CHAPTER-1

INTRODUCTION

Data mining techniques are becoming very widespread domain nowadays because of the extensive accessibility of huge quantity of data and the need for transforming such data into knowledge. In the current financial sector, core banking model and cut throat competition making banks to struggling to gain a competitive edge over each other. The face to face interaction with customer is no more exists in the modern banking world. Banking systems collect huge amounts of data on day to day basis, be it customer information, transaction details like deposits and withdrawals, loans, risk profiles, credit card details, credit limit and collateral details related information. Thousands of decisions are taken in a bank on daily basis. In recent years the ability to generate, capture and store data has increased enormously. The information contained in this data can be very important. The wide availability of huge amounts of data and the need for transforming such data into knowledge encourage IT industry to use data mining. Lending is the primary business of the banks. Trust Risk Management is one of the most important and critical factor in banking world. Without proper credit risk management banks will face huge losses and lending becomes very tough for the banks. Data mining techniques are greatly used in the banking industry which helps them compete in the market and provide the right product to the right customer with less risk. Credit risks which account for the risk of loss and loan defaults are the major source of risk encountered by banking industry. Data mining techniques like classification and prediction can be applied to overcome this to a great extent. In this paper we introduce an effective prediction model for the bankers that help them predict the credible customers who have applied for loan. Decision Tree Induction Data Mining Algorithm is applied to predict the attributes relevant for credibility. A prototype of the model is described in this thesis which can be used by the organizations in making the right decision to approve or

customer types. The following chapters discuses about the introduction of data mining and other details.

1.1 DATA MININGINTRODUCTION

mining or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information. The information that can be used to increase revenue, cuts cost, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Although data mining is a relatively a new term, the technology is not new. Companies have used powerful computers to sift through volumes of supermarket scanner data and Data analyze market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.

Data: Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- Operational or transactional data such as, sales, cost, inventory, payroll, and accounting.
- Nonoperational data, such as industry sales, forecast data, and macro economic data
- Meta data - data about the data itself, such as logical database design or data dictionary definitions

Information: The patterns, associations, or relationships among all this *data* can provide *information*. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

Knowledge: Information can be converted into *knowledge* about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data. With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

II. Hierarchical Agglomerative methods

The hierarchical agglomerative clustering methods are most commonly used. The construction of an hierarchical agglomerative classification can be achieved by the following general algorithm.

1. Find the 2 closest objects and merge them into a cluster
2. Find and merge the next two closest points, where a point is either an individual object or a cluster of objects.
3. If more than one cluster remains , return to step 2

Individual methods are characterized by the definition used for identification of the closest pair of points, and by the means used to describe the new cluster when two clusters are merged. There are some general approaches to implementation of this algorithm, these being stored matrix and stored data, are discussed below:

- In the second matrix approach, an N*N matrix containing all pairwise distance values is first created, and updated as new clusters are formed. This approach has at least an O(n*n) time requirement, rising to $O(n^3)$ if a simple serial scan of dissimilarity matrix is used to identify the points which need to be fused in each agglomeration, a serious limitation for large N.

- The stored data approach required the recalculation of pairwise dissimilarity values for each of the N-1 agglomerations, and the O (N) space requirement is therefore achieved at the expense of an $O(N^3)$ time requirement.

1.2.2. Use of Clustering in Data Mining

Clustering is often one of the first steps in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships. This technique supports the development of population segmentation models, such as demographic-based customer segmentation. Additional analyses using standard analytical and other data mining techniques can determine the characteristics of these segments with respect to some desired outcome. For example, the buying habits of multiple population segments might be compared to determine which segments to target for a new sales campaign. For example, a company that sale a variety of products may need to know about the sale of all of their products in order to check that what product is giving a wide scope and which is lacking. This is done by data mining techniques. But if the system clusters the products that are giving lesssale then only the cluster of such products would have to be checked rather than comparing the sales value of all the products. This is actually to facilitate the mining process.

1.3 CLASSIFICATION

Classification is a data mining (machine learning) technique used to predict group membership for data instances. Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high Trust risks. Classification models are tested by comparing the predicted values to known target values in a set of test data. The historical data for a classification project is typically divided into two data sets: one for building the model; the other for testing the model. Classification algorithms in data mining and machine learning are given a set of inputs and they come up with a specific class associated with those inputs.

Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. Classifier performance depends greatly on the characteristics of the data to be classified. There is no single classifier that works best on all given problems (a phenomenon that may be explained by the no-free-lunch theorem). Various empirical tests have been performed to compare classifier performance and to find the characteristics of data that determine classifier performance. Determining a suitable classifier for a given problem is however still more an art than a science.

1.3.1 Decision tree induction

Decision tree induction is the learning of decision trees from class-labeled training Tuples. s. Decision tree induction is the learning of decision trees from class-labeled training tuples. The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy. However, successful use may depend on the data at hand. Decision tree induction algorithms have been used for classification in many application areas, such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology. Decision trees are the basis of several commercial rule induction systems. A decision tree is a flowchart-like tree structure, where each internal node (nonleaf 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 topmost node in a tree is the root node. “How is decision trees used for classification?” Given a tuple, X, for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree.

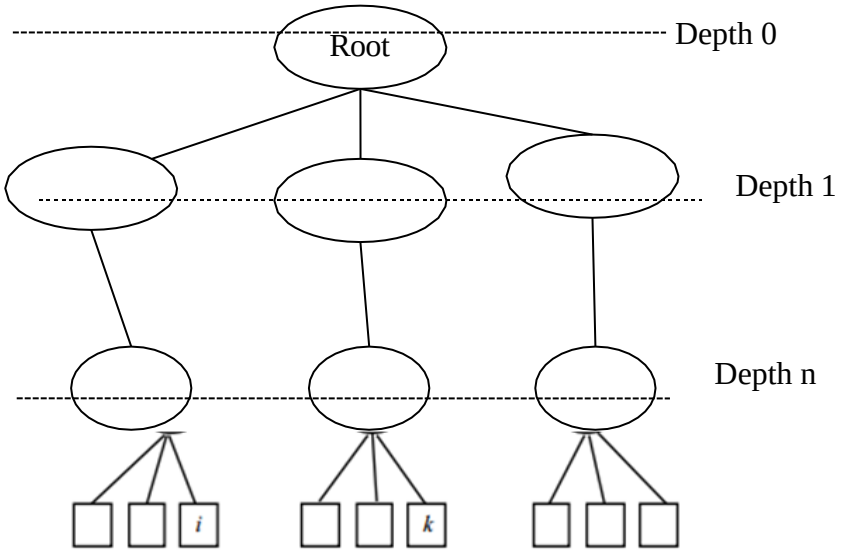


Fig1.3Decision Tree Model

Decision tree induction is used to minimize recommendation errors by making recommendation only for customers who are likely to buy recommended products. Making recommendation only for customers who are likely to buy recommended products could be a solution to avoid the false positives of the poor recommendation. This phase performs the tasks of selecting such customers based on the decision tree induction. The decision tree induction uses both the model set and the score set generated from customer records.

The majority of the oblique and univariate decision tree induction algorithms perform a top-down strategy for growing the tree, relying on an impurity-based measure for splitting nodes. Decision trees are used for classification and regression tasks. The training set used for inducing the tree must be labeled. However, acquiring a labeled data set is a costly task. Therefore, we believe that using a decision model which requires examples of one class only is highly preferable. It performs at least as accurate as the well-known decision tree data-linkage model, while incorporating the advantages of a one-class solution. A decision tree was induced using these five attributes. An individuals name, gender, date of birth, location, and the relationships between the individuals. Decision tree to determine which records should be matched to one another.

CHAPTER 2

LITERATURE REVIEW

2.1 Ralambondrainy, Henri. "A conceptual version of the K-means algorithm."*Pattern Recognition Letters* 16.11 (1995): 1147-1157.

H. Ralambondrainy [1995] proposed a hybrid numeric symbolic method that integrates an extended version of the K-Means algorithm for cluster determination and a complementary conceptual characterization algorithm for cluster description.

2.2 Calis, Asli, Ahmet Boyaci, and Kasim Baynal. "Data mining application in banking sector with clustering and classification methods." *Industrial Engineering and Operations Management (IEOM), 2015 International Conference on*. IEEE, 2015.

Calis, Asli, Ahmet Boyaci (2009), performed a study in form of application of data mining, in analysis of unstructured data. As a result of converting the unstructured data by using the methods of text and web mining and their contribution to the access of the model after inclusion and being converted into structured from, were analyzed. Models built by using C5.0 algorithm which is one of the decision tree methods were compared with each other and the best model is determined.

2.3 Ngai, Eric WT, Li Xiu, and Dorothy CK Chau. "Application of data mining techniques in customer relationship management: A literature review and classification." *Expert systems with applications* 36.2 (2009): 2592-2602.

Ngai, Eric WT, Li Xiu, by benefiting from data base containing personalized sales behavior according to the customers of a retail enterprise, aimed at, making a sales analysis containing, a detailed and relative measurement results. The classification type formed, benefited from CART decision tree technique for the sales forecasting model. At the end of CART decision tree technique application, k the customers were divided into classes according to their amount of spending. By doing so, it was possible to determine he target voids formed in scale success and to determine if at what degree the relative contributions of different factors were in this

2.4 Huang, Zhexue, and Michael K. Ng. "A note on k-modes clustering." *Journal of Classification* 20.2 (2003): 257-261.

Huang and Ng [1999] describes extensions to the fuzzy k-means algorithm for clustering categorical data. By using a simple matching dissimilarity measure for categorical objects and modes instead of means for clusters, a new approach is developed, which allows the use of the k-means paradigm to efficiently cluster large categorical data sets. A fuzzy k-modes algorithm is presented and the effectiveness of the algorithm is demonstrated with experimental results.

2.5 Morzy, Tadeusz, Marek Wojciechowski, and Maciej Zakrzewicz. "Scalable hierarchical clustering method for sequences of categorical values."*Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, 2001. 282-293.

Morzy et. al [2001] introduces a problem of clustering categorical data sequences and present an efficient scalable algorithm to solve the problem. This algorithm implements the general idea of agglomerative hierarchical clustering and uses frequently occurring subsequences as features describing data sequences. The algorithm not only discovers a set of high quality clusters containing similar data sequences but also provides descriptions of the discovered clusters.

2.6 Kanungo, Tapas, et al. "An efficient k-means clustering algorithm: Analysis and implementation." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 24.7 (2002): 881-892.

Kanungo et. al [2002] present a simple and efficient implementation of Lloyd's k-means clustering algorithm, which they call the filtering algorithm. This algorithm is easy to implement, requiring a kd-tree as the only major data structure. They establish the practical efficiency of the filtering algorithm in two ways. First, they present a data-sensitive analysis of the algorithm's running time, which shows that the algorithm runs faster as the separation between clusters increases. Second, they present a number of empirical studies both on synthetically generated data and on real data sets from applications in color quantization, data compression, and image segmentation.

CHAPTER 3
SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

Many researches proposed classification and prediction methods based on data mining in the field of financial and banking sector. This section presents about the existing techniques which are used in financial risk management and the loan repayment ability finding.

In existing K-means clustering methods are used.

- Using the k-means clustering the customers are classified into the predefined number of classes.
- After the classification process the rules were formed for identifying potential customers. This has been performed using decision tree algorithms.
- Based on the splitting criteria, customer classification is performed. But the existing decision trees and classification algorithms need more iterations.
- But in the proposed system, the optimal suggestions and schedules are identified for the repayment process.

Data Mining can be used to derive credit behavior of individual borrowers with parameters card loans, mortgage value, repayment and using characteristics such as history of credit, employment period and length of residency. A score is thus produced that allow a lender to evaluate the customer and decide whether the person is a good candidate for a loan, or if there is a tendency to become high risk of default. Customers who have been with bank for a longer periods of time, remained better with bank and have good credit history and have higher salaries/wages, are more likely to receive a loan than a new customer who has no credit history with the bank, or who earns low salaries/wages. Bank can reduce the risk factors to maintain a better position by knowing the chances of a customer to become default.

Disadvantages

- Identifying every users behavior using predefined test data is always not accurate.
- The existing system suffers in providing accurate decisions and suggestions to the customers.
- Failed to perform the effective decision over dynamic banking dataset.

A number of studies have been carried out on the application of data mining techniques for bank data sets. The Data Mining techniques applied on Bank data include k-means, bi clustering, k nearest neighbor, Neural Networks (NN) Support Vector Machine (SVM), Naive Bayes Classifier and Fuzzy c-means. As can be seen the appropriateness of data mining techniques is to a certain extent determined by the different types of bank data or the problems being addressed

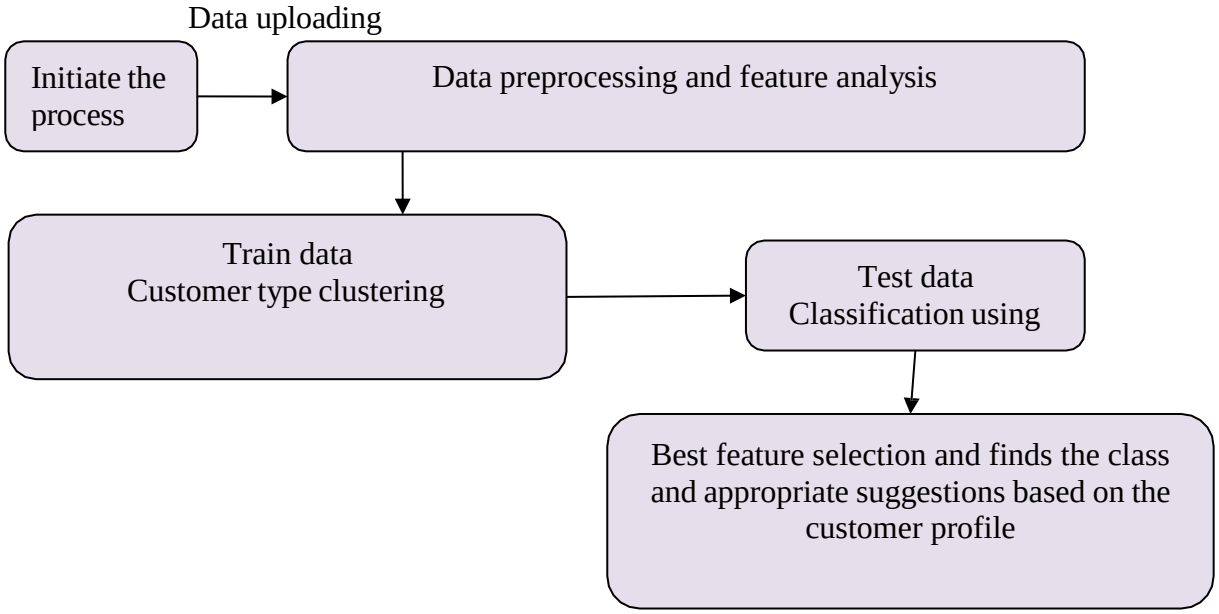
3.2 PROPOSED SYSTEM

The proposed work employed data mining techniques and data analysis systems with decision support techniques to discover new knowledge from banking data. In order to reduce the un-payment problems and credit risks in the banking domain, the proposed system introduces a set of data mining techniques to find an optimal decision on banking to increase repayment. The proposed system uses two types of implementations, which finds the credit risk and classifies the bank customers into different categories. Selecting features for the classification and decision support with improved clustering and classification techniques. This chapter describes the tools, techniques and the algorithms are used in the proposed system.

Analyzing the credit risk and finding best solution is the ultimate aim of this thesis. To perform the above, extended K-means algorithm and Enhanced Decision Tree (Ek-EDT) algorithms has been proposed. Ek-EDT involves incrementally building an ensemble by training each new model instance to emphasize the training instances that previous models wrongly classified and misplaced. The proposed Ek-EDT often leads to a dramatic improvement in predictive accuracy by providing an effective feature selection, when a data is classified and decision selected. The all applicable rules are found and help to calculate the priority; based on the priority the selection has been made. This finally improves both the interpretability of rule sets and their predictive accuracy.

3.2.1 Advantages of the proposed System

- Ek-EDT has been developed to get better accuracy than existing techniques.
- It is an iterative decision support detection tool, which helps to the bankers for appropriate decision making.
- The proposed system overcomes the problem of over-fit the training data. And reduces training overhead.
- The proposed EDT algorithm has the ability to handle missing values.
    - Identifies the best plan which is suitable for the customers and insist them to repay the loan
    - Brings the appropriate decisions to the customers based on the above features.



3.2.1 Data collection and uploading

The data used in this study was prepared from the Agri department. The dataset included soil details based on the area and weather details, location attributes (latitude, longitude), etc. Data was collected for Coimbatore district.

