A Decision Support System for Predicting Bank Target Clients: Comparison between Decision Tree and Artificial Neural Network

Sukontip Wongpun[1], Tharitsaya Kongkaew[2]

[1]Information Systems Department, Faculty of Business Administration
[2]Accounting and Finance System Department, Faculty of Business Administration
Rajamangala University of Technology Thanyaburi
Pathumthani, Thailand
e-mail: sukontip_w@rmutt.ac.th[1], tharitsaya@rmutt.ac.th[2]

Abstract— nowadays, many banks face a high cost of operation for finding bank clients via telemarketing. A possible solution to solve this problem is finding an effective technology to support work processes and decision making. The effective technology to support bank operation helps banking business continuously growing, survive and make a profit. Therefore, this work proposed a model for prediction an opportunity of the customer to be a target customer of a new bank campaign to reduce the operation cost of finding a new customer. The results from the model were used to develop a decision support system program for predicting target bank clients to promote marketing campaigns via phone call. The system focuses on to predict a possible rate of the customer to subscribe to a bank deposit. The proposed model has compared an accuracy rate between decision tree classification algorithm and artificial neural network (ANN) to predict bank target clients who subscribe bank deposit. The results showed that the decision tree algorithm to predict a bank target client which improved the effectiveness by applying a fuzzy set was high accuracy rate at 88.38%.

Keywords-bank target clients; classification; decision tree; artificial neural network, fuzzy set; decision support system

I. INTRODUCTION

Nowadays, the number of market bank campaign is increasing dramatically because all of the bank businesses need to receive a profit and money to support their business processes. Many banks try to provide new services such as e-banking, e-payment, insurance, credit card, loan, and deposit account to their customers and try to find a new method for sale those services.

Many bank marketing campaigns have created to provide for their customers. A direct sale via phone call (telemarketing) is a sample of marketing strategy that banks try to use to enhance sales balance. However, there are not easy to increase sale balance under economic pressure and high competition. Many banks face with a costly and wasted time of operation for finding target customer. Most customers have denied buying a new campaign via phone call sale. These problems lead to a high cost of operation and less profit of banking business.

Therefore, banking business needs computerized support of managerial decision making to show who will have a chance to buy a new bank campaign before operate telemarketing, it will help to reduce the operation cost. Decision support system or business intelligence systems are required for banking business to survive and reduce operation cost. A business intelligence system proposes a suitable solution for supporting a decision making of human in operating the business process and an estimate of what is going to happen in the future [1].

Many machine-learning models were used to analyze customers personal and behavioral data of customers to predict customer who is expected to buy/cancel or have a retention rate to be a target bank customer [2]-[4]. For example, the problem of customer churn which is an important issue of all bank firms. Research of [2] compared ten analytical techniques for predicting a customer who expected to churn. A decision tree algorithm shows the high accuracy rate with 90%, random forest and ADA boost provided the highest accuracy rate with 96%.

Another issue of performing banking business is to find a new target customer for buying or subscribing their new products. Research of [3] proposed a decision tree model for analyzing factors of bank customers who were subscribed to a fixed deposit. The results showed that influenced factors of a bank customer who will deposit were a number of employees, duration and month. This research used a decision tree algorithm to find a list of influenced factors which the results confirmed that decision tree provided a high accuracy rate. Research of [4] proposed a comparing the most known machine learning models to optimize for predicting a telemarketing target calls for selling bank long-term deposits. Their

experimental found that Artificial Neural Network (ANN) and Support Vector Machines (SVM) have the best performance.

Consequently, this research proposes to use data mining techniques for predicting target bank customers by comparing the most effective data mining techniques between decision tree and ANN. The results can use for creating a plan and help to decision making for calling to a high chance of target bank customer who will buy a new bank product via telemarketing. The data set to use in this research was a Portuguese banking institution [5].

II. METHODOLOGY

A goal of this work is to propose a model for predicting target bank customers to promote marketing campaigns via telemarketing. This research used an open historical data of bank client. The data collected from a Portuguese banking institution which open public on UCI Machine Learning Repository [5]. The number of Instances was 45,211 records which have 16 input variable and 1 output variable input variables: 16 attributes and 1 Output variable as shown in table I.

A proposed prediction model has applied the Cross-Industry Standard Process for Data Mining (CRISP-DM) [6] as show in figure 1. The proposed prediction model has six steps as follow.

