Naipunnya®
To reach the unreachable

Naipunnya Institute of Management & Information Technology

Affiliated to the University of Calicut, ISO 9001-2015 Certified
Accredited by NAAC with B++ grade

FACULTY ENHANCEMENT PROGRAMME

Naipunnya®
To reach the unreachable

Date of event: 15/06/2023

Faculty In-charge: Ms. Rinku K Vithayathil

NAIPUNNYA INSTITUTE OF MANAGEMENT AND INFORMATION TECHNOLOGY
Naipunnya
To reach the unreachable
Pongam, Koratty East
Thrissur District
Kerala-680 308

REPORT

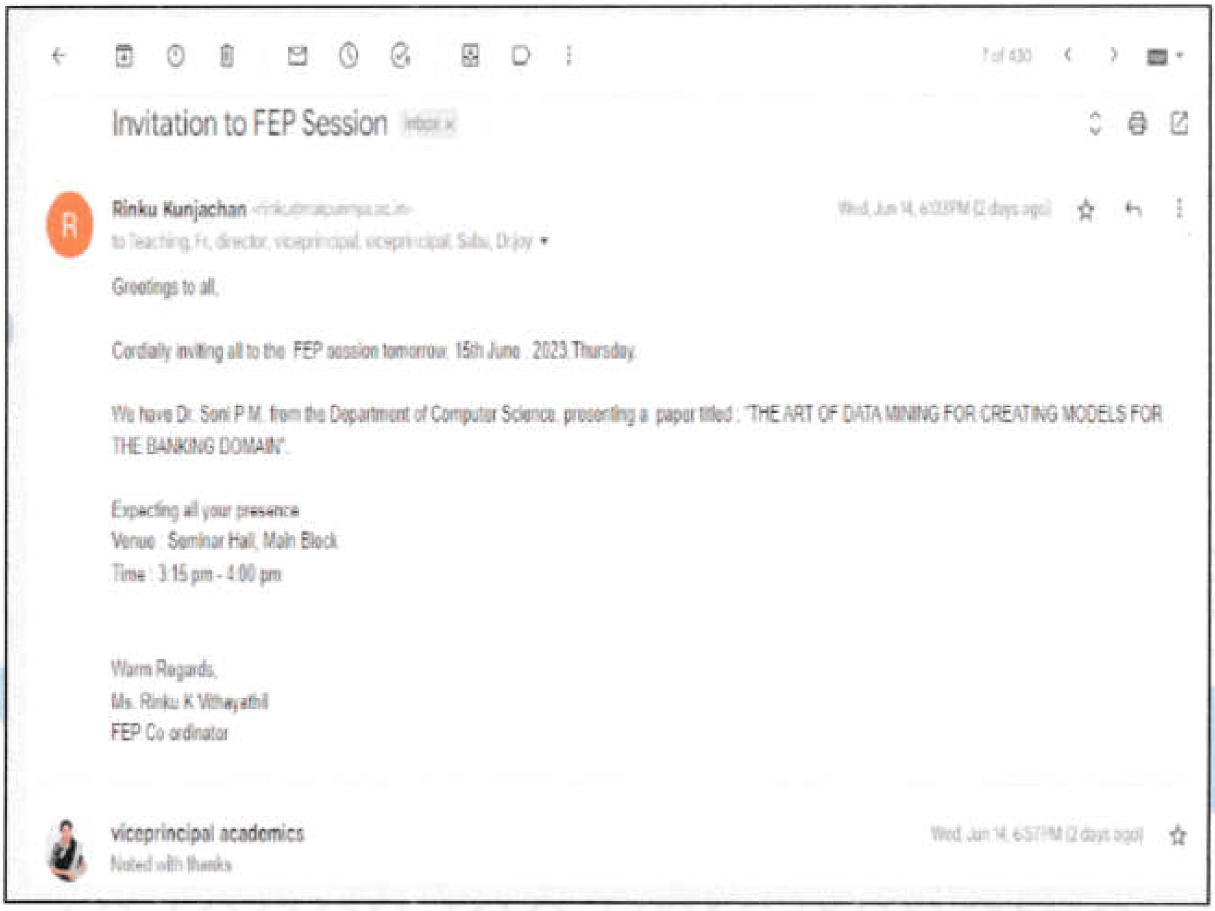
The Faculty Enhancement Program for the month of June was conducted on 15/06/2023, Thursday at 3:15 p.m. at Seminar hall, Main block. Dr. Soni P M of the Computer Science Department presented a paper titled "The Art of Data Mining for Creating Models for Banking Domain". 51 members of the faculty from various departments attended the program. Dr. Joy Joseph Puthussery, Dr. Sabu Varghese and other members of the faculty raised queries and made the session more interactive. The program concluded at 4:00 p.m. with a thanks note by Ms. Rinku K Vithayathil, FEP Coordinator.

Prepared by:	Verified by:	Approved by:
Ms. Rinku K Vithayathil	Dr.Sabu Varghese	Rev.Fr.Dr. Paulachan K J
(FEP Co-ordinator)	(Director, IT/HRD Cell)	(Principal)

Naipunnya
To reach the unreachable

NAIPUNNYA INSTITUTE OF MANAGEMENT AND INFORMATION TECHNOLOGY
Naipunnya
To reach the unreachable
Pongam, Koratty East
Thrissur District
Kerala-680 308

SCREENSHOT OF E-MAIL



Invitation to FEP Session

Rinku Kunjachan <rinku@naipunnya.ac.in>
to Teaching, Ft, director, viceprincipal, viceprincipal, Sabu, Drjoy
Wed, Jun 14, 6:03PM (2 days ago)

Greetings to all,

Cordially inviting all to the FEP session tomorrow, 15th June , 2023 Thursday.

