

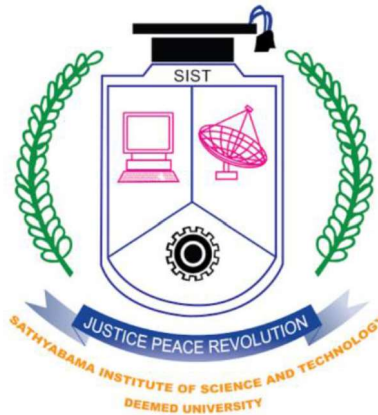
LOAN FRAUD DETECTION USING MACHINE LEARNING ALGORITHM

Submitted in partial fulfillment of the requirements for
the award of
Bachelor of Engineering degree in Computer Science and Engineering

by

PUTLA SUBHASH CHANDRA

(Reg No. 38110563)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

**Accredited with Grade "A" by NAAC
JEPPIAAR NAGAR, RAJIV GANDHI
SALAI, CHENNAI 600 119**

MARCH - 2022



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with "A" grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this project report is the bonafide work of **PUTLA SUBHASH CHANDRA(Reg.No. 38110563)**, project who carried out the project entitled "**LOAN FRAUD DETECTION USING MACHINE LEARNING ALGORITHM**" under my supervision from November 2021 to March 2022.

Internal Guide

Dr. CM. SUJA, M.E., Ph.D.,

Head of the Department

Dr. S. Vigneswari, M.E., Ph.D.,

Submitted for Viva voce Examination held on_____

Internal Examiner

External Examiner

DECLARATION

I **PUTLA SUBHASH CHANDRA(38110563)** hereby declare that the Project Report entitled “**LOAN FRAUD DETECTION USING MACHINE LEARNING ALGORITHM**” is done by us under the guidance of **Dr CM. SUJA, ME.,Ph.D.**, Department of Computer Science and Engineering at Sathyabama Institute of Science and Technology is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering.

DATE:

PLACE: CHENNAI

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala, M.E., Ph.D., Dean**, School of Computing, **Dr. S. Vigneswari, M.E., Ph.D.,** and **Dr. L. Lakshmanan, M.E., Ph.D., Heads of the Department** of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. CM. SUJA, M.E., Ph.D.,** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

Abstract

The banking sector is a very important sector in our present-day generation where almost every human has to deal with the bank either physically or online. In dealing with the banks, the customers and the banks face the chances of being trapped by fraudsters. Examples of fraud include insurance fraud, credit card fraud, accounting fraud, etc. Detection of fraudulent activity is thus critical to control these costs. This paper hereby addresses bank fraud detection via the use of machine learning techniques; association, clustering, forecasting, and classification to analyze the customer data in order to identify the patterns that can lead to frauds. Upon identification of the patterns, adding a higher level of verification/authentication to banking processes can be added.

TABLE OF CONTENTS

CHAPTER No	TITLE	PAGE No
	ABSTRACT	Vi
	LIST OF FIGURES	Viii
1.	INTRODUCTION	1
	1.1 OVERVIEW	1
2.	LITERATURE SURVEY	3
3.	SYSTEM DESIGN	7
	3.1 SCOPE OF THE PROJECT	7
	3.2 OBJECTIVE	7
	3.3 PROBLEM DEFINITION	7
	3.4 EXISTING SYSTEM	7
	3.5 PROPOSED SYSTEM	8
4.	METHODS & ALGORITHMS USED	10
	4.1 HARDWARE REQUIREMENTS	10
	4.2 SOFTWARE REQUIREMENTS	10
	4.3 SYSTEM ARCHITECTURE	17
	4.4 MODULE OVERVIEW	17
5.	SYSTEM IMPLEMENTATION	21
6.	CONCLUSION	35

LIST OF FIGURES

FIG No	FIGURE NAME	PAGE No
4.1	JAVA COMPILATION	12
4.2	COLLECTION FRAME WORK	14
4.3	SYSTEM ARCHITECTURE	17
4.4	DECISION TREE DATA-FLOW	18
4.5	CLUSTERING DATA FLOW DIAGRAM	19
5.1	DATA FLOW DIAGRAM 1	21
5.2	DATA FLOW DIAGRAM 2	22
5.3	DATA FLOW DIAGRAM 3	23
5.4	USE CASE DIAGRAM	24
5.5	SEQUENCE DIAGRAM	25
5.6	ACTIVITY DIAGRAM	26
6.1	USER LOGIN PAGE	36
6.2	ADMINISTRATOR LOGIN PAGE	36
6.3	LOGIN DETAILS VERIFICATION	37
6.4	LOAN REQUESTS	37
6.5	PAST-LOAN DETAILS	38
6.6	PERSON DETAILS	38
6.7	FRAUD HISTORY	39

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

Fraud detection is a set of activities undertaken to prevent money or property from being obtained through false pretenses. Fraud detection is applied to many industries such as banking or insurance. In banking, fraud may include forging checks or using stolen credit cards. Other forms of fraud may involve exaggerating losses or causing an accident with the sole intent for the payout. With an unlimited and rising number of ways someone can commit fraud, detection can be difficult. Activities such as reorganization, downsizing, moving to new information systems or encountering a cybersecurity breach could weaken an organization's ability to detect fraud. Techniques such as real-time monitoring for fraud is recommended. Organizations should look for fraud in financial transactions, locations, devices used, initiated sessions and authentication systems.

Fraud typically involves multiple repeated methods, making searching for patterns a general focus for fraud detection. For example, data analysts can prevent insurance fraud by making algorithms to detect patterns and anomalies. Fraud detection can be separated by the use of statistical data analysis techniques or artificial intelligence (AI).

Statistical data analysis techniques include:

- calculating statistical parameters
- regression analysis
- probability distributions and models
- data matching

AI techniques used to detect fraud include:

- Data mining classifies, groups and segments data to search through millions of transactions to find patterns and detect fraud.

- Neural networks learn suspicious-looking patterns and use those patterns to detect them further.
- Machine learning automatically identifies characteristics found in fraud.
- Pattern recognition detects classes, clusters and patterns of suspicious behavior.