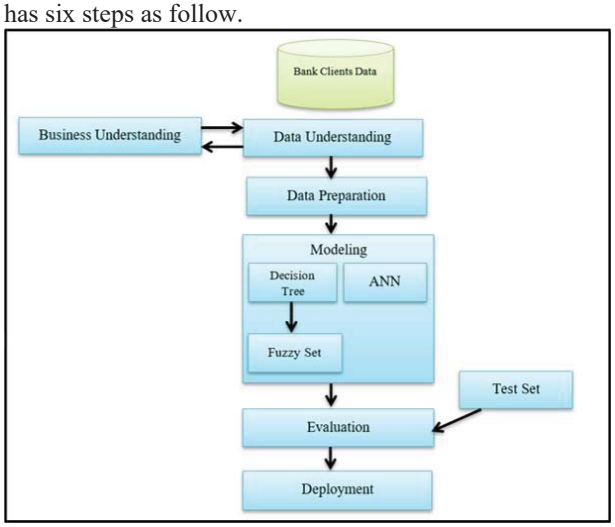

Figure 1. A model for prediction of target bank customers

- *Step 1: Data Understanding*: This experiment used a data set of a Portuguese banking institution.
- *Step 2: Data Preparation:* To prepare the data set, the data set was prepared by removing an incomplete record.
- *Step 3: Model Building:* The step has a goal to build a prediction model by comparing an accuracy rate with two data mining techniques of Decision Tree C4.5 and ANN.
- *Step 4: Evaluation:* This step is to evaluate an accuracy rate between two data mining techniques include decision tree and ANN. It is an important step to identify which algorithm suited to use for predicting bank target customers via telemarketing. The evaluation shows the most effective algorithm and the proportions of training and test data set to optimize in the model.
- *Step 5: Deployment:* After the model was generated and provided the results, this step uses that results and knowledge to deploy in real use.

TABLE I. ATTRIBUTE TYPE AND VALUES

Attribute	Data Type	Values
1 - age	N	
2 - job	C	'admin.','blue-collar', 'entrepreneur','housemaid','management','retired',..
3 - marital status	C	'divorced','married','single','unknown'
4 - education	C	'basic.4y','basic.6y','basic.9y','high.school','illiterate',..
5 - default: has credit in default	C	'no','yes','unknown'
6 - balance: average yearly balance	C	'no','yes','unknown'
7 - housing loan	C	'no','yes','unknown'
8 - personal loan	C	'no','yes','unknown'
9 - contact:	C	'cellular','telephone'
10 - month: last contact month of year	C	'jan', 'feb', 'mar', ..., 'nov', 'dec'
11 - day: last contact day	I	1-31
12 – duration: last contact duration	N	second
13 - number of contacts performed during this campaign and for this client	N	includes last contact
14 - pdays: number of days that passed by after the client was last contacted from a previous campaign	N	999 means clients were not previously contacted
15 - previous: number of contacts performed before this campaign	N	
16 - outcome of the previous campaign	C	'failure','nonexistent', 'success'
17- desired target	B	"yes","no"

**Note; N is numeric attribute, C is categorical attribute, I is integer attribute, and B is binary attribute.*

III. EXPERIMENT AND RESULTS

This work uses Rapid Miner Software to do an experiment. The experimental followed the step of the proposed model.

The data set of input variables was analyzed by using statistic values (e.g. frequency, min, max, and average) to understand the range and value scopes of each attribute. Then, the next step was cleaning the data set such as remove an incomplete record. Then, the experiment was set by splitting the data set into five proportions between training and test data set as below.

- Training set 50% and test set 50%
- Training set 60% and test set 40%
- Training set 70% and test set 30%
- Training set 80% and test set 20%
- Training set 90% and test set 10%

Figure 2 and 3 show a decision tree model for predicting a list of bank target customer. The decision tree algorithm was run by uses entropy to calculate for splitting the data set.