After analyzing the customer profiles and predicts of loan risks for bank, a method to predict the credit risk and also to estimate the loan repayment risk. The bank dataset of customer details which are required for data mining are collected and got familiarized with. Various attributes needed are also studied.

3.2.2 DATA PREPROCESSING

Data preprocessing steps are applied on the new set of customer loan application data and they are converted to categorical values by applying filters using unsupervised clustering algorithm named an enhanced extended K-means. After the operations are carried out, a total of input instances of individual locations are presented for analysis.

The attributes in the bank data set are filtered and the relevant attributes needed for prediction are selected. After that the incomplete and noisy records in the dataset are removed and prepared for mining.

Proposed algorithm :EK-EDT

After the segmentation, the system performs the rule set definition by implementing the mean, median and variance, the correlation is calculated using Pearson distribution.

Algorithm EK -phase1 **(labeled example S, set of variables X)**

Fig 3.2class analysis process

The graph based results are generated first, from that the system detects the customer details.

Phase 2: Customer type Clustering:

If the user known the customer type, simply they can select, otherwise the list of features will instruct them to find the customer type. The followings are the attributes used in the customer classification. Such as income, spouse income, employment, salary account type. etc., the followings are the output of customer clustering.

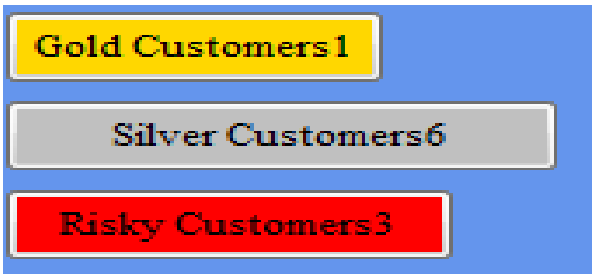


Fig 3.3Customer Classification Training Dataset

Customer Classification using enhanced EK-EDT

Input: Features of customer

Output: customer type and loan repayment risk

Steps:

1. Read training samples S.
2. For each features F in S do
3. Calculate feature score fs=unique(FSi)
4. Order the fs desc and do
5. For each class C in S, tfs=test(fs(FSi(Ci)
6. Return the scores
7. Find the max(tfs) and return the class Ck

CHAPTER 4

SYSTEM SPECIFICATION

4.1 HARDWARE REQURIEMENTS:

- Windows 10 (64 bit) or higher.
- Minimum 8GB RAM and higher.
- Stable Internet Connection Active Internet Connection Minimum Speed 1mbps and above.

4.2 SOFTWARE REQURIMENTS:

Front End : ASP.NET

Back End : SQL

Coding Language : C#

CHAPTER 5

SOFTWARE DESCRIPTION

5.1 Front End ASP.NET:

Web Forms, ASP.NET MVC, and ASP.NET Web Pages are the three web application frameworks that ASP.NET provides. Any of the three frameworks may be used to develop excellent web apps since they are all reliable and sophisticated. You may take advantage of all the capabilities and advantages of ASP.NET anywhere, regardless of the framework you select.

Every framework focuses on a distinct approach to development. The one you select will rely on the kind of application you're developing, your comfort level with the development process, and your programming assets (knowledge, abilities, and development experience).

5.2 BACK END SQL :

A programming language called structured query language (SQL) is used to store and process data in relational databases. Information is stored in tabular form in relational databases, where distinct data properties and the numerous relationships between the data values are represented by rows and columns. Information can be stored, updated, removed, searched for, and retrieved from databases using SQL commands. SQL can also be used to optimize and performance.

One common query language that is widely used in various kinds of applications is structured query language (SQL). SQL is a language that developers and data analysts learn and use because it works well with a variety of programming languages. For instance, they can create high-performing data processing apps with popular SQL database systems by integrating SQL queries into Java code.

CHAPTER 6

PROJECT DESCRIPTION

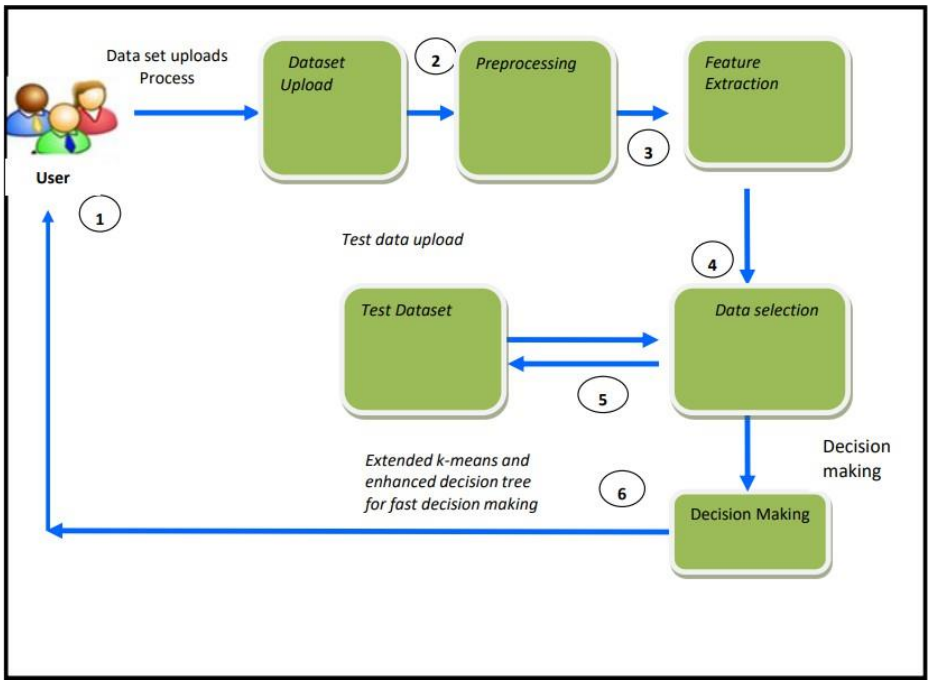
6.1 PROBLEM DEFINITION:

• In the financial sector, particularly in banking, assessing customer credit risk isa critical task. It involves evaluating the likelihood of a borrower defaulting on their loan payments based on various factors such as credit history, income, employment status, and other financial indicators. Additionally, providing appropriate loan repayment suggestions tailored to individual customers can enhance loan performance and reduce the risk of defaults.

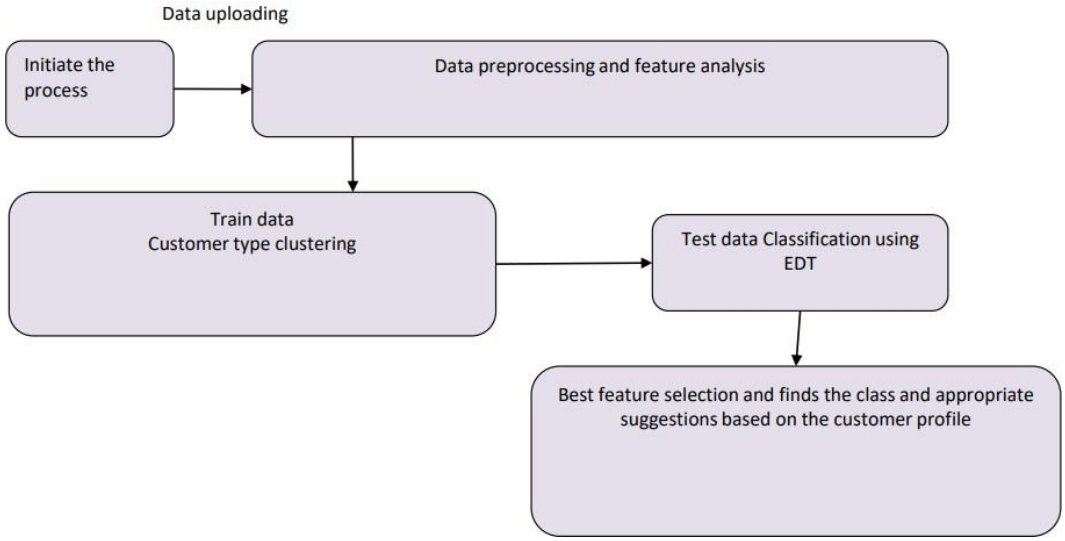
• The problem at hand is to develop an efficient system for a bank that canaccurately calculate customer credit risk and provide personalized loanrepayment suggestions. The system should be able to handle large volumes ofcustomer data, analyze it effectively, and generate actionable insights tomitigate credit risk and optimize loan repayment strategies.

6.2 INTRODUCTION TO PROPOSED SYSTEM:

6.2.1 SYSTEM ARCITECTURE:



Process Flow DIAGRAM



6.3 MODULE DESCRIPTION

6.3.1 Dataset Collection

Initially the system collects numerous customer records from UCI repository. The data set includes several fields related to the bank customer details. initially the system performs data preprocessing. Finally, the pre-processed data were transformed into a suitable format to apply data mining techniques. The first stage in the proposed work is the process initialization, where the data collection, selection and transformation process are done.

6.3.2 Data Uploading

The data used in this study was prepared from the bank sector department. The dataset included customer personal and loan details based on the loan amount and bank details, customer attributes (personal, salary in formations). After analyzing the customer profiles and predicts of loan risks for bank, a method to predict the credit risk and also to estimate the loan repayment risk..The bank dataset of customer details which are required for data mining are collected and got familiarized with. Various attributes needed are also studied.

6.3.3 Data Preprocessing

Data preprocessing steps are applied on the new set of customer loan application data and they are converted to categorical values by applying filters using unsupervised clustering algorithm named an enhanced extended K-means. After the operations are carried out, a total of input instances of individual locations are presented for analysis.

6.3.4 Data Selection

At this stage, data relevant to the analysis was decided on and retrieved from the extracted customer data set. The extracted customer data set had several attributes; their type and description are presented in this module 10

6.3.5 Customer Data Descriptions with Result

The data set used in this research is divided into training and testing data sets. All training cases are set by default taking into account the banks' guidelines for personal credit approvalVariables are the conditions or characteristics that he investigator manipulates, controls or observes.

CHAPTER 7

SYSTEM IMPLEMENTATION

7.1 MACHINE LEARNING MODELS:

7.1.1 RULE-BASED ALGORITHM:

A rules-based algorithm, also known as a rule-based system, uses a set of predefined rules to make decisions or perform actions. These rules are typically in the form of "if-then" statements, where conditions are checked, and corresponding actions are taken based on those conditions.

For example, a simple rule-based algorithm for deciding whether to approve a loan might include rules such as:

If the applicant's credit score is above 700, approve the loan.

If the applicant's income is below $30,000, reject the loan.

If the loan amount requested is above $100,000, require additional documentation.

These rules are applied sequentially, with each rule potentially influencing the final decision. Rule-based algorithms are often used in decision support systems, expert systems, and business process automation. They are relatively easy to understand and modify but can become complex as the number of rules increases.

7.1.2 PASSIVE AGGRESSIVE ALGORITHM:

A passive-aggressive algorithm is a machine learning approach used for classification and regression tasks. It's called "passive-aggressive" because it tries to minimize the loss function while being passive about correct classifications and aggressive about incorrect ones. Essentially, it updates the model in a way that corrects mistakes without making significant changes to correctly classified instances.

CHAPTER 8

RESULTS AND DISCUSSION

8.1 EXPERIMENT RESULT:

8.1.1 Dataset:

The experiment uses the real time and as well as synthetic data sets for experiments. In particular, initially the data set has been collected from different sources. The customer dataset are collected from the UCI repository etc., the dataset includes location; customer details and appropriate decision are collected from the literature. The system can have n number of tuples for experiments.

Dataset 1: (a) Customer Feature dataset with several properties for three types of customers named as Golden, silver and risky.

ID	age	job	marital	education	default	loan_in _other	contact	campaign	gender	poutcome	income	income_spouse
1	30	unemployed	married	primary	no	yes	cellular	1	Female	unknown	0	20000
2	33	services	married	secondary	no	no	cellular	1	male	failure	30000	12000
3	35	management	single	tertiary	no	no	cellular	1	male	failure	25000	0
4	30	management	married	tertiary	no	no	unknown	4	male	unknown	32000	23000
5	59	blue-collar	married	secondary	yes	yes	unknown	1	Female	unknown	35000	34000

children_count	children_depen	loan	existing_paid	physically_challenge	salary_ac_type	loan_amt	paid_amt	balance
2	yes	yes	no	no	bank	100000	20000	80000
1	yes	yes	no	no	cash	200000	50000	150000
0	no	yes	null	no	cash	500000	30000	470000
1	no	yes	no	no	cash	600000	500000	100000
1	yes	yes	no	yes	cash	200000	30000	170000

Fig 4.1 Dataset Sample

Dataset 1: (b) Customer Feature dataset with several properties for three types of customers named as Golden, silver and risky.

loan_amt	paid_amt	balance
100000	20000	80000
200000	50000	150000
500000	30000	470000
600000	500000	100000
200000	30000	170000
100000	30000	70000
600000	10000	590000
300000	20000	280000
400000	50000	350000
500000	80000	420000

Fig 8.2 Dataset Sample2

After the data set is generated, and given the number of m classes, each tuple from the synthetic dynamic database D is assigned to database Si chosen uniformly. Clearly, all datasets are formatted in a same manner. In particular, a the dataset are converted into desired type for different type of implementation sample set of the underlying data set, and the sample sets are mutually disjoint.

Experiment: An interactive software using .Net framework

The system has used Visual Studio.Net framework. And C#.Net has been used for developing the front end and SQL Server for the back end. The reason for using C#.Net is its flexibility. This can add or remove any features without editing the whole code. This separated the standalone functions like OOPS functions which are reused again and again. For the back end this needed a freely distributed and powerful database so SQL Server

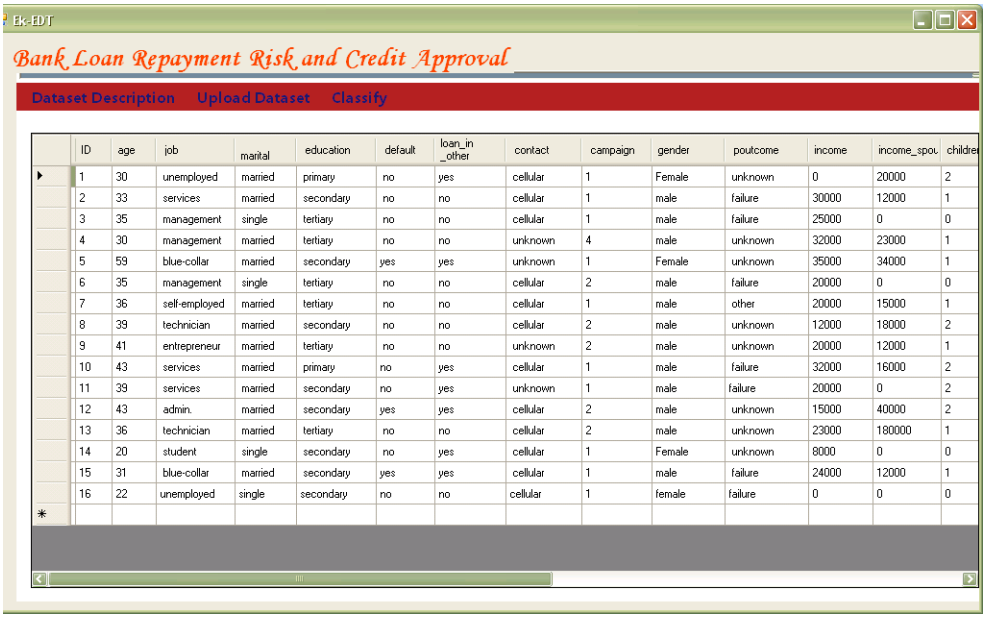
Implementation steps:

1.Data collection :

In the proposed system implementation, the customer dataset has been collected from UCI repository, the collected data has been uploaded to the database after detecting inconsistent and missing values. the collected dataset contains 24 attributes.

Fig: dataset description process.

The training dataset is identified and that will be uploaded to the database separately. The following fig represents the uploaded dataset.