We have Dr. Soni P M. from the Department of Computer Science, presenting a paper titled : "THE ART OF DATA MINING FOR CREATING MODELS FOR THE BANKING DOMAIN".

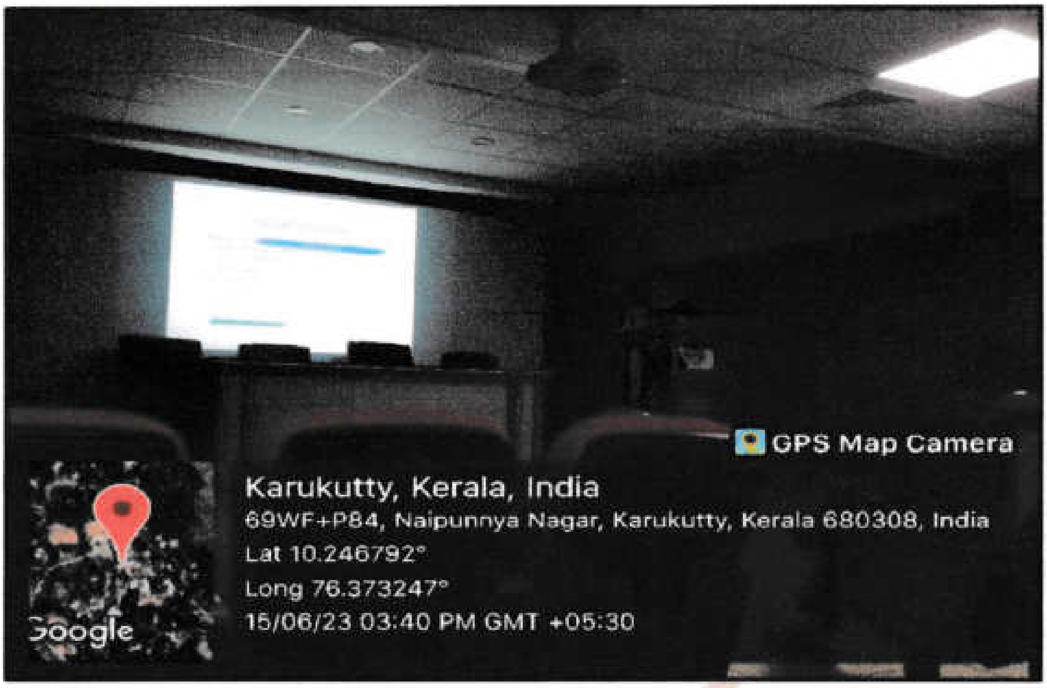
Expecting all your presence
Venue : Seminar Hall, Main Block
Time : 3.15 pm - 4.00 pm

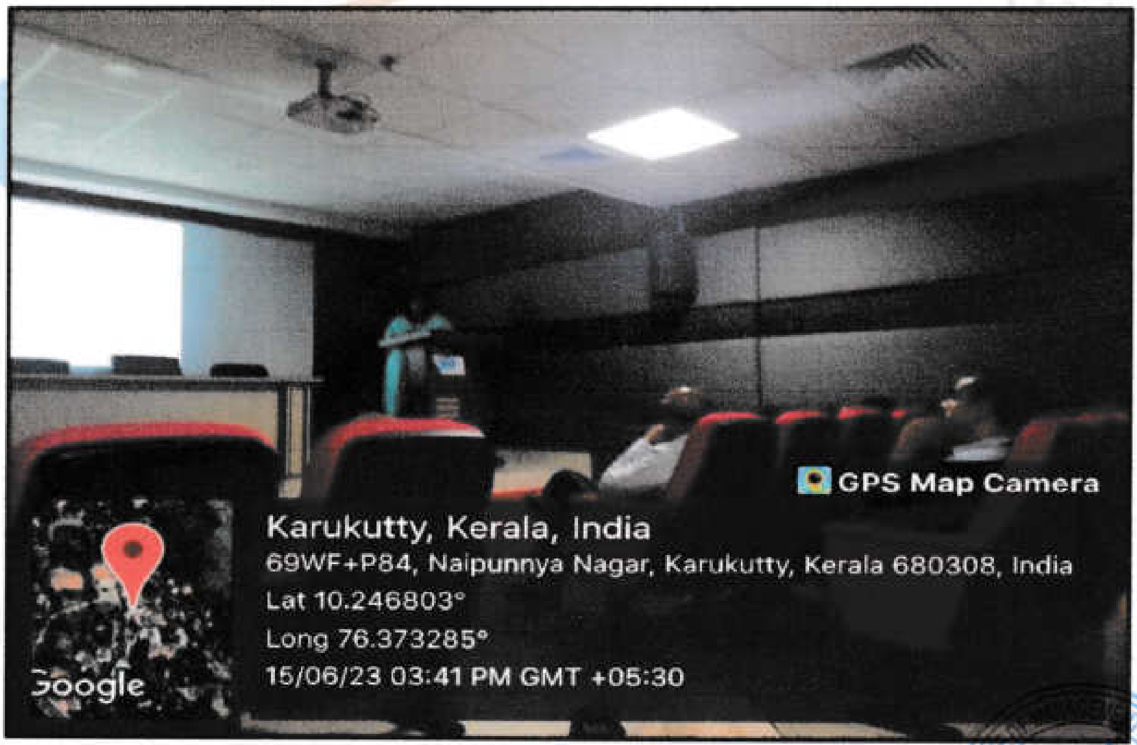
Warm Regards,
Ms. Rinku K Vithayathil
FEP Co ordinator

viceprincipal academics
Noted with thanks
Wed, Jun 14, 6:57PM (2 days ago)