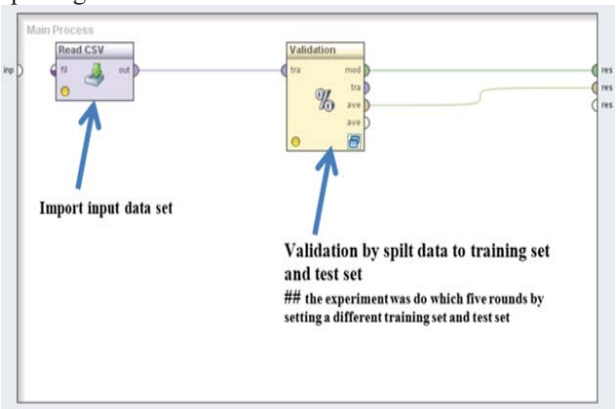

Figure 2. Import data and validation for decision tree model

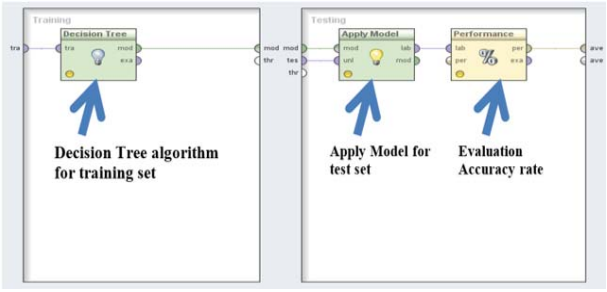

Figure 3. A decision tree model

After running the decision tree model, the results found that the split data set (training set 80% and test set 20%) has the highest accuracy rate (89.10%). Only three attributes (duration, pdays, and age) from sixteen attributes were extracted to use for classification a target customer. The rule set was extracted form Decision Tree as shown below.

Rule Set

duration > 827.500
| pdays > 495.500: no {no=3, yes=0}
| pdays ≤ 495.500: yes {no=821, yes=1137}
duration ≤ 827.500
| age > 89.500
| | age > 93.500: no {no=2, yes=1}
| | age ≤ 93.500: yes {no=0, yes=5}
| age ≤ 89.500: no {no=43096, yes=4667}

A goal of this research is to find an algorithm which has a high accuracy rate of predicting target customers. Therefore, the next step was to run the ANN algorithm. The ANN algorithm performed by setting 0.01 learning rate and two hidden layers to generate the results. Figure 4 and 5 show an ANN for predicting a list of bank target customer.

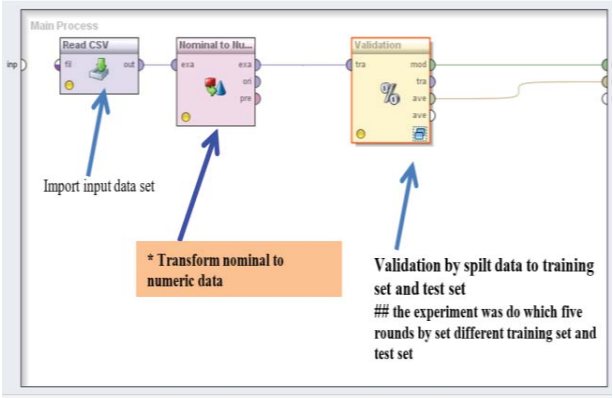

Figure 4. Import data and validation for ANN model

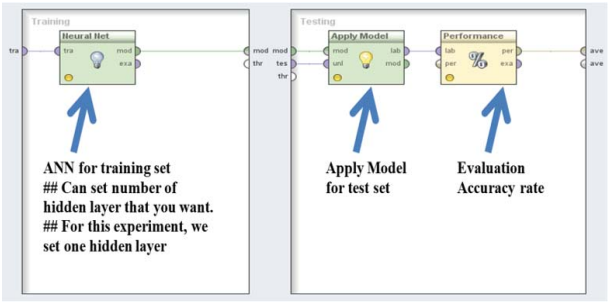

Figure 5. An ANN model

After running the ANN model, the results found that the split data set (training set 80% and test set 20%) has the highest accuracy rate (90.68%). Overall, the experiment showed that the ANN model provided a high accuracy rate (90.68%) than the Decision tree model (89.10%).

IV. EVALUATION

This research used the accuracy rate and execution time to compare the effectiveness of predicting target customers between decision tree algorithm and ANN.

"Accuracy" was defined as the overall success rate of the classifier and computed by using an equation as in (1)

$$\text{Acurracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

True Positive: TP, a number of correctly classified that an instance positive

False Positive: FP, a number of incorrectly classified that an instance is positive

False Negative: FN, a number of incorrectly classified that an instance is negative

True Negative: TN, a number of correctly classified that an instance is negative.