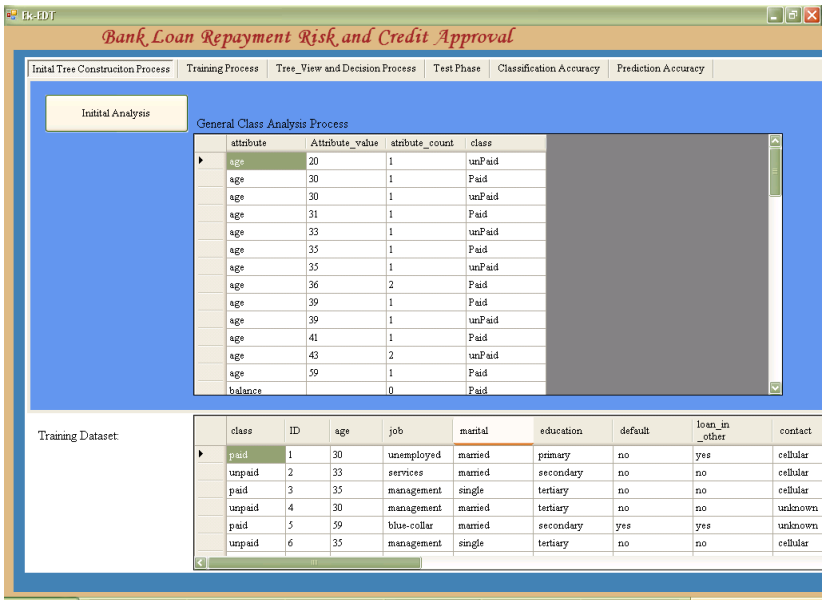


	ID	age	job	marital	education	default	loan_in _other	contact	campaign	gender	poutcome	income	income_spou	childre
▶	1	30	unemployed	married	primary	no	yes	cellular	1	Female	unknown	0	20000	2
	2	33	services	married	secondary	no	no	cellular	1	male	failure	30000	12000	1
	3	35	management	single	tertiary	no	no	cellular	1	male	failure	25000	0	0
	4	30	management	married	tertiary	no	no	unknown	4	male	unknown	32000	23000	1
	5	59	blue-collar	married	secondary	yes	yes	unknown	1	Female	unknown	35000	34000	1
	6	35	management	single	tertiary	no	no	cellular	2	male	failure	20000	0	0
	7	36	self-employed	married	tertiary	no	no	cellular	1	male	other	20000	15000	1
	8	39	technician	married	secondary	no	no	cellular	2	male	unknown	12000	18000	2
	9	41	entrepreneur	married	tertiary	no	no	unknown	2	male	unknown	20000	12000	1
	10	43	services	married	primary	no	yes	cellular	1	male	failure	32000	16000	2
	11	39	services	married	secondary	no	yes	unknown	1	male	failure	20000	0	2
	12	43	admin.	married	secondary	yes	yes	cellular	2	male	unknown	15000	40000	2
	13	36	technician	married	tertiary	no	no	cellular	2	male	unknown	23000	180000	1
	14	20	student	single	secondary	no	yes	cellular	1	Female	unknown	8000	0	0
	15	31	blue-collar	married	secondary	yes	yes	cellular	1	male	failure	24000	12000	1
	16	22	unemployed	single	secondary	no	no	cellular	1	female	failure	0	0	0
*														

Fig uploaded dataset

2.Initial analysis and tree construction process:

After successful data collection, the data will be analyzed for tree generation, which collects each attribute and performs the analysis for every class C. the proposed system has two type of classes.



	attribute	Attribute_value	atribute_count	class
▶	age	20	1	unPaid
	age	30	1	Paid
	age	30	1	unPaid
	age	31	1	Paid
	age	33	1	unPaid
	age	35	1	Paid
	age	35	1	unPaid
	age	36	2	Paid
	age	39	1	Paid
	age	39	1	unPaid
	age	41	1	Paid
	age	43	2	unPaid
	age	59	1	Paid
	balance		0	Paid

	class	ID	age	job	marital	education	default	loan_in _other	contact
▶	paid	1	30	unemployed	married	primary	no	yes	cellular
	unpaid	2	33	services	married	secondary	no	no	cellular
	paid	3	35	management	single	tertiary	no	no	cellular
	unpaid	4	30	management	married	tertiary	no	no	unknown
	paid	5	59	blue-collar	married	secondary	yes	yes	unknown
	unpaid	6	35	management	single	tertiary	no	no	cellular

Fig initial analysis and tree construction process

3. Clustering process using Extended K-means:

The proposed system extended k-means clustering process will cluster the customers in to three groups named as Golden customers, silver customers and risky customers. The gien value of k is 3 in the proposed system.

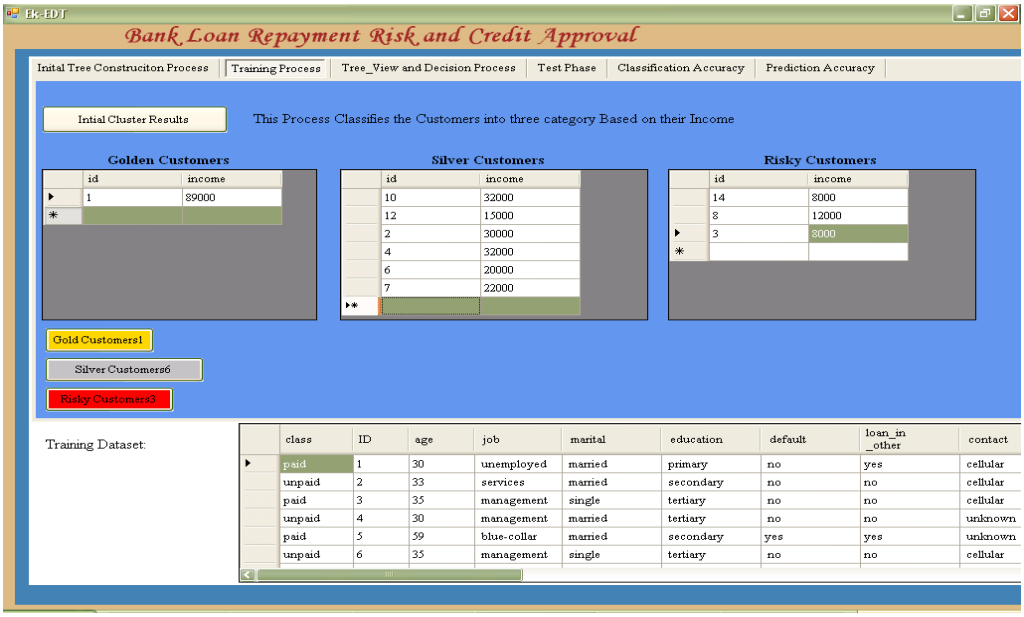


Fig Clustered result page

The above fig shows the results of the extended k-means algorithm, which finds the risky customers from their income level. This categorizes the customers into golden if the customer income is greater than 50000, and silver if their income is above 15000 and risky if their income level is below 15000.

4. Tree View and decision analysis process

After the clustering phase, for every attribute the training process creates a rule. The collected rules are used at the time of classification.

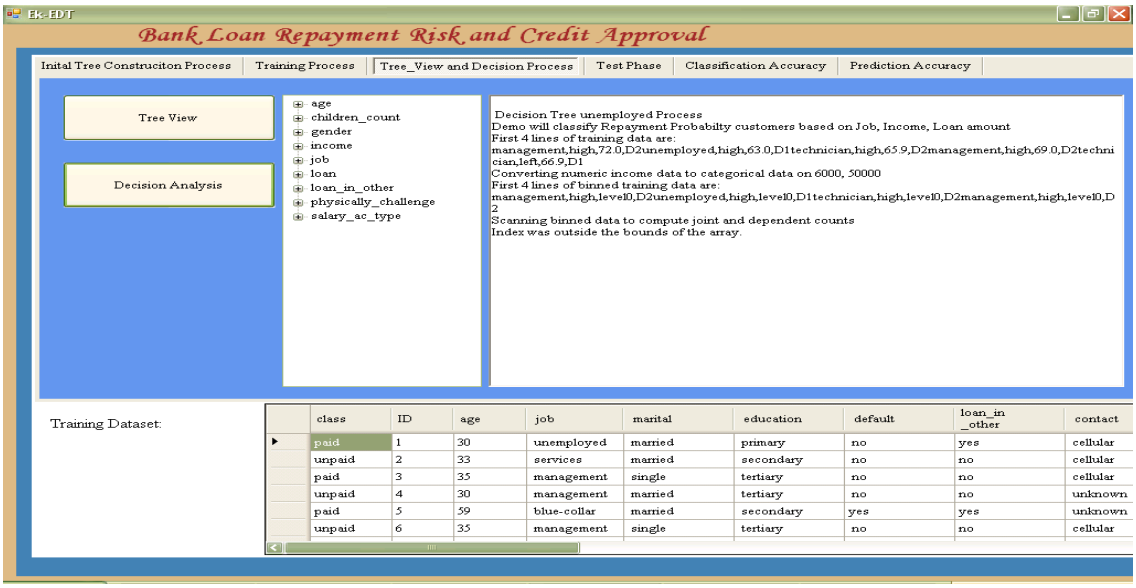


Fig tree view and decision analysis page

5. Test/Classification Phase:

Finally the training samples will be used for test phase. This phase initially asks the customer id to classify. Based on the EDT modal, the system analyses and predicts the class of the given test sample. The test sample should include all the possible values such as age, income, children count etc., based on the data, the system finds whether the customer will repay the loan amount or not. This will be identified by the classification criteria score '0.056320386964977.821". Along with the criteria, the percentage of the class will be analyzed.

8.2 Results And Discussion:

In this chapter, this evaluates the efficiency of the algorithms, in terms of time consumption against dimensionality d, number of traversal of tree, and tree generation threshold q under two distributions of different customers. This also evaluates the progressiveness of the methods under different datasets.

This section evaluates the proposed Credit risk analysis with EK-EDT data framework in terms of both tree traversal overhead and accuracy. We applied Credit risk analysis on sample customer's data for experiments.

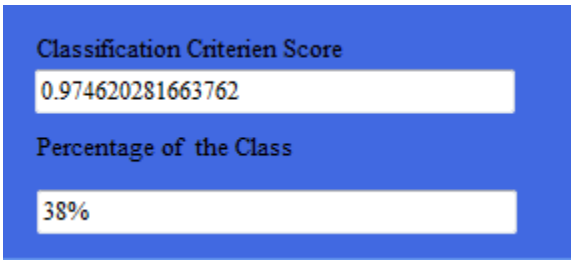


Fig:8.3 Sample output of the proposed system

The above figure represents the suggestions at every iteration based on the current customer type and details.

For the experiment, An Intel I3 2.2 GHz processor with 2 Gb RAM was used to measure the execution time and detection speed. Table 4.2 describes the execution time for varying dataset values and Table4.2 gives the accuracy for varying dataset value.

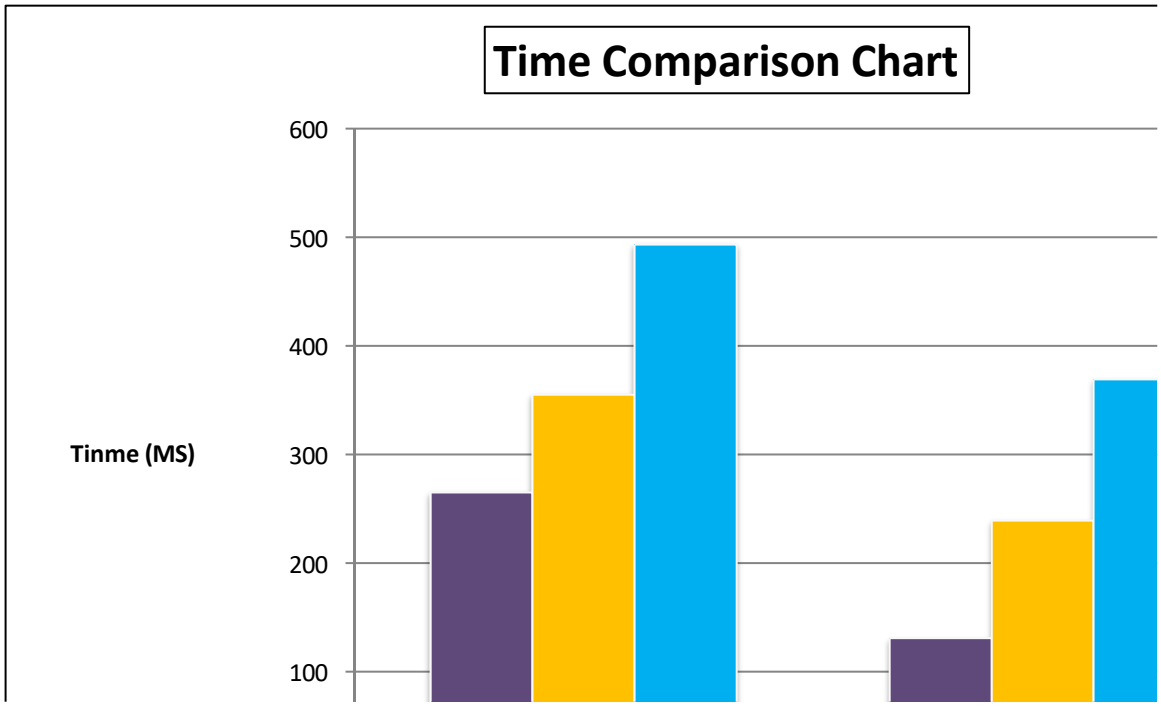
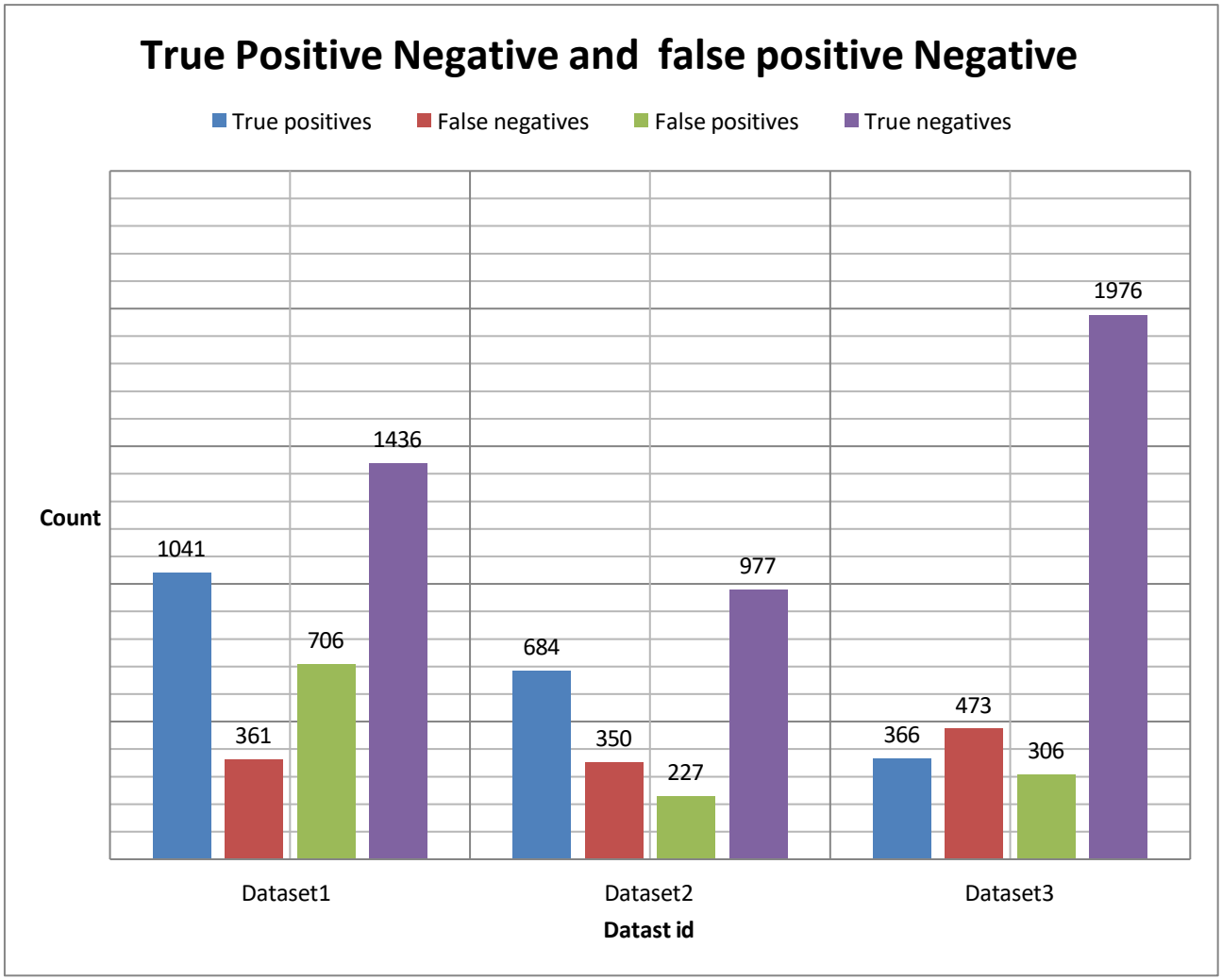
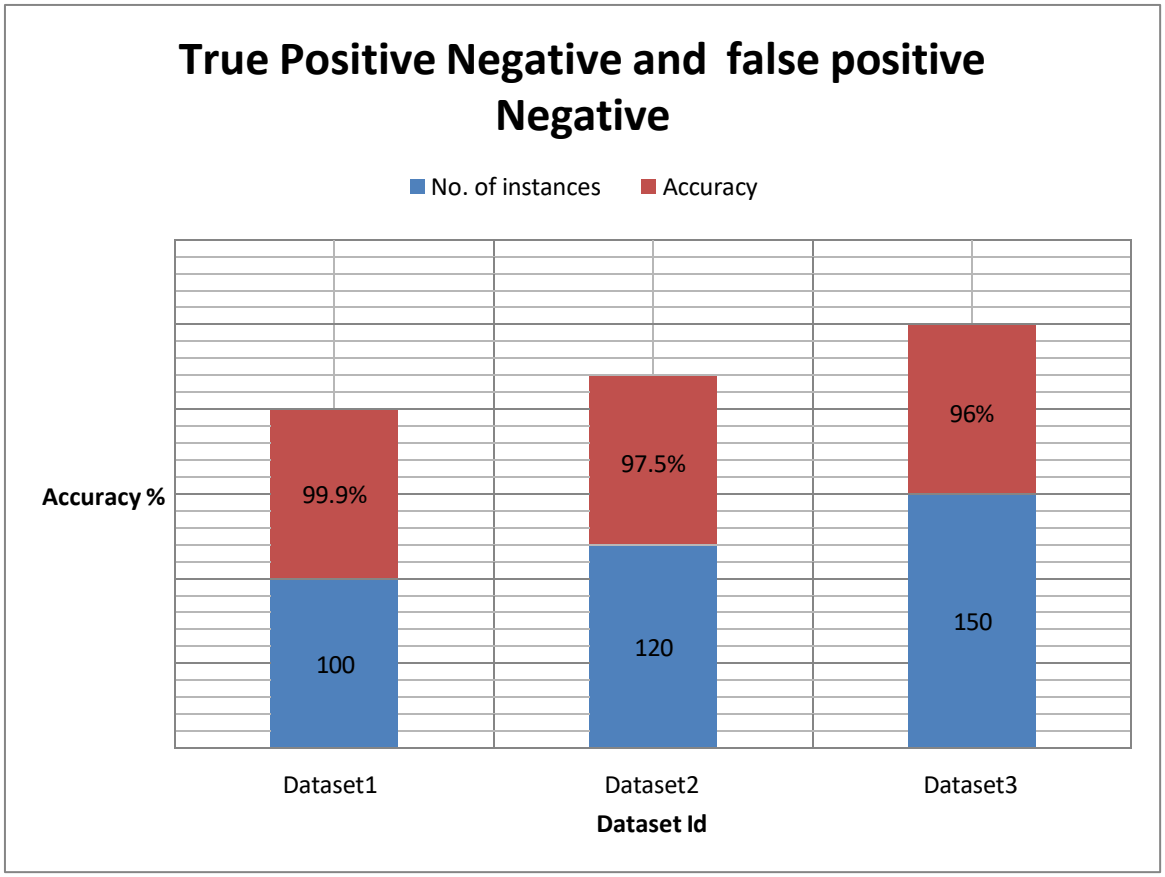