According to The American Heritage dictionary, second college edition, fraud is defined as a deception deliberately practiced in order to secure unfair unlawful gain. Fraud detection is the recognition of symptoms of fraud where no prior suspicion or tendency to fraud exists. Examples include insurance fraud, credit card fraud and accounting fraud. Data from the Nigeria Inter-Bank Settlement System (NIBSS) has revealed that fraudulent transactions in the banking sector at its peak. Fraud has evolved from being committed by casual fraudsters to being committed by organized crime and fraud rings that use sophisticated methods to take over control of accounts and commit fraud. Some 6.8 million Americans were victimized by card fraud in 2007, according to Javelin research. The Nilson Report estimates the cost to the industry to be \$4.84 billion. Javelin estimates the losses at more than six times that amount – some \$30.6 billion in 2007. Of course, fraud is not a domestic product as it's everywhere. For instance, card fraud losses cost UK economy GBP 423 million in 2006. Credit card fraud accounts for the biggest cut of the \$600 million that airlines lose each year globally.

The banking sector is a very important sector in our present-day generation where almost every human has to deal with the bank either physically or online. In dealing with the banks, the customers and the banks face the chances of been trapped by fraudsters. Examples of fraud include insurance fraud, credit card fraud, accounting fraud, etc. Detection of fraudulent activity is thus critical to control these costs. This project hereby addresses bank fraud detection via the use of machine learning techniques; association, clustering, forecasting, and classification to analyze the customer data in order to identify the patterns that can lead to frauds. Upon identification of the patterns, adding a higher level of verification/authentication to banking processes can be added.

CHAPTER 2

LITERATURE SURVEY

2. LITERATURE SURVEY:

Fraud detection has been usually seen as a data mining problem where the objective is to correctly classify the transactions as legitimate or fraudulent. For classification problems many performance measures are defined most of which are related with correct number of cases classified correctly.

A more appropriate measure is needed due to the inherent structure of credit card transactions. When a card is copied or stolen or lost and captured by fraudsters it is usually used until its available limit is depleted. Thus, rather than the number of correctly classified transactions, a solution which minimizes the total available limit on cards subject to fraud is more prominent.

Since the fraud detection problem has mostly been defined as a classification problem, in addition to some statistical approaches many data mining algorithms have been proposed to solve it. Among these, decision trees and artificial neural networks are the most popular ones. The study of Bolton and Hand provides a good summary of literature on fraud detection problems.

However, when the problem is approached as a classification problem with variable misclassification costs as discussed above, the classical data mining algorithms are not directly applicable; either some modifications should be made on them or new algorithms developed specifically for this purpose are needed. An alternative approach could be trying to make use of general purpose meta heuristic approaches like genetic algorithms.

Hussein A Abdou, Ahmed A. El-Masry in the paper on **"Neural nets versus conventional techniques in credit scoring in Egyptian banking"** found

that the number of Non-Performing Loans has increased in recent years, paralleling the current financial crisis, thus increasing the importance of credit scoring models. This study proposes a three stage hybrid Adaptive Neuro Fuzzy Inference System credit scoring model, which is based on statistical techniques and Neuro Fuzzy. The proposed model's performance was compared with conventional and commonly utilized models. The credit scoring models are tested using a 10-fold cross-validation process with the credit card data of an international bank operating in Turkey. Results demonstrate that the proposed model consistently performs better than the Linear Discriminant Analysis, Logistic Regression Analysis, and Artificial Neural Network (ANN) approaches, in terms of average correct classification rate and estimated misclassification cost. As with ANN, the proposed model has learning ability; unlike ANN, the model does not stay in a black box. In the proposed model, the interpretation of independent variables may provide valuable information for bankers and consumers, especially in the explanation of why credit applications are rejected.

Thi Huyen Thanh Dinha Stefanie Kleimeier in the paper on **A credit scoring model for Vietnams retail banking market**. As banking markets in developing countries are maturing, banks face competition not only from other domestic banks but also from sophisticated foreign banks. Given the substantial growth of consumer credit and increased regulatory attention to risk management, the development of a well-functioning credit assessment framework is essential. As part of such a framework, we propose a credit scoring model for Vietnamese retail loans. First, we show how to identify those borrower characteristics that should be part of a credit scoring model. Second, we illustrate how such a model can be calibrated to achieve the strategic objectives of the bank. Finally, we assess the use of credit scoring models in the context of transactional versus relationship lending.

D.J Hand and W.E Henley in the paper on **Statistical classification methods in consumer credit scoring**. Credit scoring is the term used to describe formal statistical methods used for classifying applicants for credit into 'good' and 'bad' risk classes. Such methods have become increasingly important with the

dramatic growth in consumer credit in recent years. A wide range of statistical methods has been applied, though the literature available to the public is limited for reasons of commercial confidentiality. Particular problems arising in the credit scoring context are examined and the statistical methods which have been applied are reviewed.

Vijay S.Desai, Jonathan N.Crook, George A.Overstreet Jr in the paper on **A comparison of neural networks and linear scoring models in the credit union environment** said that the purpose of the present paper is to explore the ability of neural networks such as multilayer perceptrons and modular neural networks, and traditional techniques such as linear discriminant analysis and logistic regression, in building credit scoring models in the credit union environment. Also, since funding and small sample size often preclude the use of customized credit scoring models at small credit unions, we investigate the performance of generic models and compare them with customized models. Our results indicate that customized neural networks offer a very promising avenue if the measure of performance is percentage of bad loans correctly classified. However, if the measure of performance is percentage of good and bad loans correctly classified, logistic regression models are comparable to the neural networks approach. The performance of generic models was not as good as the customized models, particularly when it came to correctly classifying bad loans. Although we found significant differences in the results for the three credit unions, our modular neural network could not accommodate these differences, indicating that more innovative architectures might be necessary for building effective generic models.

M Vojtek, E Koèenda in the paper on **Credit Scoring Methods. Czech Journal of Economics and Finance**. The paper reviews the best-developed and most frequently applied methods of credit scoring employed by commercial banks when evaluating loan applications. The authors concentrate on retail loans – applied research in this segment is limited, though there has been a sharp increase in the volume of loans to retail clients in recent years. Logit analysis is identified as the most frequent credit-scoring method used by banks. However, other nonparametric methods are widespread in terms of pattern recognition. The

methods reviewed have potential for application in post-transition countries.