A comparison evaluation results by using an accuracy rate as shown in Table II. As seen in Table II, all data test proportions of ANN have a higher accuracy rate than the Decision Tree algorithm. The highest accuracy rate was ANN (90.68%) with data test proportion (80/20). While the decision tree model which provided the highest accurate rate was 89.10% with data test proportion (80/20).

TABLE II. COMPARE ACCURACY RATE DECISION TREE AND ANN

Algorithm	Test Proportion (%)				
	50/50	**60/40**	**70/30**	**80/20**	**90/10**
Decision Tree	88.81	88.87	88.96	***89.10***	88.82
ANN	90.00	90.21	90.36	***90.68***	90.23

A comparison evaluation results by using an execution time as shown in Table III. As seen in Table III, the decision tree algorithm has faster execution time than ANN. The average execution time of decision tree model was only used 4 second while ANN used in 6:38 minutes.

TABLE III. COMPARE COMPARE EXECUTION TIME

Algorithm	Average Execution Time
Decision Tree	**4 Second**
ANN	6:38 Minutes

As seen in Table II, the accuracy rate of ANN (80/20) has the highest accuracy rate from all of the test proportions of decision tree model but is not so far different (Decision tree 89.10%, ANN 90.68%). While the execution time of decision tree algorithm was faster than the ANN algorithm outstandingly, as shown in Table III

This research still has some issues which might impact the effectiveness results. For example, most of the output variable of this data set was class no (a client who does not subscribe bank deposit) and duration time of contact is very short duration. Therefore, most of the rule sets of the result were the rule sets for predicting class "no". Moreover, the prediction duration time of contact for the new customer only had one rule set which might not suitable for predicting a bank target customer.

For the bank client data set, ANN is might not suitable for these data set because most of the data set are categorical data set. It has to convert to numeric which might be made the meaning of the data set was changed. While the ruleset was generated from the decision tree, have suitable for using to deploy to make a rule set in developing a real program. The influence features affect the decision of the customer to subscribe bank deposit were age, duration, and pdays. The duration was the last contact duration from the bank telemarketing. The pdays was a number of days that passed by after the client was last contacted from a previous campaign.

The rule set was extracted from the decision tree classifier in Figure 4 showed that if customer spends a time to talk with the telemarketing in last contact more than 827.50 second and a number of days passed by the last contacted from a previous campaign less than 495 days, they have a chance to be a bank target client and buy a new bank campaign. A problem of the generated rule set was a new client whose sale person never contacts before. They do not have data of duration and pdays.

Therefore, this work intends to find a new methodology for predicting a new client by using a Fuzzy Set [7]. The step of using fuzzy set starts from convert two influenced attributes (duration and pdays) to ordinal data. The duration attribute has converted by using the duration time to ordinal data which included very short duration (duration time <=500), short duration (duration time <=1500), medium duration (duration time <=2500), long duration (duration time <= 3500), and very long duration (duration time >3500). The pdays attribute has converted by using the last contacted days to ordinal data which included short days (last contacted days <=30), medium days (last contacted days <=90), and long days (last contacted days <=), long duration (duration time <= 3500), and very long duration (duration time >3500).

The next step of fuzzy set was calculated the weight membership function. This work used the trapezoidal membership function to calculate the membership ship of the duration and pdays attributes [8] as in (2)

$$f(x;a,b,c,d) = \begin{cases} 0, & x < a \\ (x-a)/(b-a), & a \le x < b \\ 1, & b \le x < c \\ (d-x)/(d-c), & c \le x < d \\ 0, & d \le x \end{cases} \quad (2)$$

An example of calculation the membership function of evaluation the very short duration value as in (3).

$$f(x;0,0,100,500) =$$
$$if\ 0\ \le X\ \le 100\ then\ 1\ (true)$$
$$if\ 100\ < X\ < 500\ then\ (500-x)/(500-100)$$
$$if\ 500\ \le x\ then\ 0\ (false) \quad (3)$$

The last step of the fuzzy set was calculated the weight membership function. The original duration time and pdays values were converted to what ordinal value, which considered by the max value of the membership function.