Fig 4.4Execution Time Comparison Chart

The following figure 4.4 shows comparison between the existing J48 and Credit risk analysis EK-EDT, from the experiments the results shows the proposed system faster than the existing system.


True Positive Negative and false positive Negative
True positives
False negatives
False positives
True negatives
Count
1041
361
706
1436
684
350
227
977
366
473
306
1976
Dataset1
Dataset2
Dataset3
Datast id


True Positive Negative and false positive Negative
No. of instances
Accuracy
Accuracy %
99.9%
97.5%
96%
100
120
150
Dataset1
Dataset2
Dataset3
Dataset Id

CHAPTER 9

CONCLUSION AND FUTURE ENHANCEMENTS

9.1 CONCLUSION

Banking credit risk analysis is became difficult to learn, when the user doesn't know anything about the customer prior. This toughness is due to the high dimensional and dynamic data. And the loan repayment options are differs customer to customer. This thesis presented an overview of the banking customer data required for decision making and provides an interactive GUI based tool to analyze the customer type, loan repayment ability, and risk score. The decisions related to the customer loan repayment activities are highlighted.

This includes the enhanced EDT algorithm for fast decision support process. This finds the appropriate solutions and decisions based on the given attributes and values. the experiments are conducted with various conditions and factors to evaluate the output of the proposed system. And the experiments are carried out using Dotnet technology. This helps the organizations in making the right decision to approve or reject the loan request of the customers based on the credit risk. . a Decision Tree Algorithm named as EDT is used for the prediction.. We obtained statistically significant linear and nonlinear models to accomplish the above.

The study proposed a new classification and prediction scheme for bank data. The system studied the main two problems in the literature, which are prediction accuracy and classification delay. The study overcomes the above two problem by applying the effective enhanced decision tree with data suggestion. The EDT represents with the effective splitting criteria which has been verified by the data suggestion. The system performs pre pruning and post pruning to eliminate irrelevant results. The system effectively identifies the disease and its sub types, the sub type which is referred as the percentage of class such as normal and disease.

The experimental results are evaluated using the C#.net. The experimental result shows that integrated extended decision tree with data suggestion shows better quality assessment compared to traditional C4.5 techniques. From the experimental results, the execution time calculated for classification object is almost reduced than the existing system.

9.2 FUTURE ENHANCEMENT

There is also a need to understand the decision support frameworks and the processes and inputs required to facilitate decision making accurately for all customers, because the current work handles only a probabilistic solution. The outcome of the project is based on the current and historical customer loan approval and rejection feature and details. The work is fully relies on the training data. If the improper customer details and risk factors are given, then it will affect the whole output. So concentrating on the accurate dataset creation can be extended in future.

The proposed framework model can be used to analyze the existing work, identify gaps and provide scope for further works. The researchers may use the model to identify the existing area of research in the field of data mining in other dataset and use of other classification algorithms. As further work, use this model as a functional base to develop an appropriate data mining system for classification performance.

APPENDIX

SOURCE CODE

```
return _doubleList;

    }

publicbool BulkInserSheet3Table(string tableName, DataTable tdataTable)
    {
bool isSuccuss;
try
        {
          SqlConnection SqlConnectionObj = newSqlConnection("server=.;Initial
Catalog=bank;Integrated Security=True;");

SqlConnectionObj.Open();
// SqlConnection SqlConnectionObj = GetSQLConnection();
          SqlBulkCopy bulkCopy = newSqlBulkCopy(SqlConnectionObj,
SqlBulkCopyOptions.TableLock | SqlBulkCopyOptions.FireTriggers |
SqlBulkCopyOptions.UseInternalTransaction, null);
          bulkCopy.DestinationTableName = tableName;
bulkCopy.WriteToServer(tdataTable);
isSuccuss = true;
SqlConnectionObj.Close();
        }
catch (Exception ex)
        {
isSuccuss = false;
        }
return isSuccuss;
```

```
        }
publicvoid Clear()
        {
dataSet = new DataSet();
        }
        #endregion
publicvoid decisionprocess()
        {
            SqlConnection co1n = newSqlConnection("server=.;Initial Catalog=bank;Integrated Security=True;");

co1n.Open();
            SqlDataAdapter da1 = newSqlDataAdapter("select * from customers1", co1n);
            DataSet ds1 = newDataSet();
da1.Fill(ds1);
//dtable = ds.Tables[0];
            dataGridView1.DataSource = ds1.Tables[0];
co1n.Close();
if (textBox3.Text != "")
            {
// textBox2.Text = "";
//  textBox1.Text = "";
string a, b, c, d, ed, f, g, h, i, jj, k, l, m,n,o,p,q,r,s,t,u,v;
int n1;

//
                n1 = Convert.ToInt32(textBox3.Text);
int nn1 = Convert.ToInt32(dataGridView1.RowCount.ToString());
if (nn1 < n1)
                {
MessageBox.Show("Data Source not found..");
```

```
                textBox3.Text = "";
            }
int rn = n1;

            a = dataGridView1.Rows[rn].Cells[2].Value.ToString();
            b = dataGridView1.Rows[rn].Cells[3].Value.ToString();
            c = dataGridView1.Rows[rn].Cells[4].Value.ToString();
            d = dataGridView1.Rows[rn].Cells[5].Value.ToString();
ed = dataGridView1.Rows[rn].Cells[6].Value.ToString();
            f = dataGridView1.Rows[rn].Cells[7].Value.ToString();
            g = dataGridView1.Rows[rn].Cells[8].Value.ToString();
            h = dataGridView1.Rows[rn].Cells[9].Value.ToString();
            i = dataGridView1.Rows[rn].Cells[10].Value.ToString();
jj = dataGridView1.Rows[rn].Cells[10].Value.ToString();
            k = dataGridView1.Rows[rn].Cells[11].Value.ToString();
            l = dataGridView1.Rows[rn].Cells[12].Value.ToString();
            m = dataGridView1.Rows[rn].Cells[13].Value.ToString();
            n = dataGridView1.Rows[rn].Cells[14].Value.ToString();
            o = dataGridView1.Rows[rn].Cells[15].Value.ToString();
            p = dataGridView1.Rows[rn].Cells[16].Value.ToString();
            q = dataGridView1.Rows[rn].Cells[17].Value.ToString();
            r = dataGridView1.Rows[rn].Cells[18].Value.ToString();
            s = dataGridView1.Rows[rn].Cells[19].Value.ToString();
            t = dataGridView1.Rows[rn].Cells[20].Value.ToString();
            u = dataGridView1.Rows[rn].Cells[21].Value.ToString();
            v = dataGridView1.Rows[rn].Cells[22].Value.ToString();
string results = Classify_edt(newstring [] { a, b, c, d, ed, f, g, h,i,jj,k,l,m,n,o,p,q,r,s,t,u,v });
        }
    }
string customer_type;
string final_decision;
```

```
publicstring Classify_edt(string[] obj)
      {
int a = Convert.ToInt32(obj[12].ToString());
if(a >=50000)
         {
            customer_type="Golden";

         }
elseif((a >=15000) && (a < 50000))
         {
            customer_type="silver";

         }
elseif ((a < 6000))
         {
            customer_type = "risky";
int b = Convert.ToInt32(obj[21].ToString());
int remainamt = Convert.ToInt32(obj[20].ToString()) / 2;
if (remainamt != 0)
            {
if (b >= remainamt)
               {
                  final_decision = "2% Discount  on interest";

               }
elseif ((b < remainamt) || (b > 10000))
               {
                  final_decision = "1% Discount  on interest";
               }

else
```

```
                {
                    final_decision = "EMI Facilities";

                }
            }

        }

if (customer_type == "Golden")
        {
int b= Convert.ToInt32(obj[21].ToString());
int remainamt=Convert.ToInt32(obj[20].ToString())/2;
if (remainamt != 0)
            {
if (b >= remainamt)
                {
                    final_decision = "2% Discount  on interest";

                }
elseif ((b < remainamt) || (b > 10000))
                {
                    final_decision = "1% Discount  on interest";
                }

else
                {
                    final_decision = "EMI Facilities";

                }
            }
        }
```

```
elseif (customer_type == "silver")
        {
int b=0;
try
            {
               b = Convert.ToInt32(obj[21].ToString());
            }
catch (Exception ex)
            { }
int remainamt = Convert.ToInt32(obj[20].ToString()) / 3;
if (remainamt != 0)
            {
if (b >= remainamt)
               {
                  final_decision = "1% Discount  on interest";

               }
elseif ((b < remainamt) || (b > 10000))
               {
                  final_decision = "EMI options";
               }

else
               {
                  final_decision = "Payment data can be extended";

               }
            }
        }
if((final_decision != "")&&(final_decision != null ))
```

SNAP SHOTS

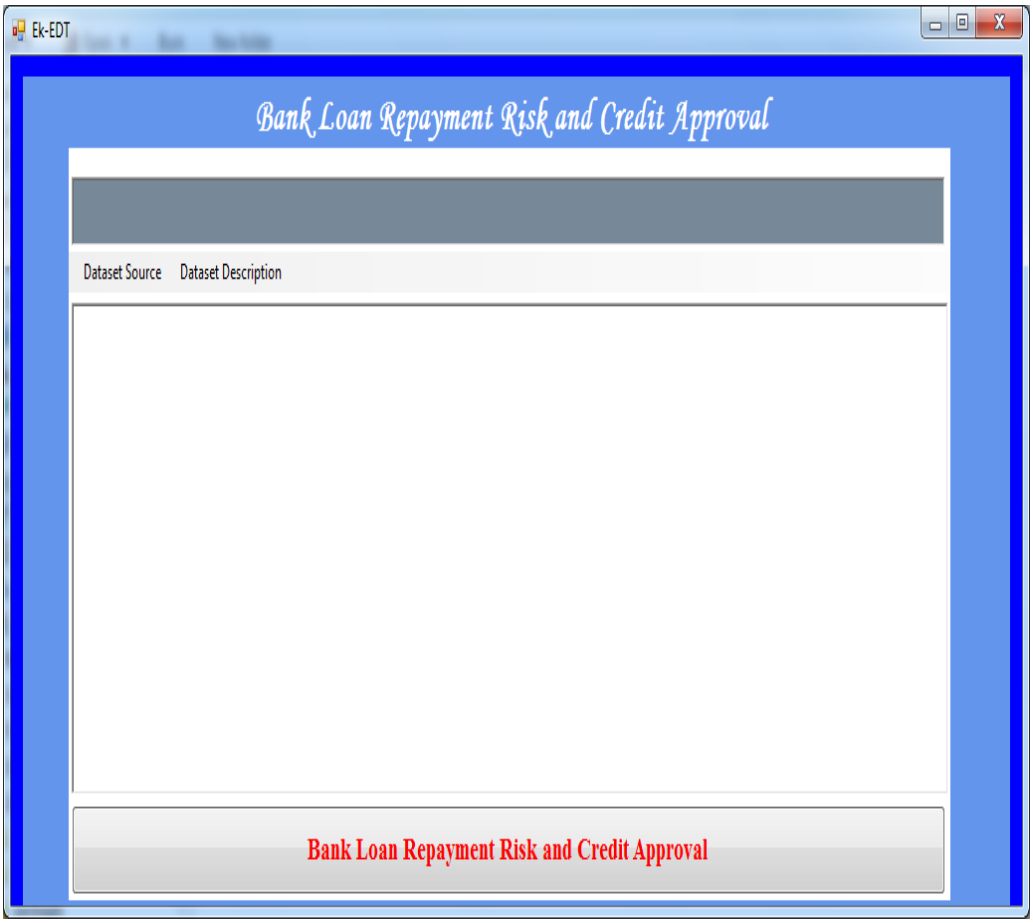


Figure 9.1

Login Successful Windows

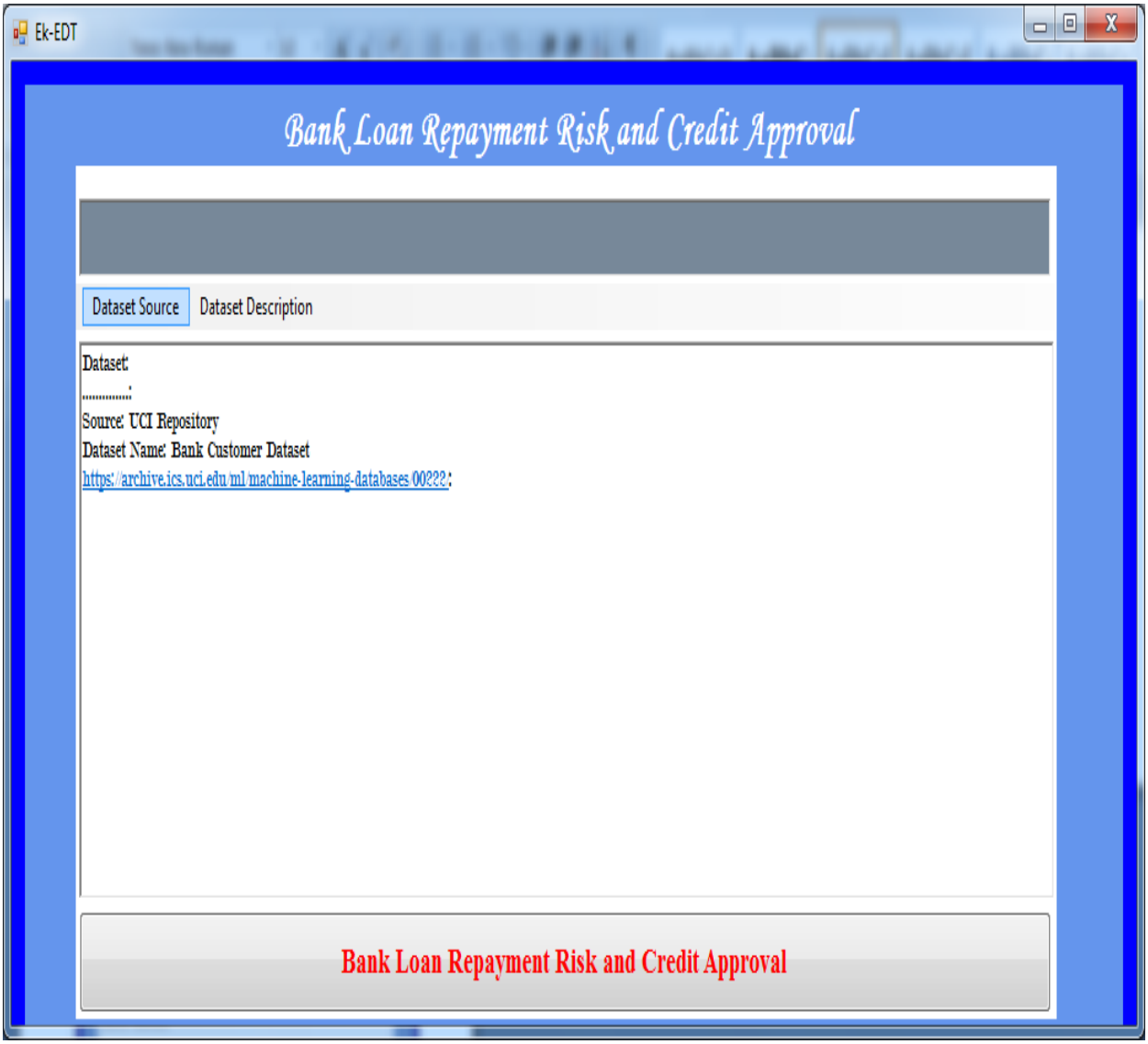

Ek-EDT
Bank Loan Repayment Risk and Credit Approval
Dataset Source
Dataset Description
Dataset:
Source: UCI Repository
Dataset Name: Bank Customer Dataset
https://archive.ics.uci.edu/ml/machine-learning-databases/00222/
Bank Loan Repayment Risk and Credit Approval

