of to remove the dominant class value in the unbalanced data set. So the Naïve Bayes classification uses the balanced data set for analysis.

So in order to conclude this paper if the bank sector uses the data mining application it is better to use the decision tree because the decision tree produce good result to understand easily the characteristics of loan customer behavior specially loan defaulter.

## REFERENCE

1. 1. Bhambri, V. (2011). Application of Datamining in Banking Sector. *2* (2).
2. T.Larose, D. (2005). *Discovering Knowledge in Data: An Intriduction to Data Mining.* John Wiley & Sons Incorporation.
3. Qiwei Gan, B. L. (2008). Identifying Potential Default loan applicants: A case study of Consumer Credit Decision for Chinese Commercial Bank.
4. M.J.Xavier, S. P. (n.d.). Improving prediction accuracy of loan default: A Case in rural Credit.
5. Dass, R. (n.d.). Data Mining in Banking and Finance: A Note Book for Bankers.
6. Han, J. (2006). *Data Mining: Concepts and Techniques* (Second edition ed.). Elsevier Incorporation.
7. Two Crows Corporation.1999. Introduction To Data Mining And Knowledge Discovery.
8. T.Domingo, R. (2011, February). *Applying Data Mining to Banking.* Retrieved from www.rtdonline.com: www.rtdonline.com/BMA/BSM/4.html
9. S.Linoff, G. (2004). *Data Mining Techniques: For Marketting, sales and Customer Relationship Management* (Second edition ed.). Indianapolis, Indiana: Wiley Publishing Incorporation.
10. Bao, H. (2003). *Knowledge Discovery and Data Mining Techniques and Practice*. Retrieved from www.netnam.vn: https://www.netnam.vn/unescocourse/knowledge/3-1.htm

.