Then, a new experiment was operated by decision tree algorithm again by using the new value of both duration and pdays attributs.

The results found that the accuracy rate of the decision tree algorithm which used converted data set to fuzzy term set was 88.38%. Although the accuracy rate of the converted data set was lower than using the original data set, the rule sets form the decision tree which used fuzzy term set provided more understandable and reasonable. The new rule sets from the decision tree which used the fuzzy term set described as below.

New Rule Set (decision tree which used fuzzy term set)

age > 89.500
| age > 93.500: no {no=2, yes=1}
| age ≤ 93.500: yes {no=0, yes=6}
age ≤ 89.500
| fduration = long duration
| | age > 29.500
| | | age > 35
| | | | age > 36.500: no {no=6, yes=6}
| | | | age ≤ 36.500: yes {no=0, yes=3}
| | | age ≤ 35: no {no=2, yes=0}
| | age ≤ 29.500: yes {no=0, yes=3}
| fduration = medium duration: yes {no=79, yes=125}
| fduration = short duration
| | age > 83.500: yes {no=0, yes=7}
| | age ≤ 83.500: no {no=3005, yes=2130}
| fduration = very long duration: no {no=2, yes=1}
| fduration = very short duration: no {no=36826, yes=3007}

From the new rule set, attributes were used to create a rule set includes only age and duration. Consequently, for new customers who have no duration attribute need to find a solution to fulfill this value. The next step was to run the decision tree model for predicting a duration value. The results identified that the duration value of a new customer should set to be a "very short duration" value.

The last step was transformed the class label of each customer from yes/no to 4 categories includes: very-high chance, high chance, medium chance and low chance by using probabilities (error rate) value from rule set. To calculate the probability value, we use formula as in (4)

$$probality\ of\ class\ yes\ of\ class\ yes\ = \frac{tp}{tp} + fp$$
$$probality\ of\ class\ yes\ of\ class\ no\ = 100 - \left(\frac{tn}{tn} + fn\right) \quad (4)$$

The detail of transformation from output variable (Yes /no) as show in Table IV.

TABLE IV. TRANFORM DETAILS FROM YES/NO TO 4 CATEGORIES

Probability to class yes	Result
0-40%	low chance
41-60%	middle chance
61-80%	high chance
81-100%	very high

V. DEVELOPMENT A DECISION SUPPORT SYSTEM

To development a decision support system for predicting bank target client, the rule sets were inputted to the web-based program. Output result of a DSS program, we propose to show a predict result of target customers who are taken into consideration for bank deposit subscription. Outputs have 4 categories includes: very-high chance, high chance, middle chance and low chance. The DSS for predicting bank target clients were developed by using PHP language and MYSOL server. The system has program abilities are include:

- Security Mode: Users have to login before using application
- Import file
  - Import old customer file by csv format for predict probability to repeat purchased in another bank deposit product.
  - Import new customer file by csv format for predict probability to purchase bank deposit product.
- Show Result of prediction
  - Search Customer.
  - Search by Customer name
- Search Data
- Summary Statistic
- Summary Report the number of target customer in each probability

The system interface

First page of this web application is Login Page. Users have to login before used this system, as show in Figure 6

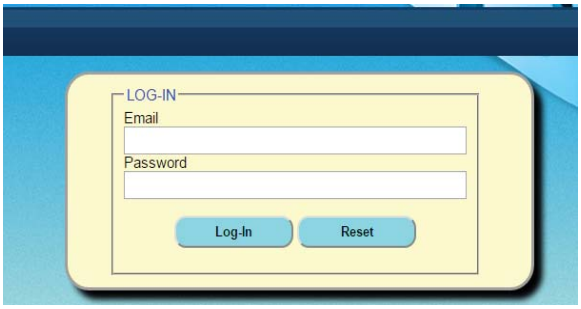

Figure 6. Login Page

If users input correct email and password, a homepage will show. The homepage of this system includes with six menus are include Home, Predict Old Customer, Predicting New Customer, Search by Name, Search by Chance to buy and Summary Statistic.

Figure 7. Home Page

Figure 8 show the Predicting Old Customer page. Users can import csv file for predicting result by click browse button then click upload button. Then the predicting result will show as Figure 9.