Figure 9.2

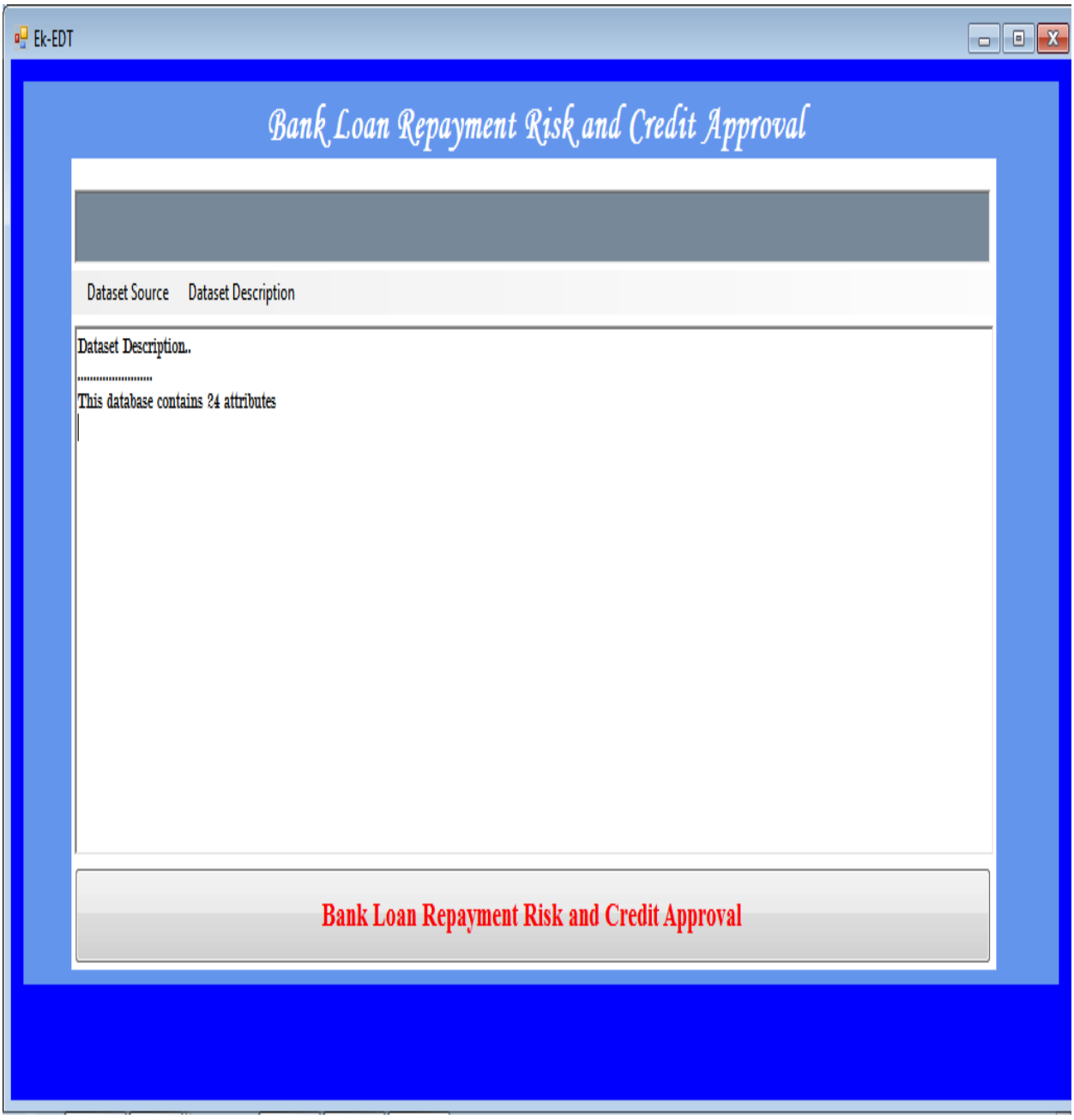

Ek-EDT
Bank Loan Repayment Risk and Credit Approval
Dataset Source
Dataset Description
Dataset Description..
This database contains 24 attributes
Bank Loan Repayment Risk and Credit Approval

Figure 9.3

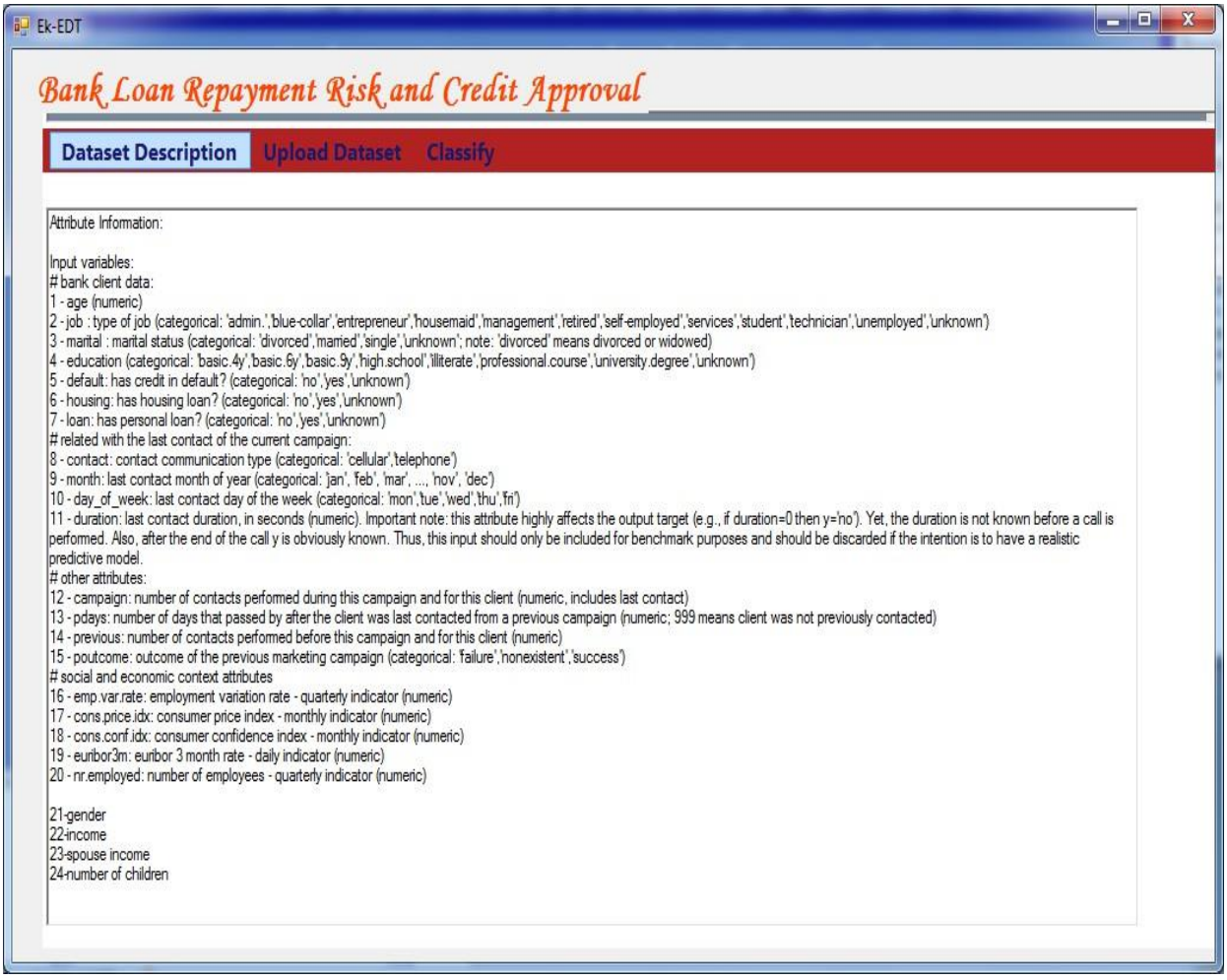
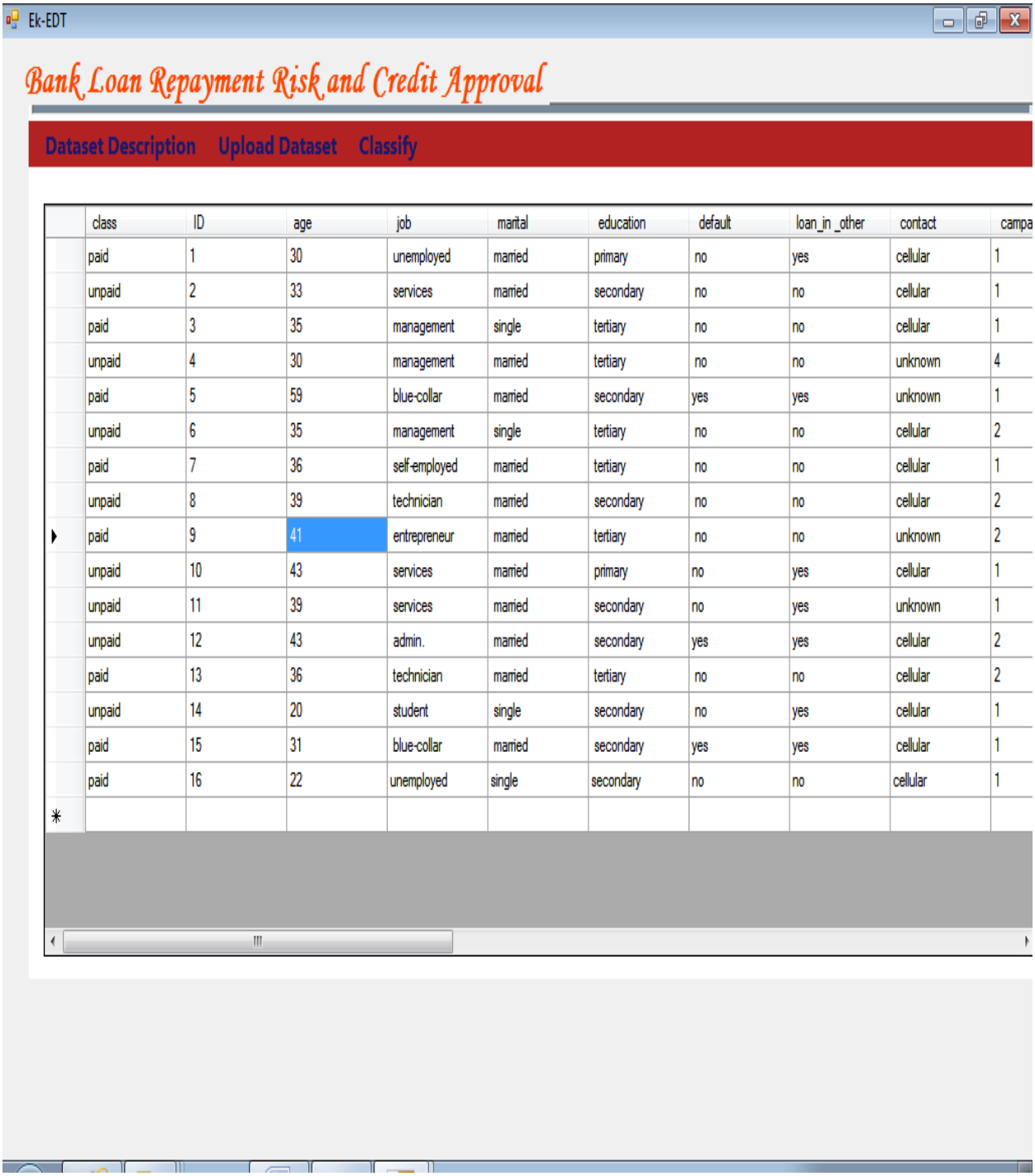

Ek-EDT
Bank Loan Repayment Risk and Credit Approval
Dataset Description
Upload Dataset
Classify
Attribute Information:
Input variables:
bank client data:
1 - age (numeric)
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5 - default: has credit in default? (categorical: 'no','yes','unknown')
6 - housing: has housing loan? (categorical: 'no','yes','unknown')
7 - loan: has personal loan? (categorical: 'no','yes','unknown')
related with the last contact of the current campaign:
8 - contact: contact communication type (categorical: 'cellular','telephone')
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
other attributes:
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14 - previous: number of contacts performed before this campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
social and economic context attributes
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
17 - cons.price.idx: consumer price index - monthly indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - nr.employed: number of employees - quarterly indicator (numeric)
21-gender
22-income
23-spouse income
24-number of children

Figure 9.4



class	ID	age	job	marital	education	default	loan_in _other	contact	campa
paid	1	30	unemployed	married	primary	no	yes	cellular	1
unpaid	2	33	services	married	secondary	no	no	cellular	1
paid	3	35	management	single	tertiary	no	no	cellular	1
unpaid	4	30	management	married	tertiary	no	no	unknown	4
paid	5	59	blue-collar	married	secondary	yes	yes	unknown	1
unpaid	6	35	management	single	tertiary	no	no	cellular	2
paid	7	36	self-employed	married	tertiary	no	no	cellular	1
unpaid	8	39	technician	married	secondary	no	no	cellular	2
paid	9	41	entrepreneur	married	tertiary	no	no	unknown	2
unpaid	10	43	services	married	primary	no	yes	cellular	1
unpaid	11	39	services	married	secondary	no	yes	unknown	1
unpaid	12	43	admin.	married	secondary	yes	yes	cellular	2
paid	13	36	technician	married	tertiary	no	no	cellular	2
unpaid	14	20	student	single	secondary	no	yes	cellular	1
paid	15	31	blue-collar	married	secondary	yes	yes	cellular	1
paid	16	22	unemployed	single	secondary	no	no	cellular	1