NAIPUNNYA INSTITUTE OF MANAGEMENT AND INFORMATION TECHNOLOGY
Naipunnya
To reach the unreachable
Pongam, Koratty East
Thrissur District
Kerala-680 308

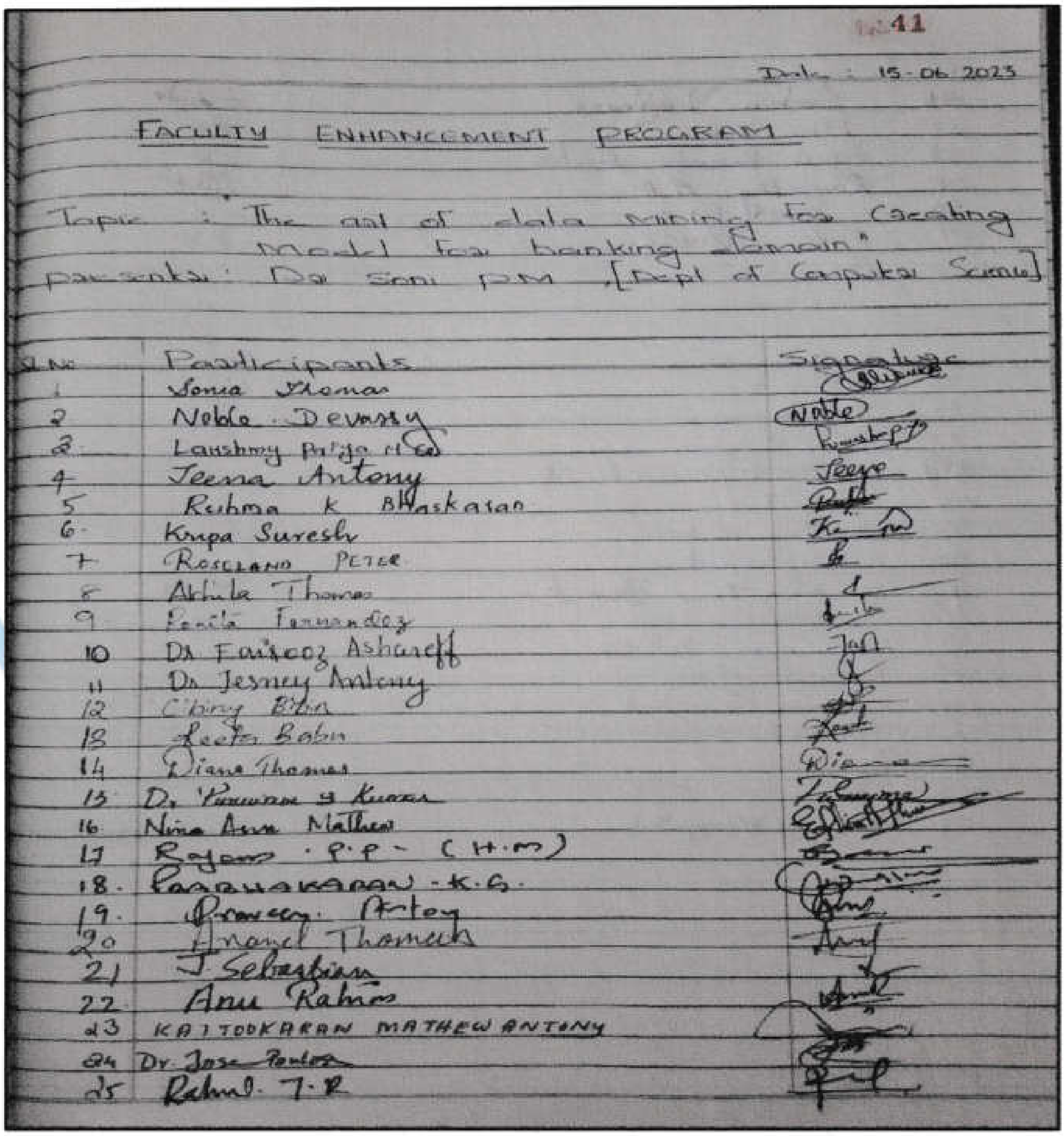
PHOTOGRAPHS


GPS Map Camera
Karukutty, Kerala, India
69WF+P84, Naipunnya Nagar, Karukutty, Kerala 680308, India
Lat 10.246792°
Long 76.373247°
15/06/23 03:40 PM GMT +05:30
Google


GPS Map Camera
Karukutty, Kerala, India
69WF+P84, Naipunnya Nagar, Karukutty, Kerala 680308, India
Lat 10.246803°
Long 76.373285°
15/06/23 03:41 PM GMT +05:30
Google