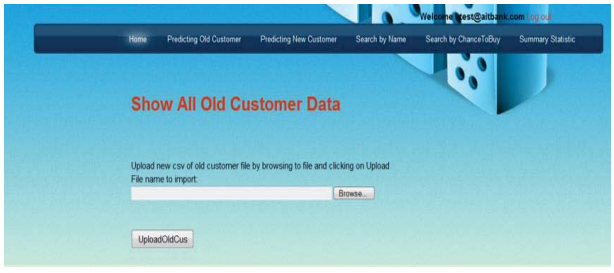

Figure 8. Predicting Old Customer Page

ID	FullName	Gender	Age	PhoneNumber	ChanceToBuy
1	Nicki Sergeant	Male	58	970819085	low chance
2	Epifania Whetsel	Male	44	970819086	low chance
3	Claudie Pendelton	Male	33	970819087	low chance
4	Vernie Bowley	Male	47	970819088	low chance
5	Thuy Critchfield	Male	33	970819089	low chance
6	Jeffry Satcher	Male	35	970819090	low chance
7	Annie Samons	Female	28	970819091	low chance

Figure 9. Predicting Results

The next page is search customer by a chance to buy. Users can search customer by selecting a chance that they want, as show in Figure 10.

ID	FullName	Gender	Age	PhoneNumber	ChanceToBuy
1	Kizzy Weidner	Female	53	970819122	high chance
2	Jeremy Osteen	Female	46	970819144	high chance
3	Marianne Cisco	Male	40	970819237	high chance
4	Isabelle Maynez	Female	39	970819531	high chance
5	Marie Cohnor	Female	34	970819559	high chance

Figure 10. Search Page

VI. Conclusion and Discussion

This work intends to find data mining techniques which have more effective with this data set by comparing between decision tree and ANN. The result showed that the ANN (80/20) has the highest accuracy rate (90.68%) more than the decision tree algorithm. The splitting data set was provided the highest rate both decision tree and ANN was training set 80% and test data set 20%.

In this case, if to bring the knowledge and results deploy to real use, there is no need to choose only the algorithm with the highest accuracy rate. It is reasonable to choose the data mining algorithm which provided a perform with speed time and also high accuracy rate for

achieving the goal set and superior to business competitors. Moreover, this work has been developed a real decision support system for predicting bank target clients by using the knowledge results from the prediction model. Therefore, it is required to find a rule set for predicting both an old contact client and a new client who will be buying a bank campaign. Thus, the fuzzy term set was used to apply to find a suitable rule set and more understandable and reasonable.

The DSS for predicting bank target clients has been developed cover with the basic functions. The telemarketing employee and bank sale managers can use this system to find customers who have a high chance to subscribe to a bank deposit. The system will help to reduce the operation cost to find a new customer.

In future research work, it is important to use more data set of bank clients which covered with a class yes (customers buy bank deposits) to improve the effectiveness of results such as data set fairly (provide suitable output class no and yes), accuracy rate and execution time.

REFERENCES

[1] Moro s., Cortez, P. and Rita, Paulo . Business intelligence in banking:A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation, Contents lists available at ScienceDirect : Expert Systems with Applications 42 (2015) 1314–1324.

[2] Sabbeh, S. F. (2018). Machine-Learning Techniques for Customer Retention: A Comparative Study. (IJACSA) International Journal of Advanced Computer Science and Applications, 9(2).

[3] Guo, J., & Hou, H. (2019, January). Statistical Decision Research of Long-Term Deposit Subscription in Banks Based on Decision Tree. In *2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)* (pp. 614-617). IEEE.

[4] Koumétio, C. S. T., Cherif, W., & Hassan, S. (2018, October). Optimizing the prediction of telemarketing target calls by a classification technique. In 2018 6th International Conference on Wireless Networks and Mobile Communications (WINCOM) (pp. 1-6). IEEE.

[5] UCI Machine Learning Repository. (2012) Bank Marketing Data Set [Data file]. Retrieved from http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

[6] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. CRISP-DM 1.0 - Step-by-step data mining guide, CRISP-DM Consortium, 2000

[7] Khan, N and Khan ,D. Fuzzy Based Decision making for promotional marketing campaigns, International Journal of Fuzzy Logic Systems (IJFLS) Vol.3, No1, January 2013

[8] Zadeh, L. A. (1965). Fuzzy sets. Information and control, 8(3), 338-353.