Figure 9.5

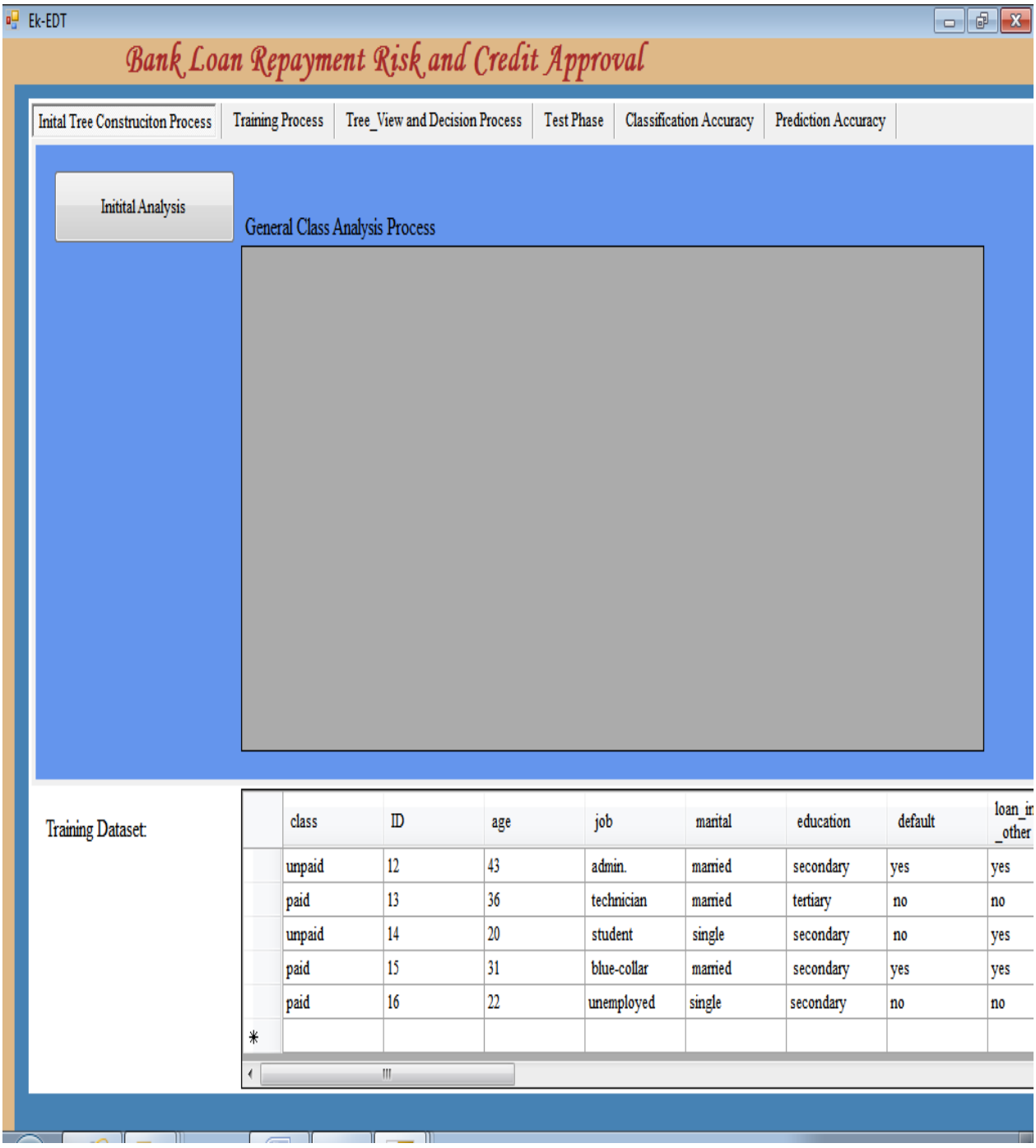

Ek-EDT
Bank Loan Repayment Risk and Credit Approval
Inital Tree Construciton Process
Training Process
Tree_View and Decision Process
Test Phase
Classification Accuracy
Prediction Accuracy
Inititial Analysis
General Class Analysis Process
Training Dataset:
class
ID
age
job
marital
education
default
loan_in
_other
unpaid 12 43 admin. married secondary yes yes
paid 13 36 technician married tertiary no no
unpaid 14 20 student single secondary no yes
paid 15 31 blue-collar married secondary yes yes
paid 16 22 unemployed single secondary no no

Figure 9.6

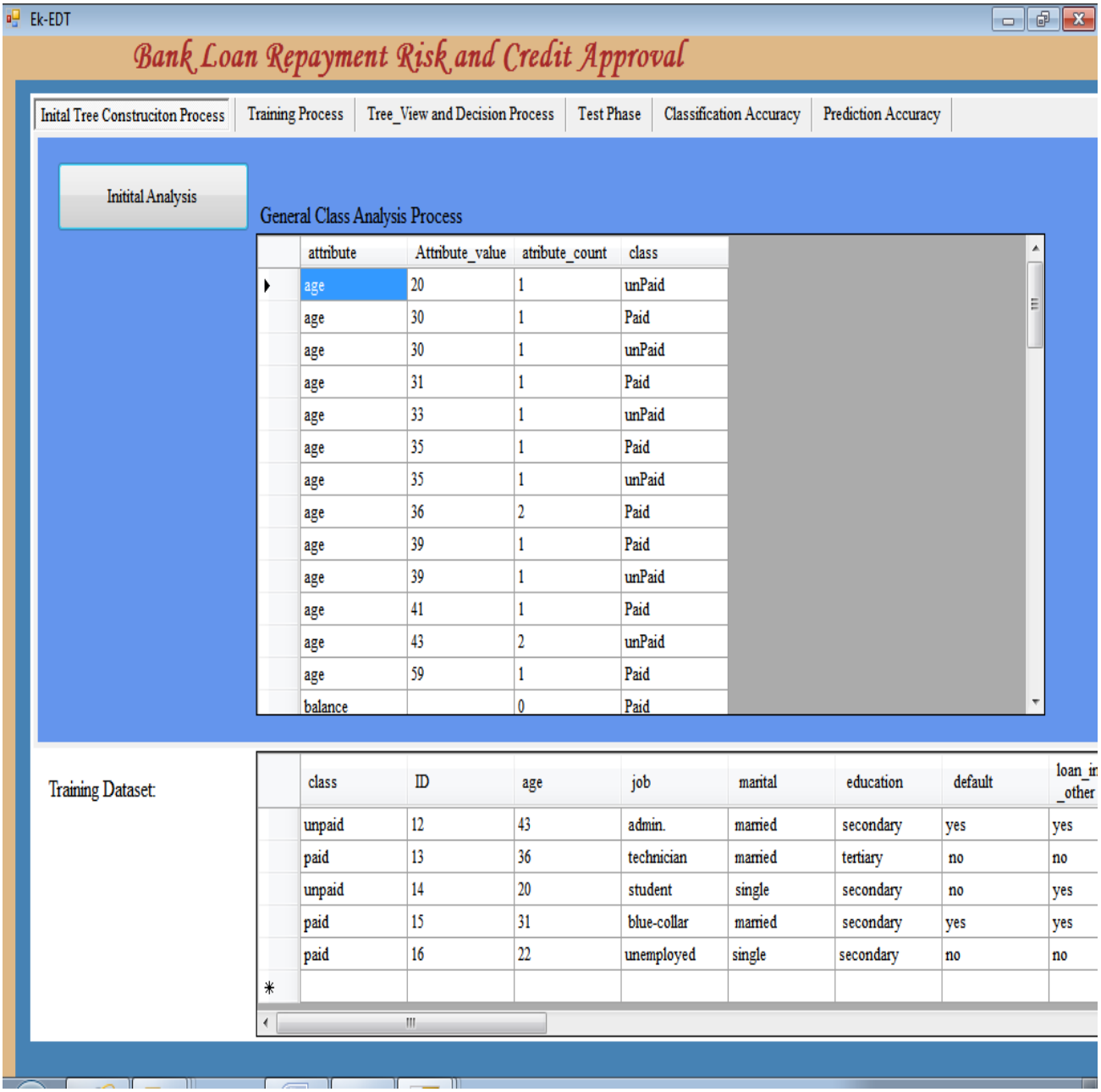

Ek-EDT
Bank Loan Repayment Risk and Credit Approval
Inital Tree Construciton Process
Training Process
Tree_View and Decision Process
Test Phase
Classification Accuracy
Prediction Accuracy
Initital Analysis
General Class Analysis Process
attribute Attribute_value atribute_count class
age 20 1 unPaid
age 30 1 Paid
age 30 1 unPaid
age 31 1 Paid
age 33 1 unPaid
age 35 1 Paid
age 35 1 unPaid
age 36 2 Paid
age 39 1 Paid
age 39 1 unPaid
age 41 1 Paid
age 43 2 unPaid
age 59 1 Paid
balance 0 Paid
Training Dataset:
class ID age job marital education default loan_in _other
unpaid 12 43 admin. married secondary yes yes
paid 13 36 technician married tertiary no no
unpaid 14 20 student single secondary no yes
paid 15 31 blue-collar married secondary yes yes
paid 16 22 unemployed single secondary no no

Figure 9.7

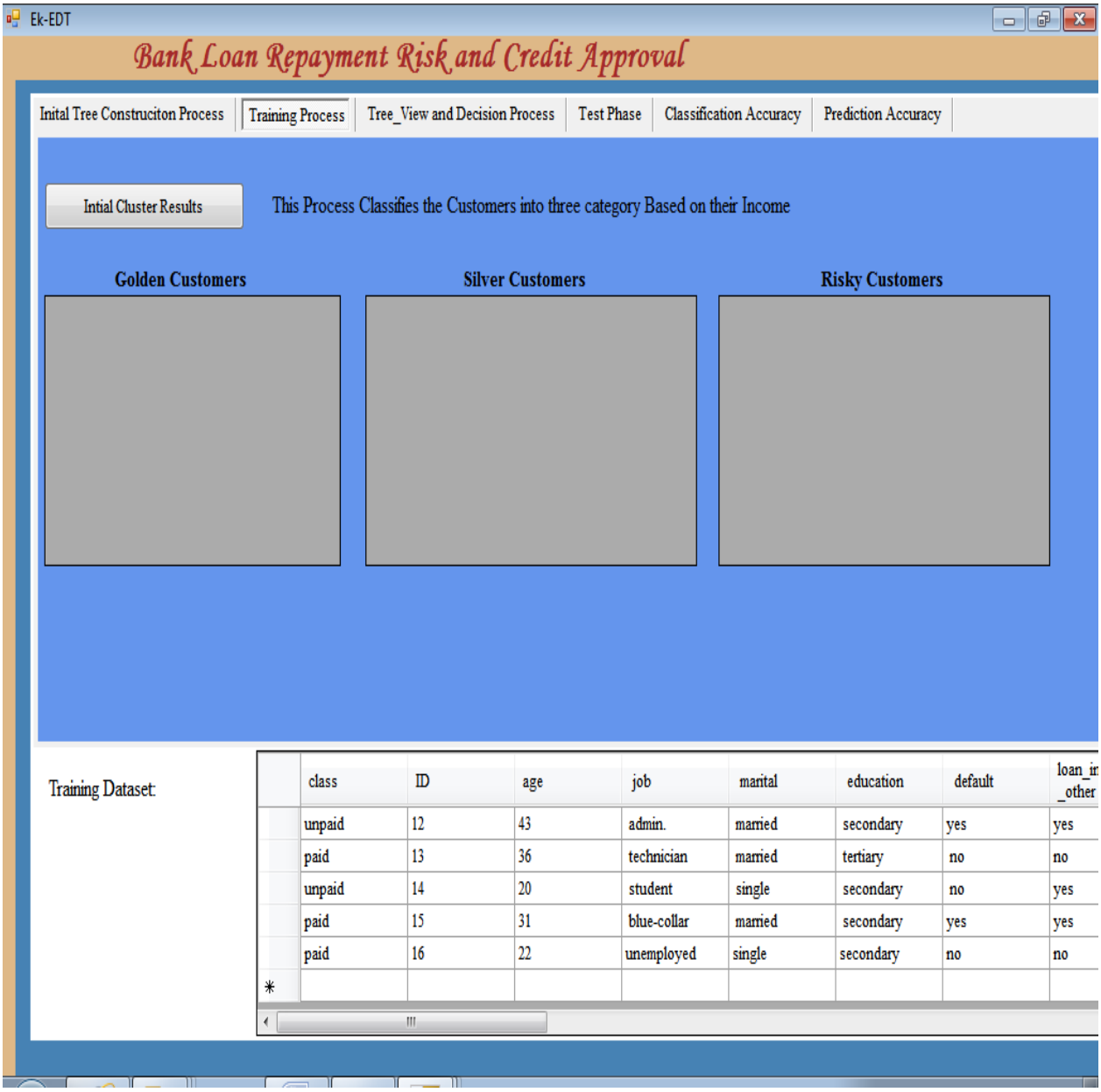

Ek-EDT
Bank Loan Repayment Risk and Credit Approval
Inital Tree Construciton Process
Training Process
Tree_View and Decision Process
Test Phase
Classification Accuracy
Prediction Accuracy
Intial Cluster Results
This Process Classifies the Customers into three category Based on their Income
Golden Customers
Silver Customers
Risky Customers
Training Dataset:
class
ID
age
job
marital
education
default
loan_in
_other
unpaid 12 43 admin. married secondary yes yes
paid 13 36 technician married tertiary no no
unpaid 14 20 student single secondary no yes
paid 15 31 blue-collar married secondary yes yes
paid 16 22 unemployed single secondary no no

Figure 9.8

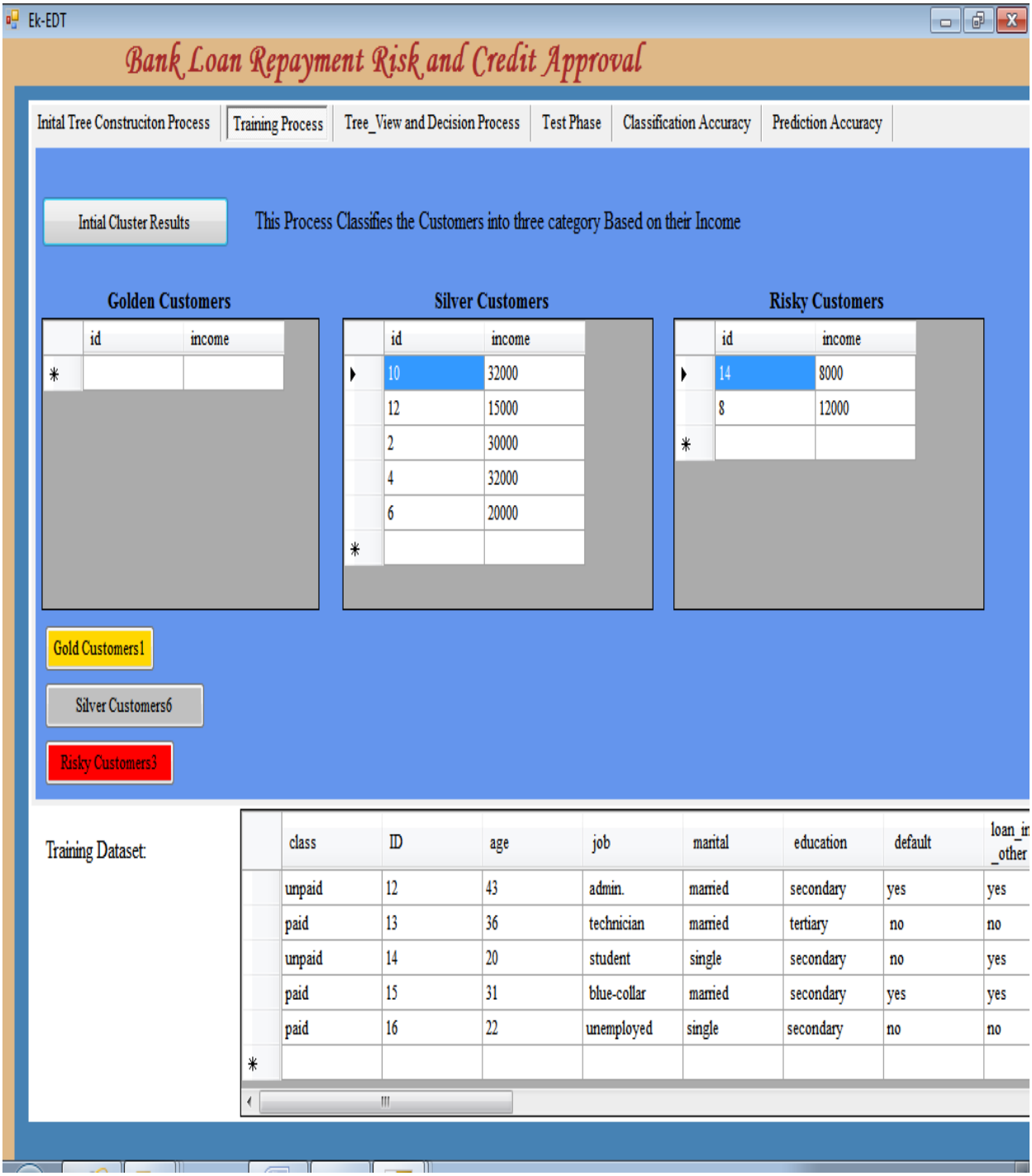

Ek-EDT
Bank Loan Repayment Risk and Credit Approval
Inital Tree Construciton Process
Training Process
Tree_View and Decision Process
Test Phase
Classification Accuracy
Prediction Accuracy
Intial Cluster Results
This Process Classifies the Customers into three category Based on their Income
Golden Customers
Silver Customers
Risky Customers
id
income
10
32000
12
15000
2
30000
4
32000
6
20000
14
8000
8
12000
Gold Customers1
Silver Customers6
Risky Customers3
Training Dataset:
class
ID
age
job
marital
education
default
loan_in
_other
unpaid
12
43
admin.
married
secondary
yes
yes
paid
13
36
technician
married
tertiary
no
no
unpaid
14
20
student
single
secondary
no
yes
paid
15
31
blue-collar
married
secondary
yes
yes
paid
16
22
unemployed
single
secondary
no
no