Lyn C.Thomas in the paper on **A survey of credit and behavioral scoring: forecasting: financial risk of lending to customers**. Credit scoring and behavioural scoring are the techniques that help organisations decide whether or not to grant credit to consumers who apply to them. This article surveys the techniques used — both statistical and operational research based — to support these decisions. It also discusses the need to incorporate economic conditions into the scoring systems and the way the systems could change from estimating the probability of a consumer defaulting to estimating the profit a consumer will bring to the lending organisation — two of the major developments being attempted in the area. It points out how successful has been this under-researched area of forecasting financial risk. 2000 Elsevier Science B.V. All rights reserved.

CHAPTER 3

SYSTEM DESIGN

3.1 SCOPE OF THE PROJECT

The main objective of this project is to help banking sector to improve its security from fraudsters. Now a days, fraudsters were finding different ways to do frauds, this project helps to minimize some of those fraud techniques.

3.2 OBJECTIVE

The main objective of this project is to help banking sector to improve its security from fraudsters.

3.3 PROBLEM DEFINATION

The banking sector is a very important sector in our present-day generation where almost every human has to deal with the bank either physically or online. In dealing with the banks, the customers and the banks face the chances of been trapped by fraudsters.

3.4 EXISTING SYSTEM

- In case of bank fraud detection, the existing system is detect the fraud after fraud has been happen. Existing system maintain the large amount of data when customer comes to know about inconsistency in transaction he/she made complaint and then fraud detection system start it working. It first tries to detect that fraud has actually occur after that it transactions that was used to fraud detection mechanism developed by master and visa cards
- A machine learning paradigm classification, with Bank Fraud Detection being the base.

- Intrusion detections to track fraud location and so on. In case of existing system there is no confirmation of recovery of fraud and Customer satisfaction.
- Secure electronic system used to analyze the behaviour of legitimate users.
- Data Mining mechanisms to classify and preprocess the user's data.
- Genetic algorithms.

3.4.1 DISADVANTAGES OF EXISTING SYSTEM

- Each payment system has its limits regarding the maximum amount in the account, the number of transactions per day and the amount of output.
- If Internet connection fails, you can not get to your online account.
- If you follow the security rules the threat is minimal. The worse situation when the system of processing company has been broken, because it leads to the leak of personal data on cards and its owners.
- The information about all the transactions, including the amount, time and recipient are stored in the database of the payment system. And it means the intelligence agency has an access to this information. Sometimes this is the path for fraudulent activities.

3.5 PROPOSED SYSTEM

In proposed methodology, Detection of fraudulent activity is thus critical to control these costs. This paper hereby addresses bank fraud detection via the use of machine learning techniques; association, clustering, forecasting, and classification to analyze the customer data in order to identify the patterns that can lead to frauds. Upon identification of the patterns, adding a higher level of verification/authentication to banking processes can be added. These kinds of frauds can be credit card fraud, insurance fraud, accounting fraud, etc. which may lead to the financial loss to the bank or the customers. Thus, detection of these

kinds of frauds are very important. Fraud detection in banking sector is based on the machine learning techniques and their collective analysis from the past experiences and the probability of how the fraudsters can steal from customers and banks. Therefore this project addresses the analysis of data mining techniques of how to detect frauds and overcoming it in banking sector.

ADVANTAGES OF PROPOSED SYSTEM

- To eliminate real time fraud to the lowest level.
- To increase the confidence of customers in the banking system especially for online transactions.
- To discourage fraudsters (both present and intending ones)

CHAPTER 4

METHODS AND ALGORITHMS USED

4.1 HARDWARE REQUIREMENT

The hardware requirements may serve as the basis for a contract for the implementation of the system and should therefore be a complete and consistent specification of the whole system. They are used by software engineers as the starting point for the system design. It shows what the system does and it also shows how it should be implemented. It specifies the speed of the system that should be used in this project. The hardware requirement for this project is mentioned below.

System	: Pentium IV 2.4 GHz
Hard Disk	: 40 GB
Floppy Drive	: 44 Mb
Monitor	: 15 VGA Colour
Ram	: 8 GB

4.2 SOFTWARE REQUIREMENT

The software requirements document is the specification of the system. It should include both a definition and a specification of requirements. It is a set of what the system should do rather than how it should do it. The software requirements provide a basis for creating the software requirements specification. It is useful in estimating cost, planning team activities, performing tasks and tracking the team and tracking the team's progress throughout the development activity. The software requirement specifies the application software that is being

used in the project. The software needed to develop this project is mentioned below.

Operating system	: Windows 10
IDE	: Visual Studio
Coding Language	: Java

INTRODUCTION

Java is one of the world's most important and widely used computer languages, and it has held this distinction for many years. Unlike some other computer languages whose influence has wearied with passage of time, while Java's has grown.

APPLICATION OF JAVA

Java is widely used in every corner of world and of human life. Java is not only used in software's but is also widely used in designing hardware controlling software components. There are more than 930 million JRE downloads each year and 3 billion mobile phones run java.

Following are some other usage of Java:

- Developing Desktop Applications
- Web Applications like LinkedIn.com, Snapdeal.com etc
- Mobile Operating System like Android
- Embedded Systems
- Robotics and games etc.

FEATURES OF JAVA

The prime reason behind creation of Java was to bring portability and security feature into a computer language. Beside these two major features, there were many other features that played an important role in moulding out the final form of this outstanding language. Those features are;

- *Simple*

Java is easy to learn and its syntax is quite simple, clean and easy to understand. The confusing and ambiguous concepts of C++ are either left out in Java or they have been re-implemented in a cleaner way.

Eg: Pointers and Operator Overloading are not there in java but were an important part of C++.

- *Object-Oriented*

In java everything is Object which has some data and behaviour. Java can be easily extended as it is based on Object Model.

3) Robust

Java makes an effort to eliminate error prone codes by emphasizing mainly on compile time error checking and runtime checking. But the main areas which Java improved were Memory Management and mishandled Exceptions by introducing automatic Garbage Collector and Exception Handling.