Naipunnya Institute of Management and Information Technology
Naipunnya
To reach the unreachable
Pongam, Koratty East
Thrissur District
Kerala-680 308

PARTICIPANT'S LIST



41

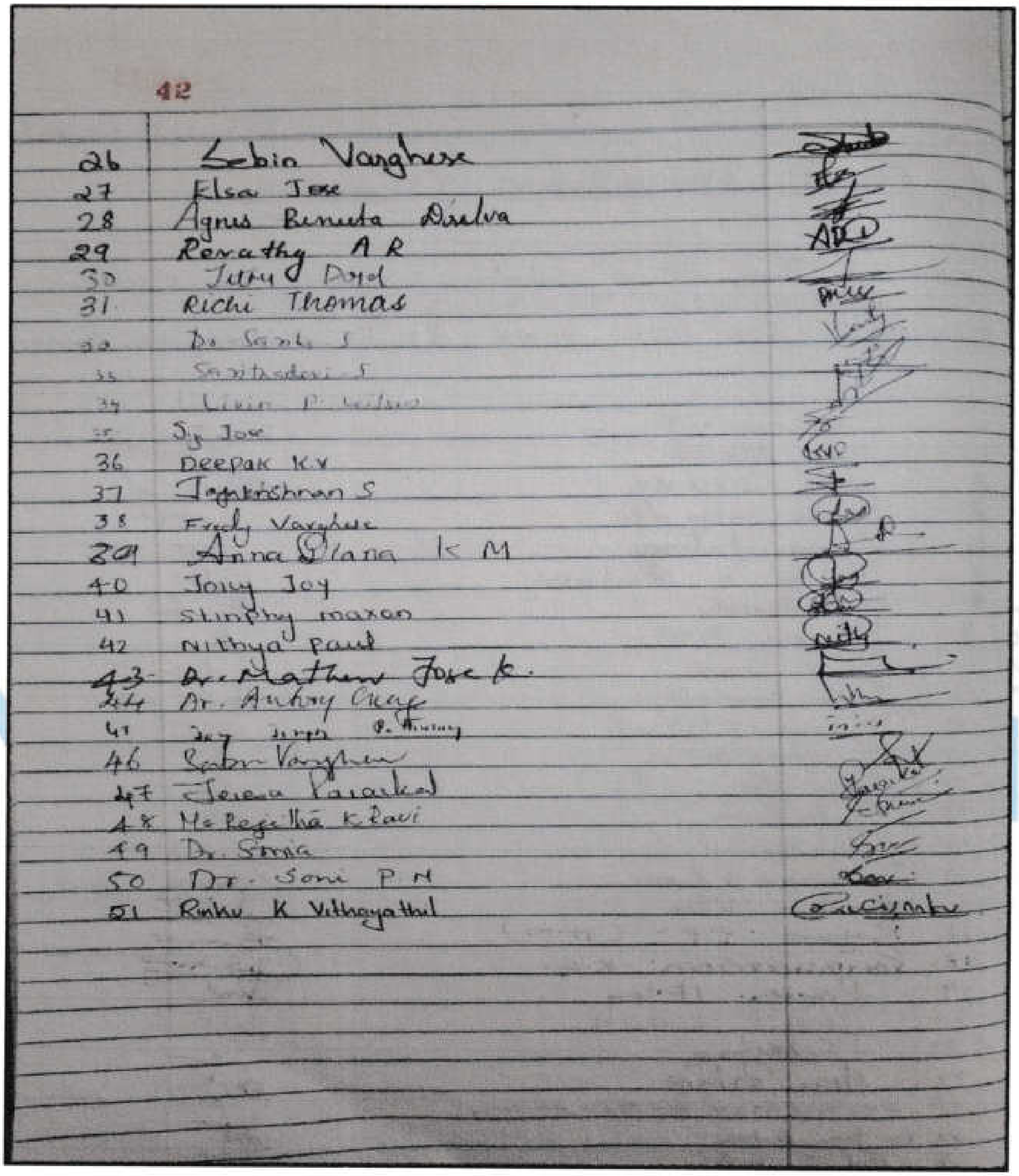
Date : 15-06-2023

FACULTY ENHANCEMENT PROGRAM

Topic : The art of data mining for Creating Model for banking domain"
presenter: Dr Soni P.M .[Dept of Computer Science]

Sl.No	Participants	Signature
1	Sonia Thomas	
2	Noble Devassy	Noble
3	Laushmy Priya M S	
4	Jeena Antony	Jeena
5	Reshma K Bhaskaran	
6.	Krupa Suresh	
7	ROSELAND PETER	
8	Akhila Thomas	
9	Fenita Fernandez	
10	Dr Fairooz Ashareff	
11	Dr Jesney Antony	
12	Cibiny Bibin	
13	Reeta Babu	
14	Diana Thomas	Dia
15	Dr Praveen S Kumar	
16	Nina Ann Mathew	
17	Rajan . P.P - (H.M)	
18.	PARBHAKARAN - K.G.	
19.	Praveen. Antony	
20	Anand Thomas	
21	J. Sebastian	
22	Anu Rahim	
23	KAITOOKARAN MATHEW ANTONY	
24	Dr. Jose Paulose	
25	Rahul. T.R	

NAIPUNNYA INSTITUTE OF MANAGEMENT AND INFORMATION TECHNOLOGY
Naipunnya
To reach the unreachable
Pongam, Koratty East
Thrissur District
Kerala-680 308