Figure 9.9

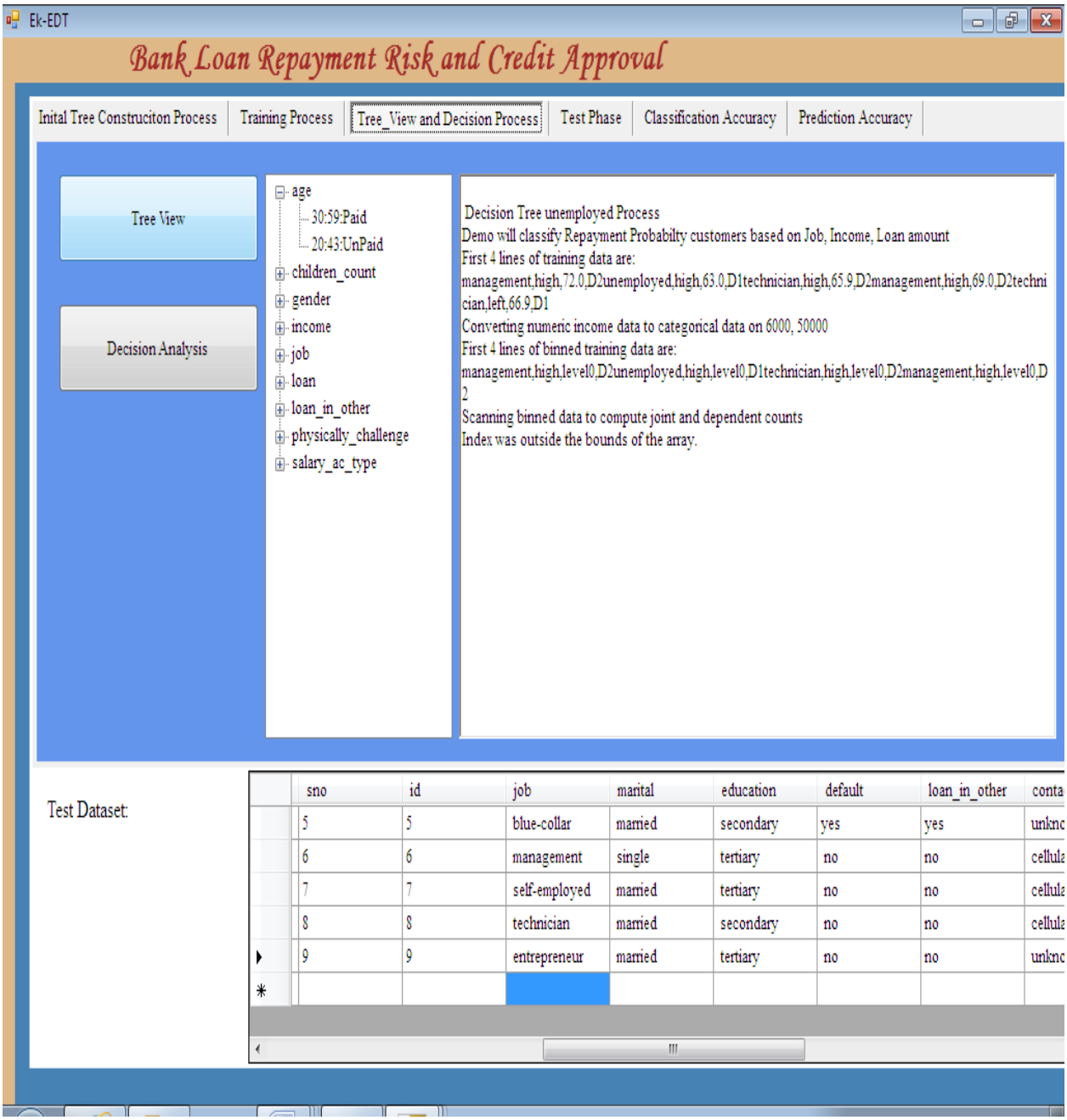

Ek-EDT
Bank Loan Repayment Risk and Credit Approval
Inital Tree Construciton Process
Training Process
Tree_View and Decision Process
Test Phase
Classification Accuracy
Prediction Accuracy
Tree View
Decision Analysis
age
30:59:Paid
20:43:UnPaid
children_count
gender
income
job
loan
loan_in_other
physically_challenge
salary_ac_type
Decision Tree unemployed Process
Demo will classify Repayment Probabilty customers based on Job, Income, Loan amount
First 4 lines of training data are:
management,high,72.0,D2unemployed,high,63.0,D1technician,high,65.9,D2management,high,69.0,D2technician,left,66.9,D1
Converting numeric income data to categorical data on 6000, 50000
First 4 lines of binned training data are:
management,high,level0,D2unemployed,high,level0,D1technician,high,level0,D2management,high,level0,D2
Scanning binned data to compute joint and dependent counts
Index was outside the bounds of the array.
Test Dataset:
sno id job marital education default loan_in_other conta
5 5 blue-collar married secondary yes yes unkno
6 6 management single tertiary no no cellula
7 7 self-employed married tertiary no no cellula
8 8 technician married secondary no no cellula
9 9 entrepreneur married tertiary no no unkno

Figure 9.10

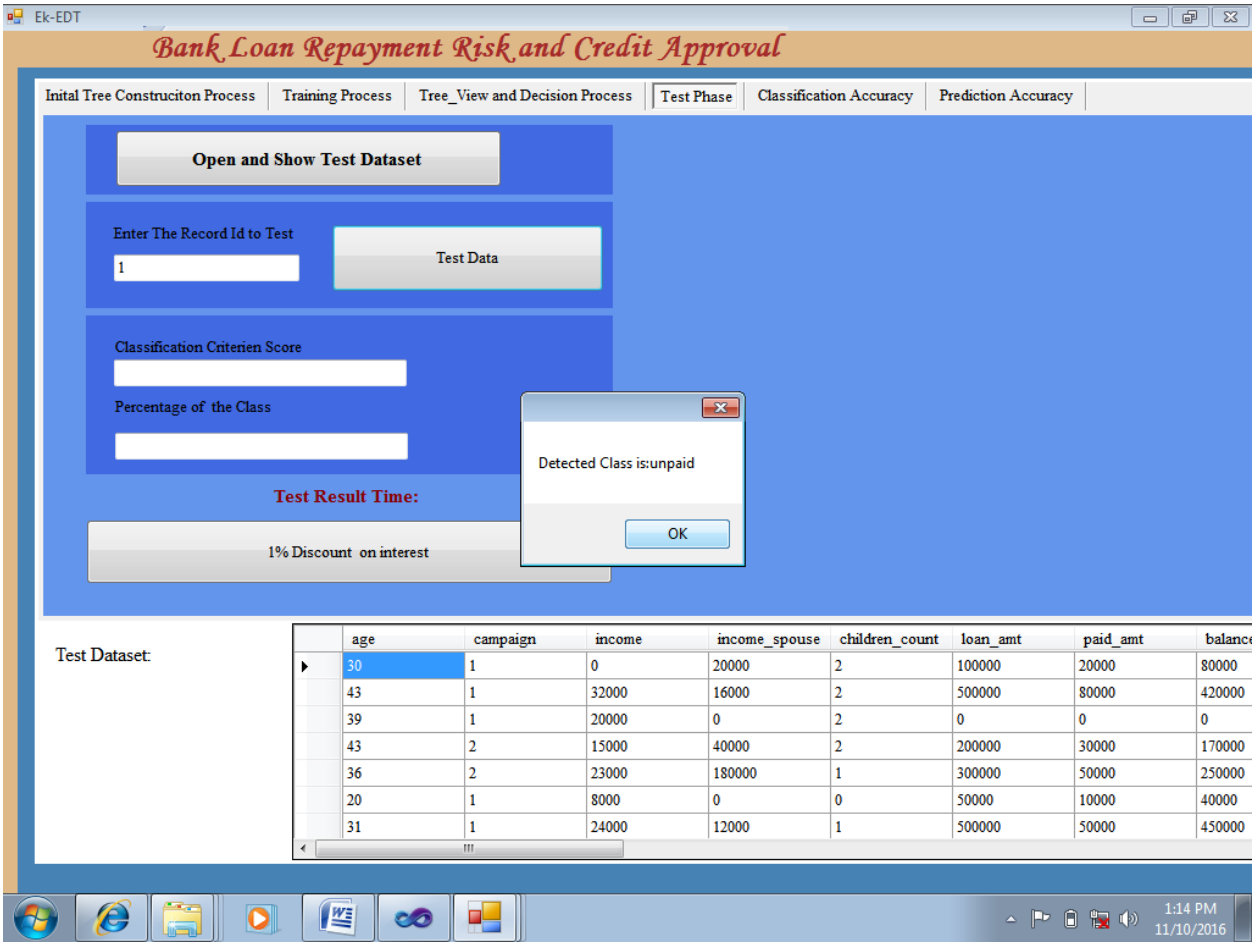

Ek-EDT
Bank Loan Repayment Risk and Credit Approval
Inital Tree Construciton Process
Training Process
Tree_View and Decision Process
Test Phase
Classification Accuracy
Prediction Accuracy
Open and Show Test Dataset
Enter The Record Id to Test
1
Test Data
Classification Criterien Score
Percentage of the Class
Test Result Time:
1% Discount on interest
Detected Class is:unpaid
OK
Test Dataset:
age campaign income income_spouse children_count loan_amt paid_amt balanc
30 1 0 20000 2 100000 20000 80000
43 1 32000 16000 2 500000 80000 420000
39 1 20000 0 2 0 0 0
43 2 15000 40000 2 200000 30000 170000
36 2 23000 180000 1 300000 50000 250000
20 1 8000 0 0 50000 10000 40000
31 1 24000 12000 1 500000 50000 450000
1:14 PM
11/10/2016

Figure 9.11

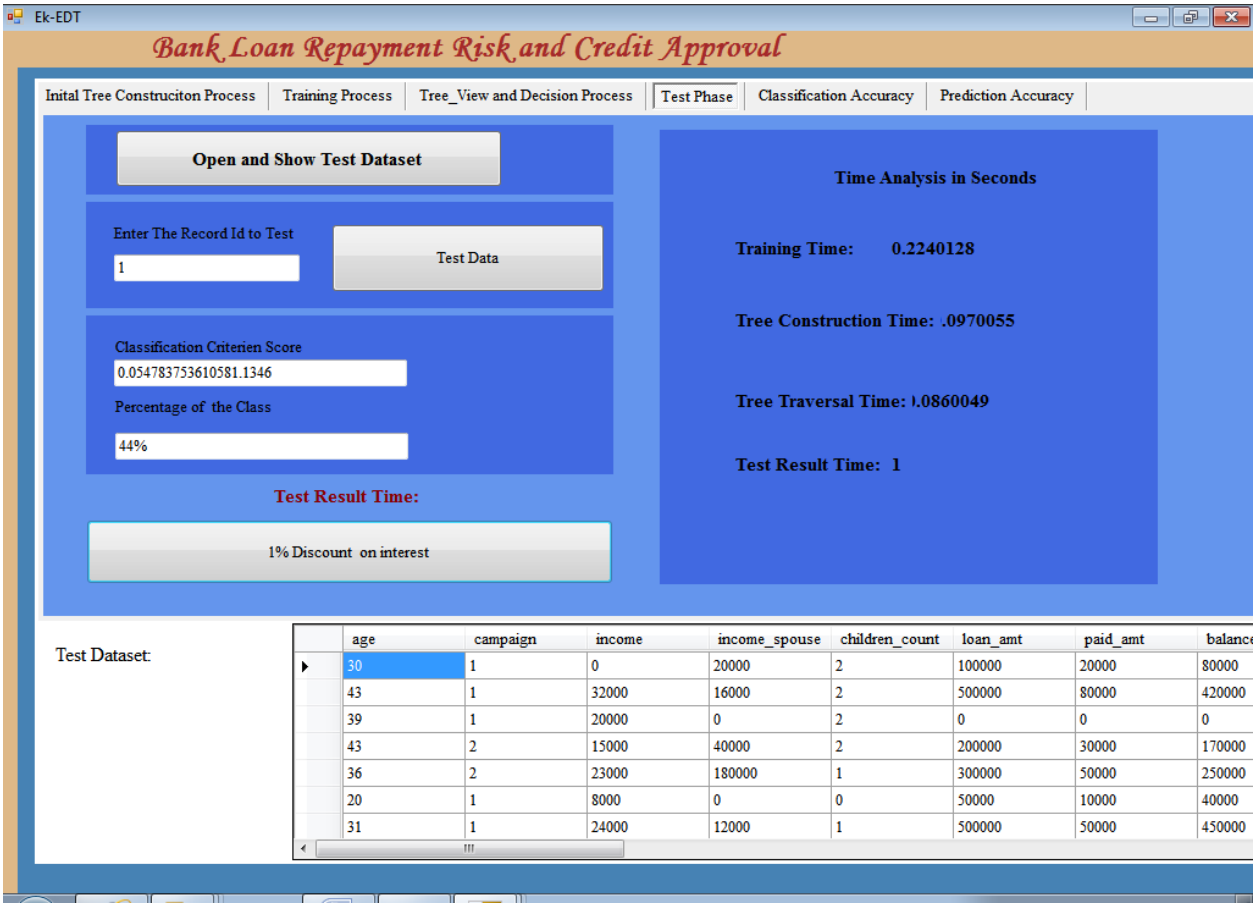

Ek-EDT
Bank Loan Repayment Risk and Credit Approval
Inital Tree Construciton Process
Training Process
Tree_View and Decision Process
Test Phase
Classification Accuracy
Prediction Accuracy
Open and Show Test Dataset
Enter The Record Id to Test
1
Test Data
Classification Criterien Score
0.0547837536105 81.1346
Percentage of the Class
44%
Test Result Time:
1% Discount on interest
Time Analysis in Seconds
Training Time: 0.2240128
Tree Construction Time: .0970055
Tree Traversal Time: .0860049
Test Result Time: 1
Test Dataset:
age campaign income income_spouse children_count loan_amt paid_amt balanc
30 1 0 20000 2 100000 20000 80000
43 1 32000 16000 2 500000 80000 420000
39 1 20000 0 2 0 0 0
43 2 15000 40000 2 200000 30000 170000
36 2 23000 180000 1 300000 50000 250000
20 1 8000 0 0 50000 10000 40000
31 1 24000 12000 1 500000 50000 450000