4) Platform Independent

Unlike other programming languages such as C, C++ etc. which are compiled into platform specific machines. Java is guaranteed to be write-once, run-anywhere language.

On compilation Java program is compiled into byte code. This byte code is platform independent and can be run on any machine, plus this byte code format also provide security. Any machine with Java Runtime Environment can run Java Programs.

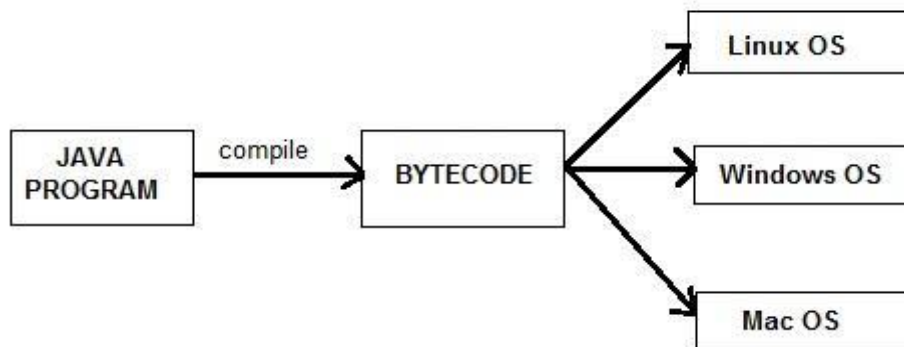


Fig no:4.1: Java Compilation

5) SECURE

When it comes to security, Java is always the first choice. With java secure features it enable us to develop virus free, temper free system. Java program always runs in Java runtime environment with almost null interaction with system OS, hence it is more secure.

6) MULTI-THREADING

Java multithreading feature makes it possible to write program that can do many tasks simultaneously. Benefit of multithreading is that it utilizes same memory and other resources to execute multiple threads at the same time, like While typing, grammatical errors are checked along.

7) ARCHITECTURAL NEUTRAL

Compiler generates byte codes, which have nothing to do with a particular computer architecture, hence a Java program is easy to interpret on any machine.

- **PORTABLE**

Java Byte code can be carried to any platform. No implementation dependent features. Everything related to storage is predefined, example: size of primitive data types.

- *HIGH PERFORMANCE*

Java is an interpreted language, so it will never be as fast as a compiled language like C or C++. But, Java enables high performance with the use of just-in-time compiler.

- *COLLECTION FRAMEWORK*

The collection framework was not part of original Java release. Collections was added to J2SE 1.2. Prior to Java 2, Java provided adhoc classes such as Dictionary, Vector, Stack and Properties to store and manipulate groups of objects. Collection framework provides many important classes and interfaces to collect and organize group of alike objects.

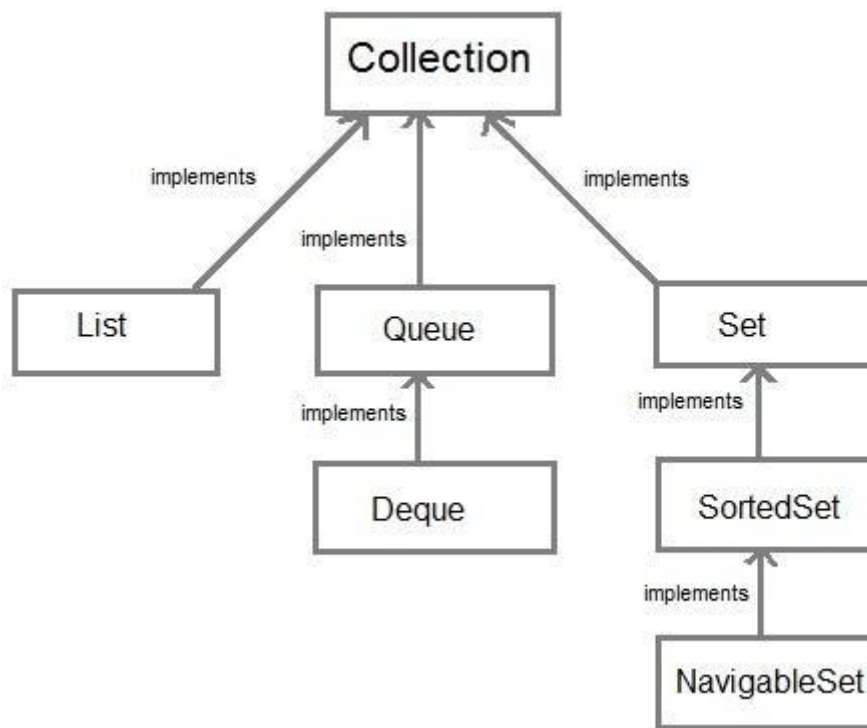


Fig no: 4.2: Collection Framework

MYSQL

MySQL, officially, but also called "My Sequel" is the world's most widely used open-source relational database management system (RDBMS) that runs as a server providing multi-user access to a number of databases, though SQLite probably has more total embedded deployments. The SQL phrase stands for Structured Query Language.

The MySQL development project has made its source code available under the terms of the GNU General Public License, as well as under a variety of proprietary agreements. MySQL was owned and sponsored by a single for-profit firm, the Swedish company MySQL AB, now owned by Oracle Corporation. MySQL is a popular choice of database for use in web applications, and is a central component of the widely used LAMP open source web application software stack (and other 'AMP' stacks).

LAMP is an acronym for "Linux, Apache, MySQL, Perl/PHP/Python." Freesoftware-open source projects that require a full-featured database management system often use MySQL.

For commercial use, several paid editions are available, and offer additional

functionality. Applications which use MySQL databases include: TYPO3, MODx, Joomla, WordPress, phpBB, MyBB, Drupal and other software. MySQL is also used in many high-profile, large-scale websites, including Wikipedia, Google (though not for searches), Facebook, Twitter, Flickr and YouTube.

ECLIPSE FOR JAVA

Eclipse IDE (integrated development environment) for Java has been the leading development environment with a market share of 65% as of today. It can be extended with an additional software component. Eclipse calls these software components as plugins, which can be grouped into features. Many companies have extended Eclipse IDE on top of the Eclipse framework. It is also available as IDE for other languages.