42

26	Sebin Varghese
27	Elsa Jose
28	Agnes Beneeta Dsilva
29	Revathy A R
30	Jerry Boyd
31	Richi Thomas
32	Dr Santhi S
33	Savithridevi S
34	Lijin P Wilson
35	Siju Jose
36	Deepak K.V
37	Jayakrishnan S
38	Fredy Varghese
39	Anna Diana K M
40	Jolly Joy
41	Shinphy maxon
42	Nithya Paul
43	Dr. Mathew Jose k.
44	Dr. Antony Cherry
45	[illegible]
46	Sebin Varghese
47	Teresa Parackal
48	Ms Regetha K Ravi
49	Dr. Sonia
50	Dr. Soni P M
51	Rinku K Vithayathil

Event coordinator

INSTITUTE OF MANAGEMENT AND INFORMATION TECHNOLOGY NAIPUNNYA
Naipunnya
To reach the unreachable
Pongam, Koratty East
Thrissur District
Kerala-680 308

THE ART OF DATA MINING FOR CREATING MODELS FOR THE BANKING DOMAIN

Soni P M [1]*Assistant Professor, Dept.of CS, NIMIT, Pongam ,Thrissur , Kerala, India*

Anna Diana K M [2]*Assistant Professor, Dept.of CS, NIMIT, Pongam ,Thrissur , Kerala, India*

Abstract

Customer Relationship Management" (CRM) is an important and commonly used Data mining application in the banking domain to interact with customers properly and effectively. As banking is considered as a service industry, the purpose of maintaining a strong and effective Customer Relationship Management is a critical issue [1]. Models created by Data mining algorithms can play a significant role in Customer Relationship Management in the banking domain especially for predicting the loan credibility behavior of a customer. Before creating the model, the quality of data being improved by data preprocessing techniques especially feature selection. The classification technique aims to predict accurately the target class such as, whether to approve or reject the loan for each case in the data. The way of Data mining process on the credit data set is demonstrated here. Different Data mining classification models are generated and evaluated to identify the loan applicants as eligible or illegible for a loan.

Keywords: *Data preprocessing, Classification, CRM, Random Forest, Feature Selection*

1. INTRODUCTION

The different applications of Data mining that can be used in the banking sector are Customer segmentation, banking profitability, Credit scoring and approval, predicting payment from Customers, Marketing, detecting fraud transactions, Cash management and forecasting operations, optimizing stock portfolios, and Ranking investments [2]. The primary goal of a bank is to lend the money generated by it from various sources. The lending of money to customers is very easy but its recovery is a hard process. Therefore, the primary objective of the banks as lenders is to ensure the profitability of the loans and advances sanctioned by them. In order to maintain CRM, grant loans to the reliable customers who can repay it from reasonably reliable sources within a stipulated time. Banks hold huge volumes of customer transaction data on daily basis. Data mining tools help to analyze these data and to convert into knowledge that can be used for the prediction of loan credibility behavior of a customer. CRM can be

Naipunnya
To reach the unreachable

NAIPUNNYA INSTITUTE OF MANAGEMENT AND INFORMATION TECHNOLOGY
Naipunnya
To reach the unreachable
Pongam, Koratty East
Thrissur District
Kerala-680 308

maintained within the banking industry by predicting the loan credibility behavior of a customer. Data mining analysis, huge data collected from the banking transactions and finally summarize it into meaningful knowledge. This knowledge helps the bankers for the proper decision making process and it leads to the smooth functioning of the organization. After the formulation of the problem statement, collect the relevant data and apply some preprocessing techniques to transform original data into a suitable form that can be applied for mining process. Finally, apply data mining functionality especially classification to categories the customer into two groups such as, those who can pay the loan amount promptly or not.

2. DATA COLLECTION

The data were collected from a UCI depository. The attributes are listed in Table -1

SI	Name of attribute
1	Checking Status
2	Duration
3	Credit History
4	Purpose
5	Credit Amount
6	Savings Status
7	Employment
8	Installment Commitment
9	Personal_Status
10	Other Parties
11	Residence Since
12	Property Magnitude
13	Age
14	Other_Payment_Plans
15	Housing
16	Existing Credits
17	Job
18	Num_Dependents

Naipunnya® To reach the unreachable

NAIPUNNYA INSTITUTE OF MANAGEMENT AND INFORMATION TECHNOLOGY
Naipunnya To reach the unreachable
Pongam, Koratty East
Thrissur District
Kerala-680 308

19	Own Telephone
20	Foreign Worker
21	Class

Table 1 – List of Attributes

3. DATA PRE PROCESSING

The customer transactions data collected from the banking domain may contain duplicate values, missing values, noise or inconsistency. This affects the reliability of mining process. If the user believes that the data are dirty, they will not trust the results of the data mining process that has been applied to this data [5]. A data mining process with high quality of data will produce accurate data mining results. To improve the quality of data and consequently the mining results, the data preprocessing has to be done on the collected data. Data preprocessing is one of the critical step in data mining process which deals with preparation and transformation from the initial data set to the final data set [4]. The following categories of data preprocessing are applied to convert initial data set to final data set.