Figure 9.12

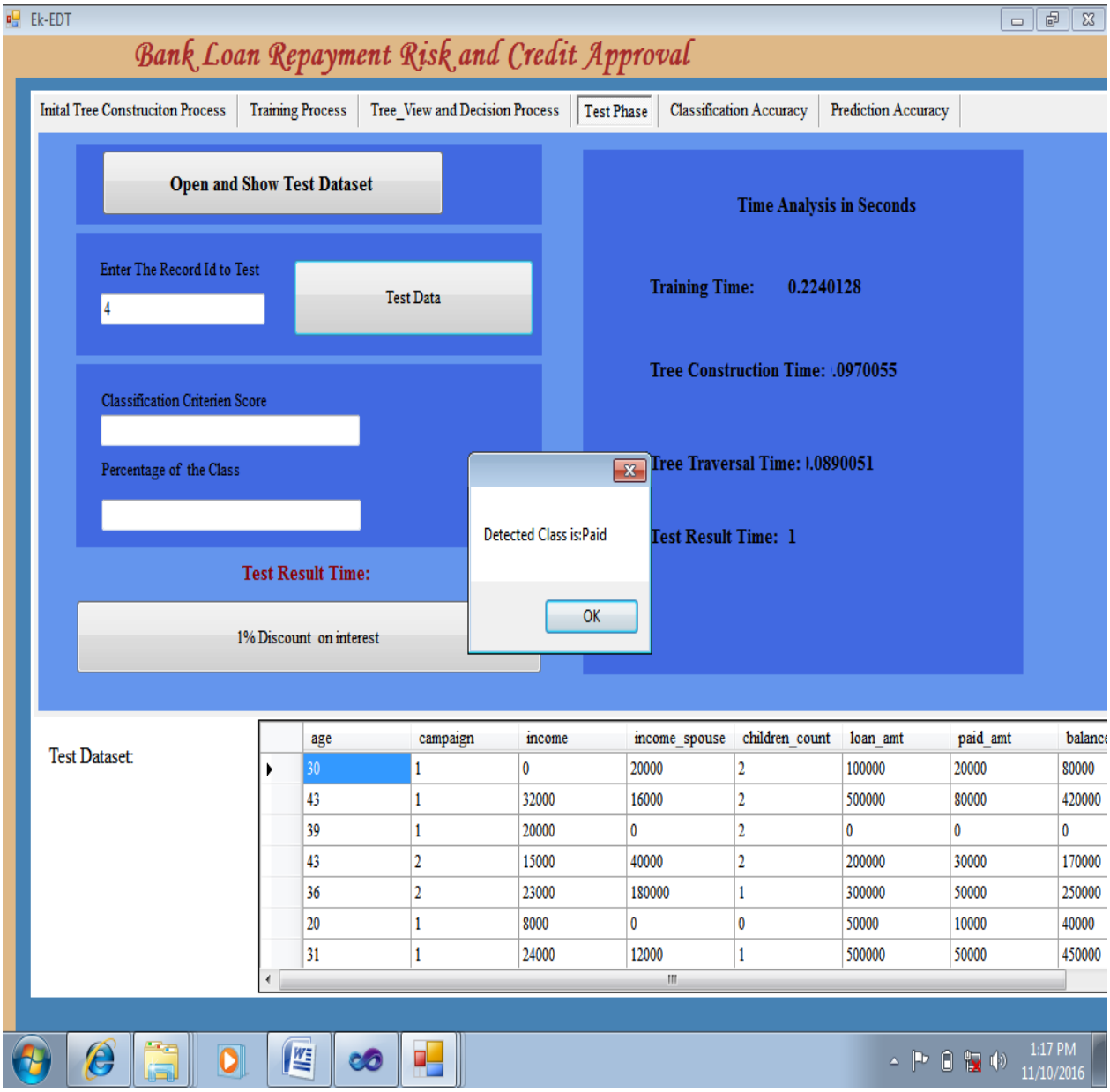

Ek-EDT
Bank Loan Repayment Risk and Credit Approval
Inital Tree Construciton Process
Training Process
Tree_View and Decision Process
Test Phase
Classification Accuracy
Prediction Accuracy
Open and Show Test Dataset
Enter The Record Id to Test
4
Test Data
Classification Criterien Score
Percentage of the Class
Test Result Time:
1% Discount on interest
Time Analysis in Seconds
Training Time: 0.2240128
Tree Construction Time: .0970055
Tree Traversal Time: .0890051
Test Result Time: 1
Detected Class is:Paid
OK
Test Dataset:
age campaign income income_spouse children_count loan_amt paid_amt balance
30 1 0 20000 2 100000 20000 80000
43 1 32000 16000 2 500000 80000 420000
39 1 20000 0 2 0 0 0
43 2 15000 40000 2 200000 30000 170000
36 2 23000 180000 1 300000 50000 250000
20 1 8000 0 0 50000 10000 40000
31 1 24000 12000 1 500000 50000 450000
1:17 PM
11/10/2016

Figure 9.13

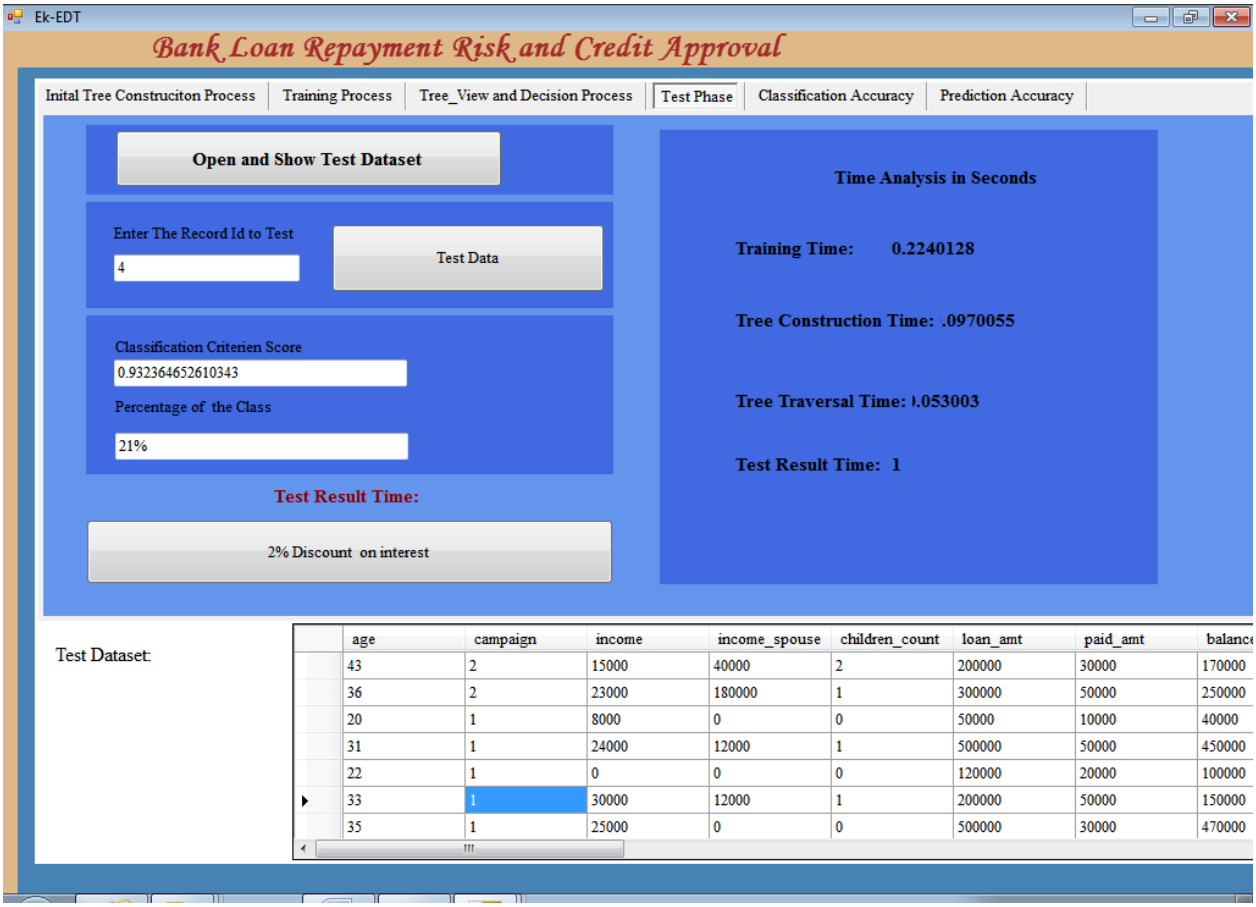


Figure 9.14

Test Phases Results

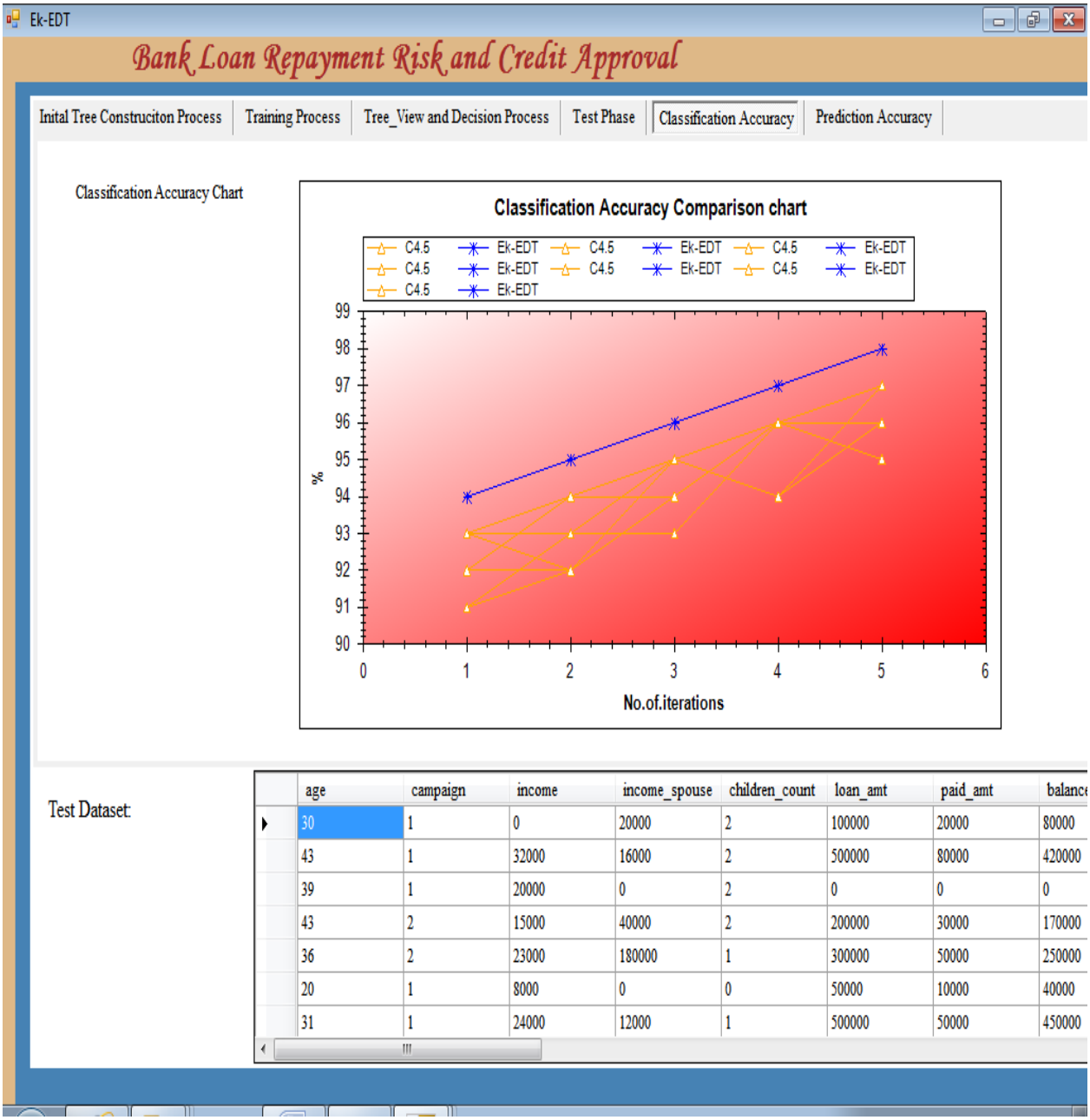
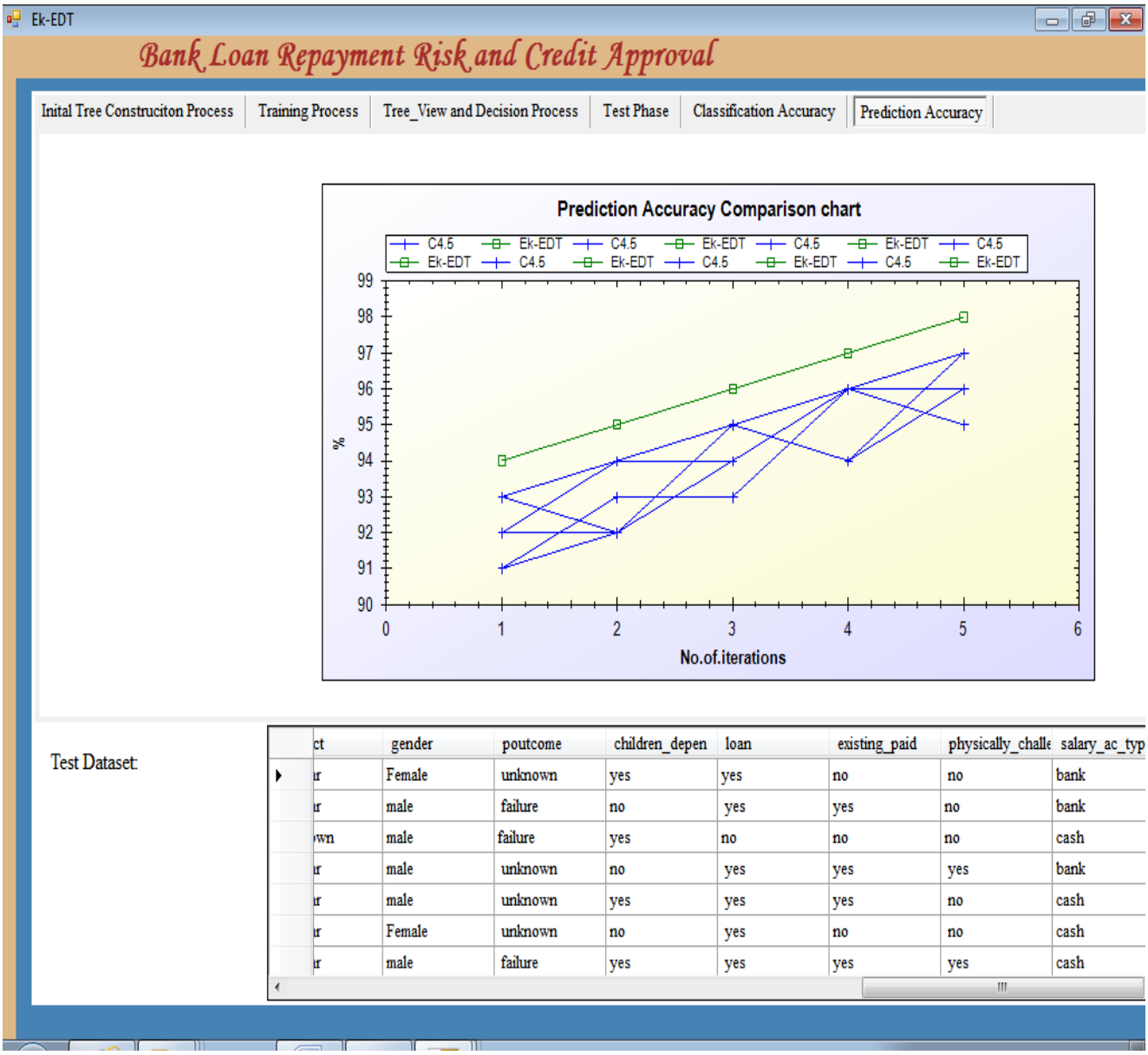


Figure 9.15

Classification Accuracy Comparison Chart



Prediction Accuracy Comparison Chart

REFERENCES

[1] Albayrak A. S., Yılmaz Ş.K., "Data mining: Decision tree algorithms and an application on data of IMKB", Süleyman Demirel University The Journal of Faculty of Economics and Administrative Sciences, vol. 14, pp. 31-52, 2009 .

[2] Aşan Z., "Examining the socioeconomic characteristics of customers using credit cards, with clustering analysis", Dumlupınar University The Journal of Social Sciences, vol.17, pp. 256-267, 2007.

[3] Atbaş A. C., A study on determining the cluster number in clustering analysis, Master Thesis, Ankara University, Graduate School of Natural Sciences, 2008.

[4] Bilen H., Data mining application for personnel selection and performance evaluation in banking sector, Master Thesis, Gazi University, Graduate School of Natural and Applied Sciences , 2009.

[5] Chien C.-F., Chen L.-F., "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry", Expert Systems with Applications,vol. 34, pp. 280-290, 2008.

[6] Ching W. K., Pong M. K.., Advances in data mining and modeling, 1st ed., World Scientific, Hong Kong, China, 2002.

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

[8] Bharati M. Ramageri, "DATA MINING TECHNIQUES AND APPLICATIONS", Indian Journal of Computer Science and Engineering Vol. 1 No. 4

[9] Vivek Bhambri "Application of Data Mining in Banking Sector", International Journal of Computer Science and Technology Vol. 2, Issue 2, June 2011

[10] P.Sundari, Dr.K.Thangadurai "An Empirical Study on Data Mining Applications", Global Journal of Computer Science and Technology, Vol. 10 Issue 5 Ver. 1.0 July 2010.