HOW TO INSTALL ECLIPSE IDE FOR JAVA

The Eclipse IDE for Java is designed for standard Java development. It contains typically required packages like support for Maven and Gradle. It also supports Git version controlling system. You can download this IDE from the official Eclipse website.

BASIC CONCEPTS

- **Workspace:** This is at a physical location of file path where you store certain meta data of development artefacts. Projects, source files, images, and other artefacts are stored either inside or outside of your workspace.
- **Views and editors:** Eclipse offers views and editors to navigate and change content. A view is used to work on a set of data, which is in a hierarchical structure. If data is changed via view, the underlying data is directly changed.
- **Eclipse project:** This is an open source, configuration and binary file that is related to a certain task. Eclipse project groups these files into a buildable and reusable units. These projects can have natures assigned to it which describes describes the purpose of this project.

STEPS TO INSTALL

1. You need to install the Java Development Kit (JDK) to setup the environment

java programming.

2. Download the latest version of Eclipse from the official Eclipse website. You can find the latest version under 'Get Eclipse IDE 2018-12'.

3. To install the Eclipse, simply unzip the downloaded file and execute the installable package.

Apache Maven is one such importance build automation tool used in various Java projects. Maven is one of the most powerful project management tools based on POM (project object model) as it makes it easier to build processes like ANT and used by Java developers to make day-to-day work easier.

The development of Maven tool started to simplify the build processes in the Jakarta Turbine project. The tool can be used for building and managing any Java project. While it is known for simplifying ANT processes, it has more advanced capabilities than just ANT, a Java library and command-line tool.

STEPS TO INSTALL MAVEN

1. Check if your system has Java installed. If not, then install the latest version of Java Environment.

2. You need to ensure that Java Environmental variable is set.

3. Download Maven from the the official website of Apache Maven.

4. Create the bin directory of of the created apache-maven-3.5.3 directory to the PATH environment variable and system variable.

5. Open cmd and run mvn -v command and the installation will get completed.

4.3 SYSTEM ARCHITECTURE

The system architecture establishes the basic structure of the system, defining the essential core design features and elements that provide the framework for the system. The systems architecture provides the architects view of the users' vision for what the system needs to be and do, and the paths along which it must be able to evolve and strives to maintain the integrity of that vision as it evolves during detailed design and implementation.

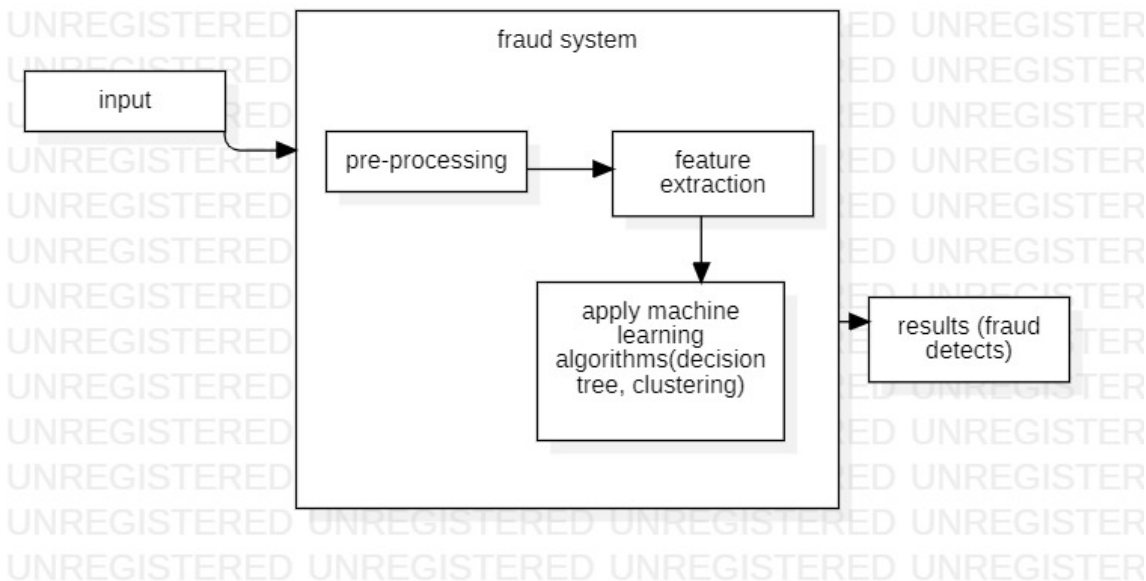


Fig no:4.3: System Architecture

MODULES

- Classification
- Clustering
- Association Rule
- Fraud Detection

4.4 MODULE OVERVIEW

CLASSIFICATION

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. The data classification process involves learning and classification.

DECISION TREE

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision

rules and each leaf node represents the outcome. In, a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

The decisions or the test are performed on the basis of features of the given dataset is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It, is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