- Data cleaning
- Data integration
- Data transformation
- Data reduction

Naipunnya
To reach the unreachable

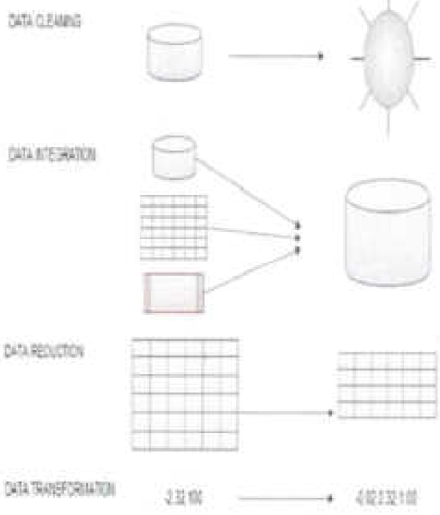


Figure - 1: Data Pre-processing methods

NAIPUNNYA INSTITUTE OF MANAGEMENT AND INFORMATION TECHNOLOGY
Naipunnya
To reach the unreachable
Pongam, Koratty East
Thrissur District
Kerala 680 308

In order to apply any of the data pre-processing technique, the data should be in proper format. Therefore, conversion of data obtained into a suitable form before applying the pre-processing steps is mandatory. The excel format has to be converted into respective formats (.csv,.arff) required for the processing of various data mining models. Data pre-processing of banking data start from processing duplicate and missing values. The missing values are substituted by another computed value such as mean median or mode. For example, all the missed "Jobs" in the dataset are replaced with the term "Business" as it is the most occurred job. Label Encoding is a method for data transformation. It will convert labels in the credit data set into the numeric form so that it can be easily transformed into a machine-readable form. The complexity of the dataset can be reduced by applying methods of data reduction. Data reduction is a technique to reduce the volume of initial data set and should produce the same performance [5].

Feature selection is an important reduction method. Feature selection has proven in both theory and practice to be effective in enhancing learning efficiency, increasing predictive accuracy and reducing the complexity of learned results. [6,7]. Table 2 represents the correctly classified accuracy and incorrectly classified accuracy obtained by each of the feature selection algorithms applied on the credit data set.

Naipunnya
To reach the unreachable

Feature Selection	**Time**	**Correctly classified**	**Incorrectly classified**
Chisquared	0.03	78.4	21.6
Filtered	0.06	74.7	25.3
InfoGain	0.03	74.7	25.3
OneR	0.03	96.1	3.9
Relief	0.06	60.4	39.6

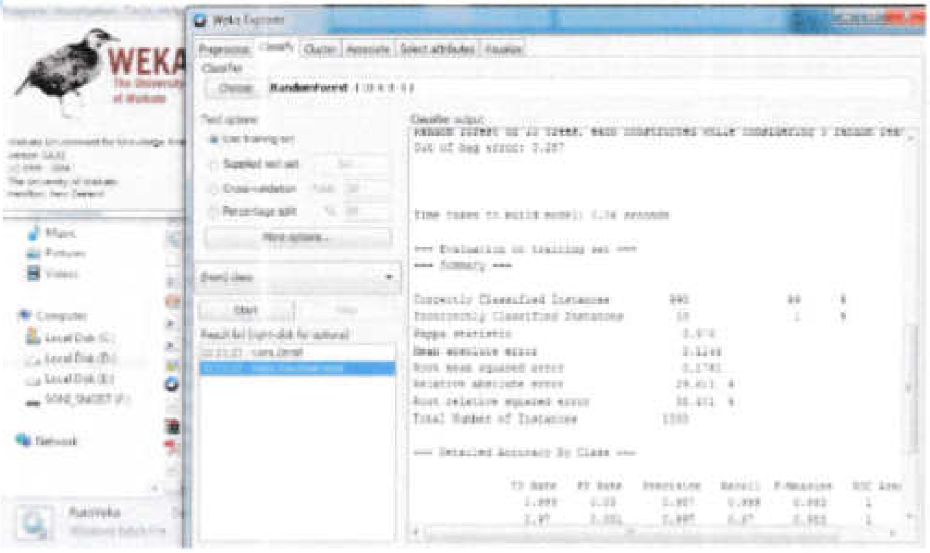
Table- 2: Performance metric of feature selection algorithms

NAIPUNNYA INSTITUTE OF MANAGEMENT AND INFORMATION TECHNOLOGY ★
Naipunnya
To reach the unreachable
Pongam, Koratty East
Thrissur District
Kerala 680 308

4. CLASSIFICATON

Classification is used in scenarios where we need to identify the category or class into which a new observation might belong. Classification is one of the data analysis methods that predict class labels [7]. There are more classification methods such as Statistical-based, Distance-based, Decision tree-based, Neural network-based, Rule-based [8]. Choosing the correct classification method, thus, becomes very important for obtaining accurate results. Random Forest is now known to be one of the most efficient classification methods [9]. In order to classify a customer as "eligible customer" or "not eligible customer" using credit dataset binary classification method is used. The process of classification divides the dataset into two parts, one for creating the model called training dataset and the other for testing the model called testing dataset.

The various classification algorithms are applied on the credit data set are JRip, ZeroR, SMO, Adaboost, Random Forest, Kstar, Ridor, and DTNB. Accuracy is a measurement to evaluate the efficiency of each classifier. The other mode of evaluating performance is, Kappa Statistic, and, Mean Absolute Error. These metrics are used to compare and evaluate which classification algorithm is better for the loan credibility prediction. The classification performance based on the above measures is described in the table and the process of classification is demonstrated in the Figure 2. From the Table 3 it is clear that the Random Forest classification algorithm produced better accuracy on the credit data set.



```
Out of bag error: [illegible]

Time taken to build model: [illegible] seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        [illegible]        [illegible] %
Incorrectly Classified Instances      [illegible]        [illegible] %
Kappa statistic                        [illegible]
Mean absolute error                    [illegible]
Root mean squared error                [illegible]
Relative absolute error               [illegible] %
Root relative squared error           [illegible] %
Total Number of Instances            [illegible]

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
                [illegible]     [illegible]       [illegible]     [illegible]     [illegible]       1
                [illegible]     [illegible]       [illegible]     [illegible]     [illegible]       1
```

Figure 2: Classification using Weka

NAIPUNNYA INSTITUTE OF MANAGEMENT AND INFORMATION TECHNOLOGY
Naipunnya
To reach the unreachable
Pongam, Koratty East
Thrissur District
Kerala-680 308

Classifiers	Accuracy (%)	Kappa	MAE
JRip	74.3	0.346	0.366
ZeroR	70	0	0.42
SMO	78.4	0.45	0.216
Adaboost	73.7	0.225	0.342
Random Forest	99	0.976	0.124
Ridor	76	0.2701	0.24
DTNB	71.1	0.394	0.362

Table 3: Classification Performance

The figure 3 represents the classification accuracy, figure 4 represents classification Kappa metric and figure 5 represents Classification MAE metric. From the above graphs, it is clear that Random Forest algorithm can perform better for classifying the customer as "eligible customer "or "not eligible customer" for issuing the loan.

Naipunnya
To reach the unreachable

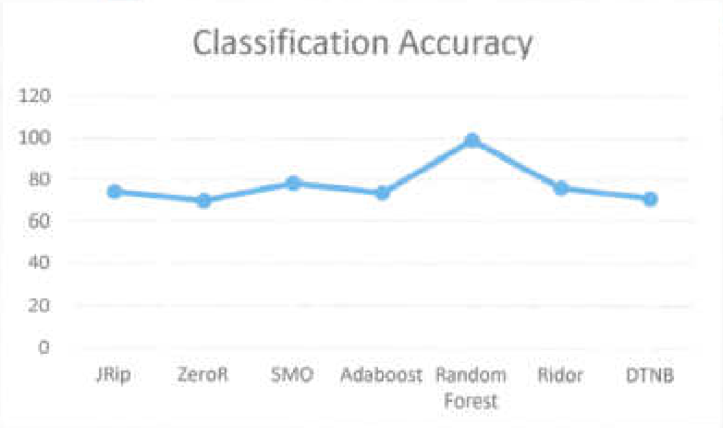


Figure 3: Classification Accuracy

NAIPUNNYA INSTITUTE OF MANAGEMENT AND INFORMATION TECHNOLOGY
Naipunnya
To reach the unreachable
Pongam, Koratty East
Thrissur District
Kerala 680 308

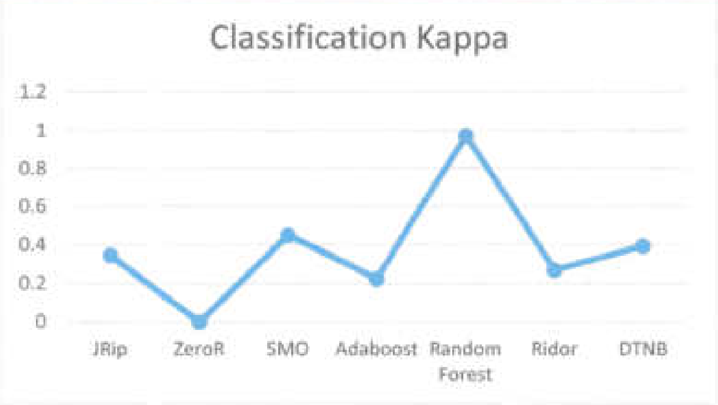


Figure 4: Classification Kappa metric

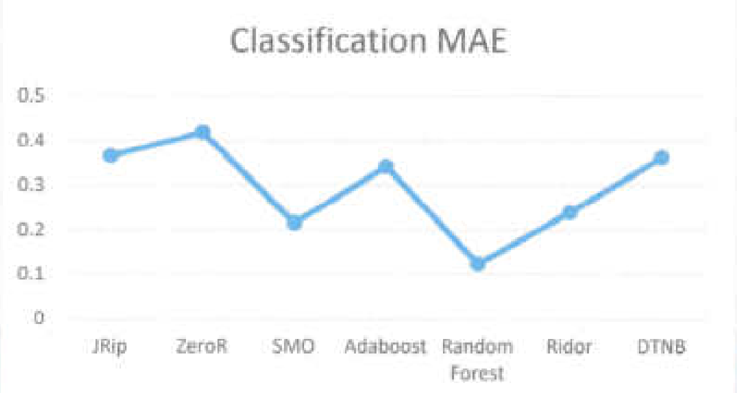


Figure 5: Classification MAE metric

Naipunnya
To reach the unreachable

5. TECHNOLOGY USED

Weka is a powerful tool that contains collection of machine learning algorithms for data mining tasks such as data preparation, classification, regression, clustering, association rules mining, and visualization. The figure 5 represents the various operations that can be performed in Weka.