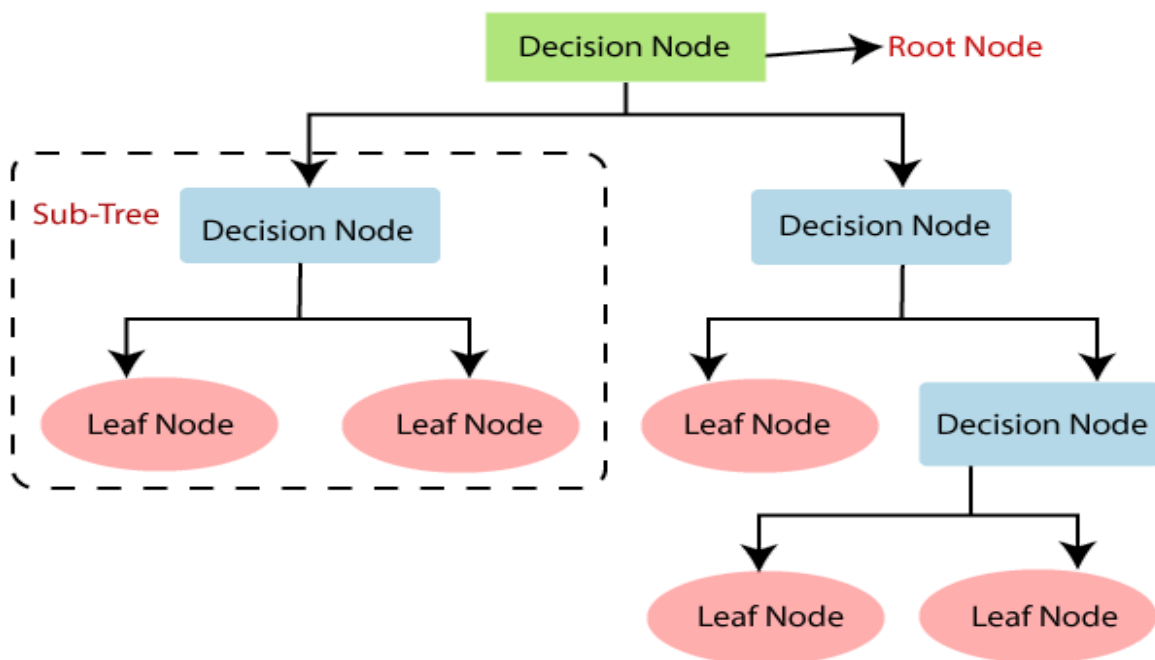


Fig no:4.4: Decision Tree Data-Flow

CLUSTERING

Clustering or cluster analysis is a machine learning technique, which groups the unlabelled dataset. It can be defined as "A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group." It does it by finding some similar patterns in the unlabelled dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns.

It is an unsupervised learning method, hence no supervision is provided to the algorithm, and it deals with the unlabeled dataset. After applying this clustering technique, each cluster or group is provided with a cluster-ID. ML system can use this id to simplify the processing of large and complex datasets.

Example: Let's understand the clustering technique with the real-world example of Mall: When we visit any shopping mall, we can observe that the things with similar usage are grouped together. Such as the t-shirts are grouped in one section, and trousers are at other sections, similarly, at vegetable sections, apples, bananas, Mangoes, etc., are grouped in separate sections, so that we can easily find out the things. The clustering technique also works in the same way. Other examples of clustering are grouping documents according to the topic.

In the process of clustering, all the different bank transactions are grouped together into one cluster. We can say that the approach of pre-processing for classification and attribute set choice is used in clustering.

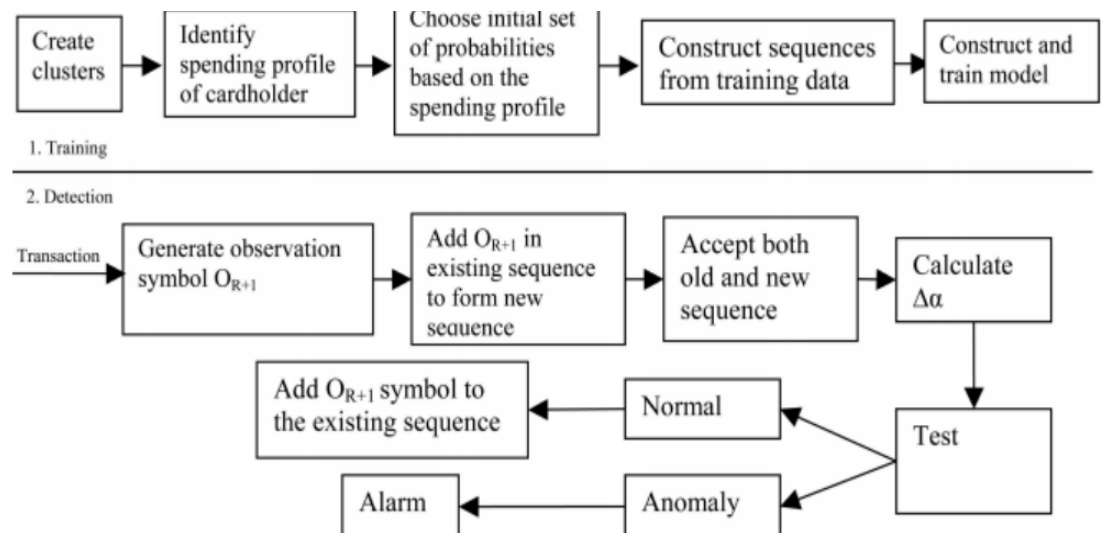


Fig no: 4.5: Clustering data flow diagram

ASSOCIATION RULE

The central task of association rule mining is to find sets of binary variables that co-occur together frequently in a transaction database, while the goal of feature selection problem is to identify groups that are strongly correlated with each other with a specific target variable.

FRAUD DETECTION

Another popular area where data mining can be used in the banking industry is in fraud detection. Being able to detect fraudulent actions is an increasing concern for many businesses; and with the help of data mining more fraudulent actions are being detected and reported.

Chapter - 5

SYSTEM IMPLEMENTATION

DATA FLOW DIAGRAM

Data Flow Diagram (DFD) is a two-dimensional diagram that describes how data is processed and transmitted in a system. The graphical depiction recognizes each source of data and how it interacts with other data sources to reach a mutual

output. In order to draft a data flow diagram one must

- Identify external inputs and outputs
- Determine how the inputs and outputs relate to each other
- Explain with graphics how these connections relate and what they result in.

Role of DFD:

It is a documentation support which is understood by both programmers and nonprogrammers. As DFD postulates only what processes are accomplished not how they are performed. A physical DFD postulates where the data flows and who processes the data. It permits analyst to isolate areas of interest in the organization and study them by examining the data that enter the process and viewing how they are altered when they leave

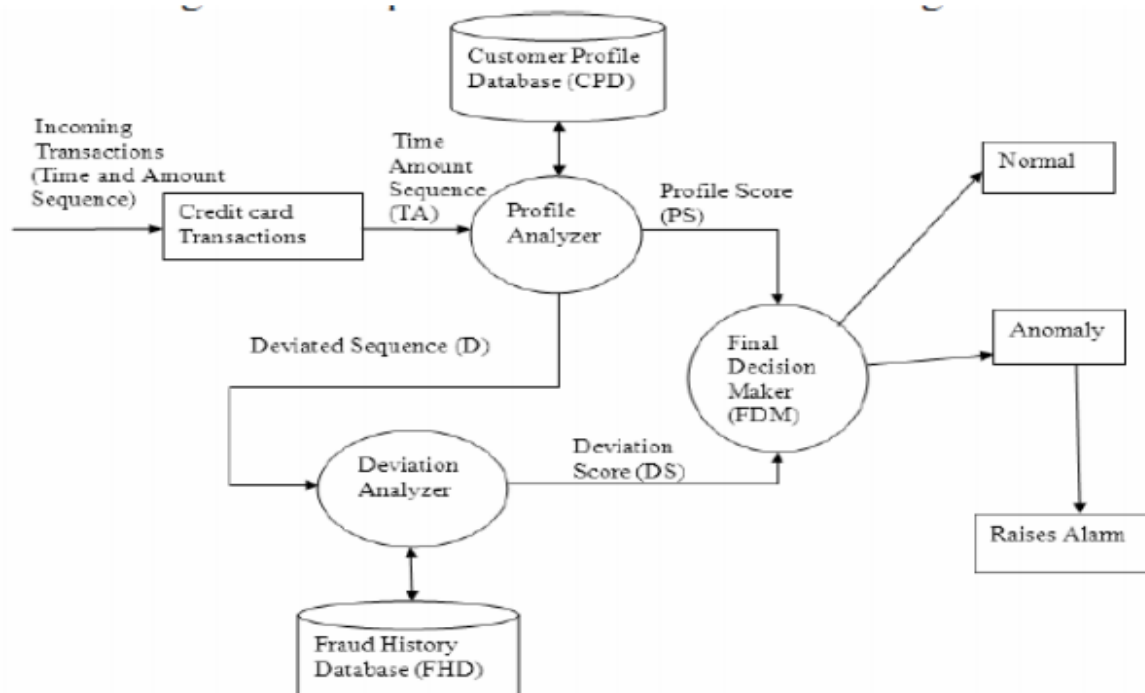
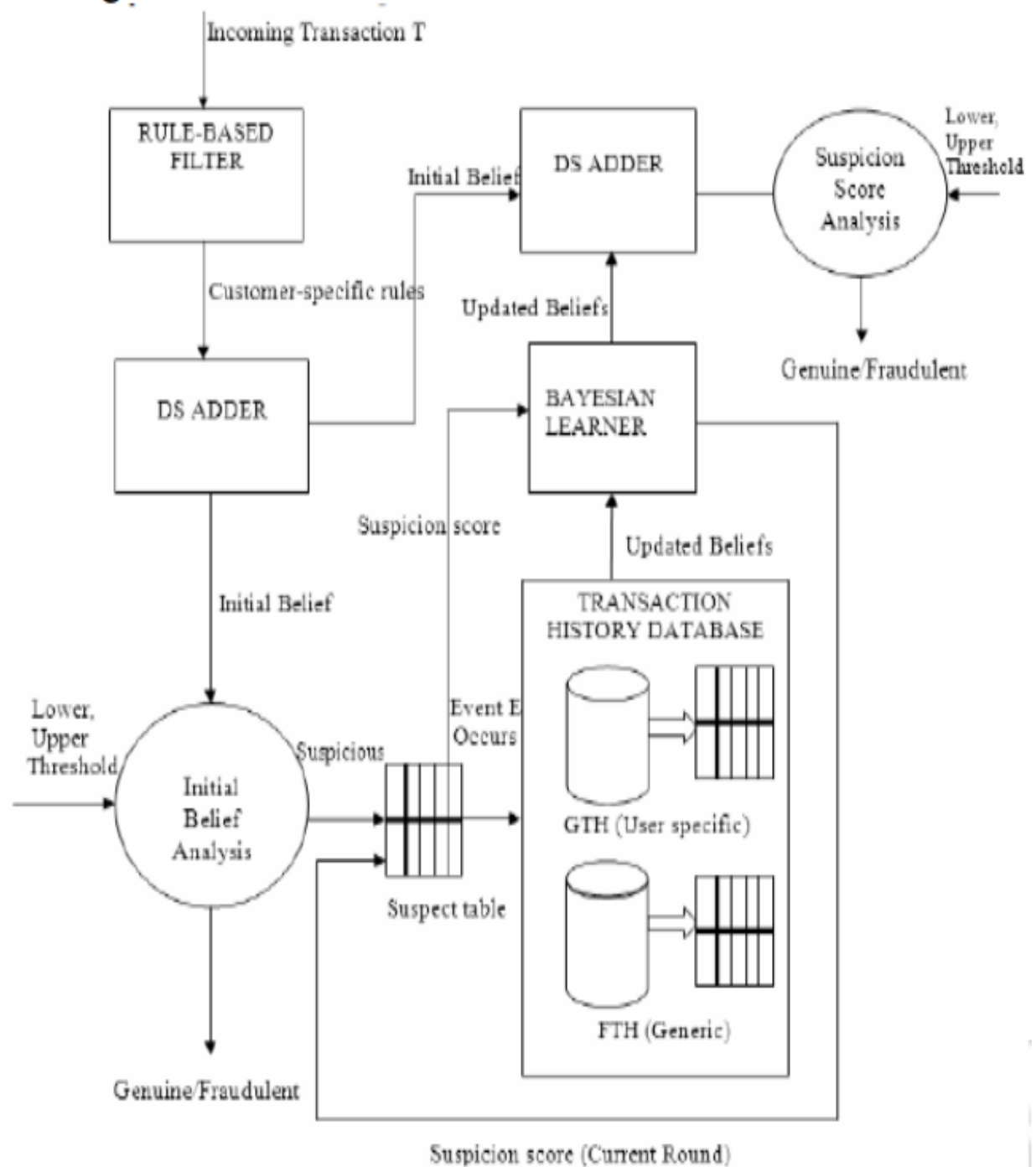