NAIPUNNYA INSTITUTE OF MANAGEMENT AND INFORMATION TECHNOLOGY
Naipunnya
To reach the unreachable
Pongam, Koratty East
Thrissur District
Kerala-680 308

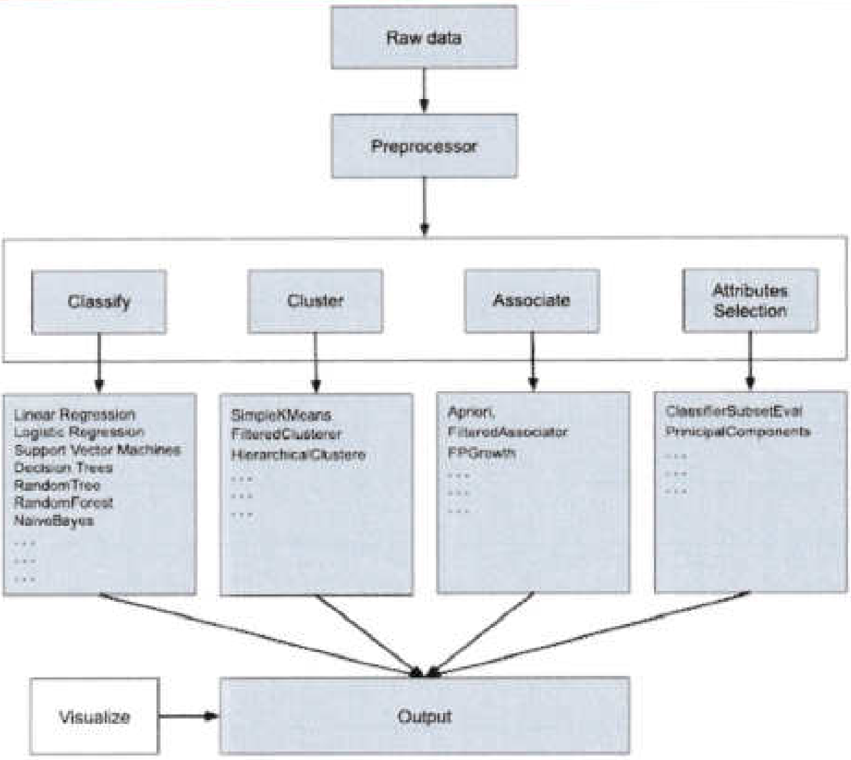


Figure **5:** Operations of Weka

Naipunnya
To reach the unreachable

The process start with the raw data set and apply preprocessing tools to clean the raw data set into preprocessed dataset that can be used for mining operations. The data mining algorithms were applied on this preprocessed dataset. The different data mining operations are **Classify, Cluster**, or **Associate.** The **Attributes Selection** allows the automatic selection of features to create a reduced dataset. Then, WEKA prepared statistical output of the model processing using visualization tools. The various models can be applied on the same dataset. Thus, WEKA results in a fast development of data mining models on the whole.

6. FUTURE SCOPE

The main focus of my work is the design of data mining models to predict the customers who repay loan amount promptly from those who do not. From the experiment it is clear that the classification algorithm Random Forest produced better accuracy after applying the feature selection methods. In future, the model

NAIPUNNYA INSTITUTE OF MANAGEMENT AND INFORMATION TECHNOLOGY
Naipunnya
To reach the unreachable
Pongam, Koratty East
Thrissur District
Kerala-680 308

creation can be extended to apply for different binary classification problems and these can also be applied to handle large amount of data using some big data technologies.

REFERENCES

[1] Frawley, W. J., Piatetsky-Shapiro, G., and Matheus, C. J. (1992). Knowledge discovery in databases: An overview. *AI Magazine*, 13(3):57.

[2] Dileep B. Desai, Dr. R.V.Kulkarni "A Review: Application of Data Mining Tools in CRM for Selected Banks", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (2) , 2013, 199 – 201

[3] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, "Data Preprocessing for Supervised Leaning", *International Journal of Computer Science*, 2006, Vol 1 N. 2, pp 111–117.

[4] Soni P M, Varghese Paul, Sudheep Elayidom "An Efficient Data Preprocessing Frame Work for Loan Credibility Prediction System "IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308

[5] Ms. Neethu Baby1, Mrs. Priyanka L.T "Customer Classification And Prediction Based On Data Mining Technique", International Journal of Emerging Technology and Advanced Engineering, (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 2, Issue 12, December 2012)

[6] Almuallim and T. G. Dietterich. "Learning boolean concepts in the presence of many irrelevant features," Artificial Intelligence, vol. 69, no. 1-2, pp. 279–305, 1994

[7] Sudhamathy G (2016), "Credit Risk Analysis and Prediction Modelling of Bank Loans Using R", Vol 8 No 5.

[8] K. Chitra, B.Subashini (2013), "An Efficient Algorithm for Detecting Credit Card Frauds", Proceedings of State Level Seminar on Emerging Trends in Banking Industry

[9] Bernard, S., Heutte, L., Adam, S. (2009), "On the selection of decision trees in Random Forests", International Joint Conference on Neural Network , pp. 302–307.

Naipunnya
To reach the unreachable

NAIPUNNYA INSTITUTE OF MANAGEMENT AND INFORMATION TECHNOLOGY
Naipunnya
To reach the unreachable
Pongam, Koratty East
Thrissur District
Kerala-680 308