Fig no: 5.1: DATA FLOW DIAGRAM 1



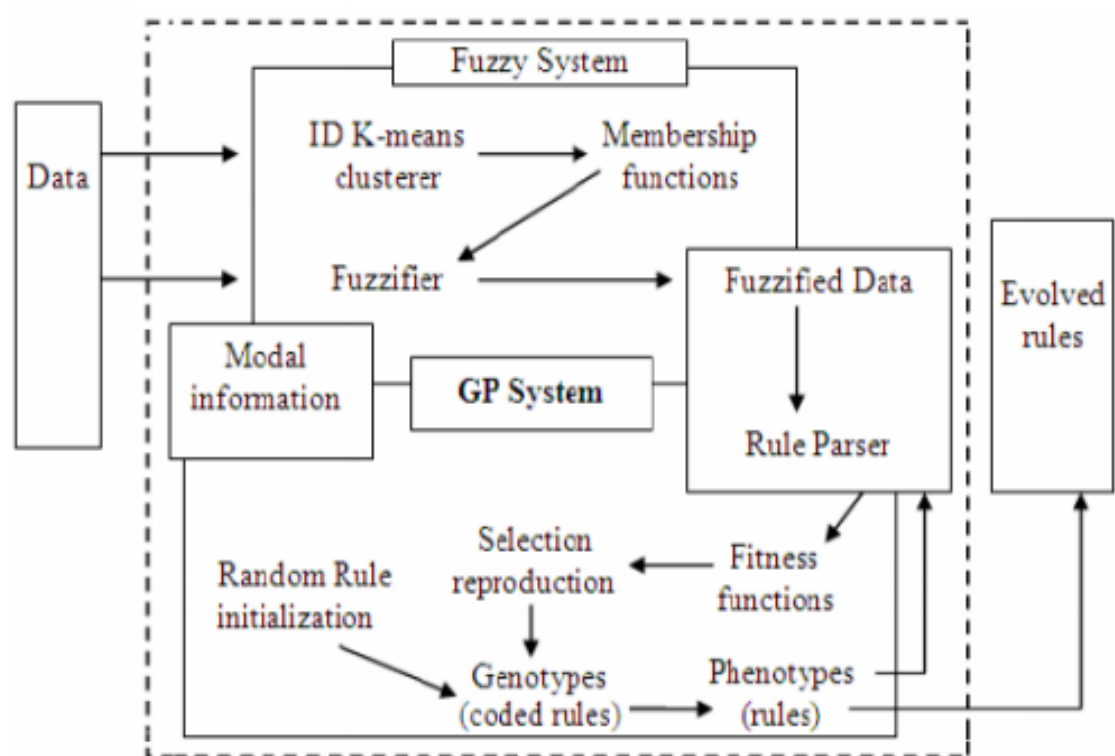


Fig no:5.3: DATA FLOW DIAGRAM 3

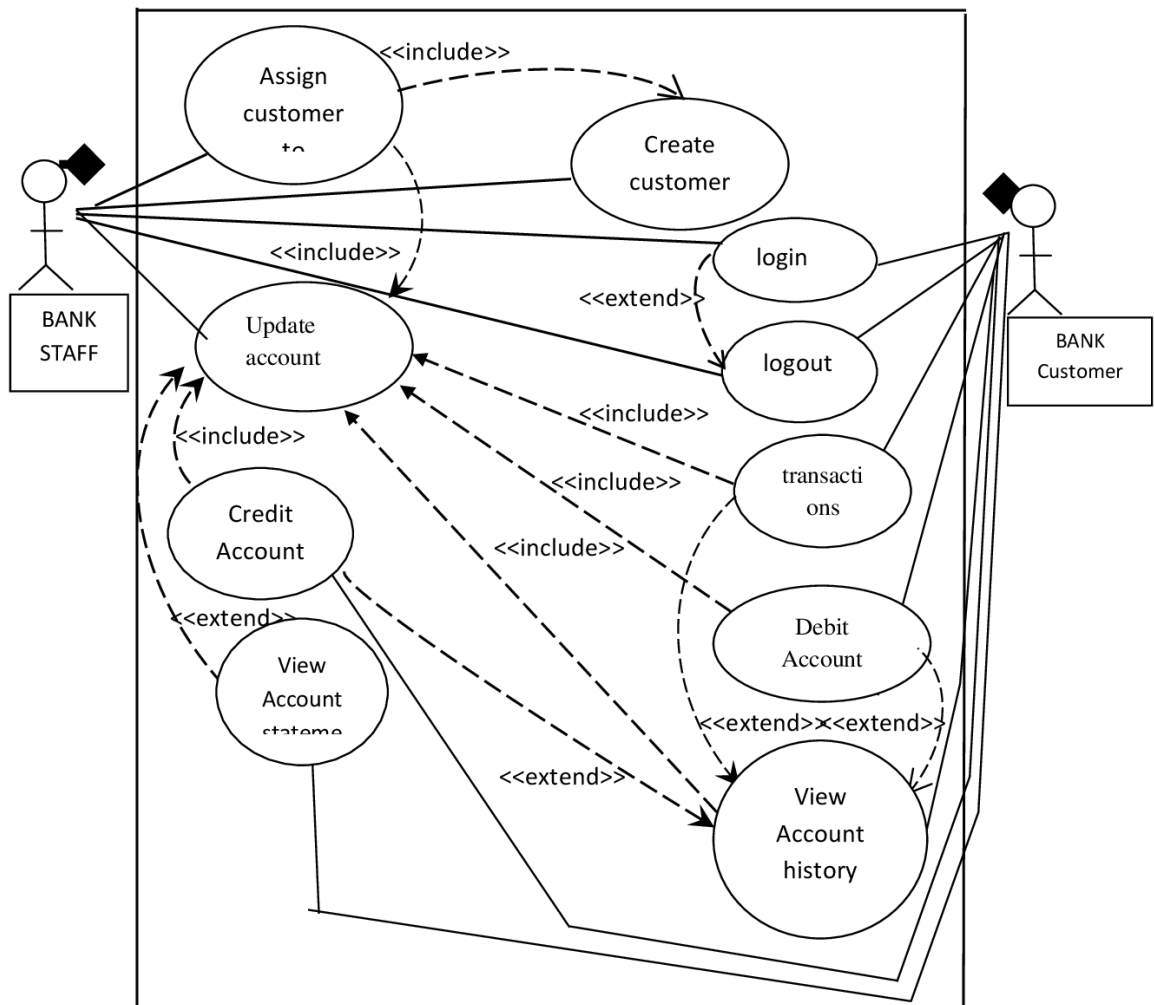


Fig no: 5.4: USE CASE DIAGRAM

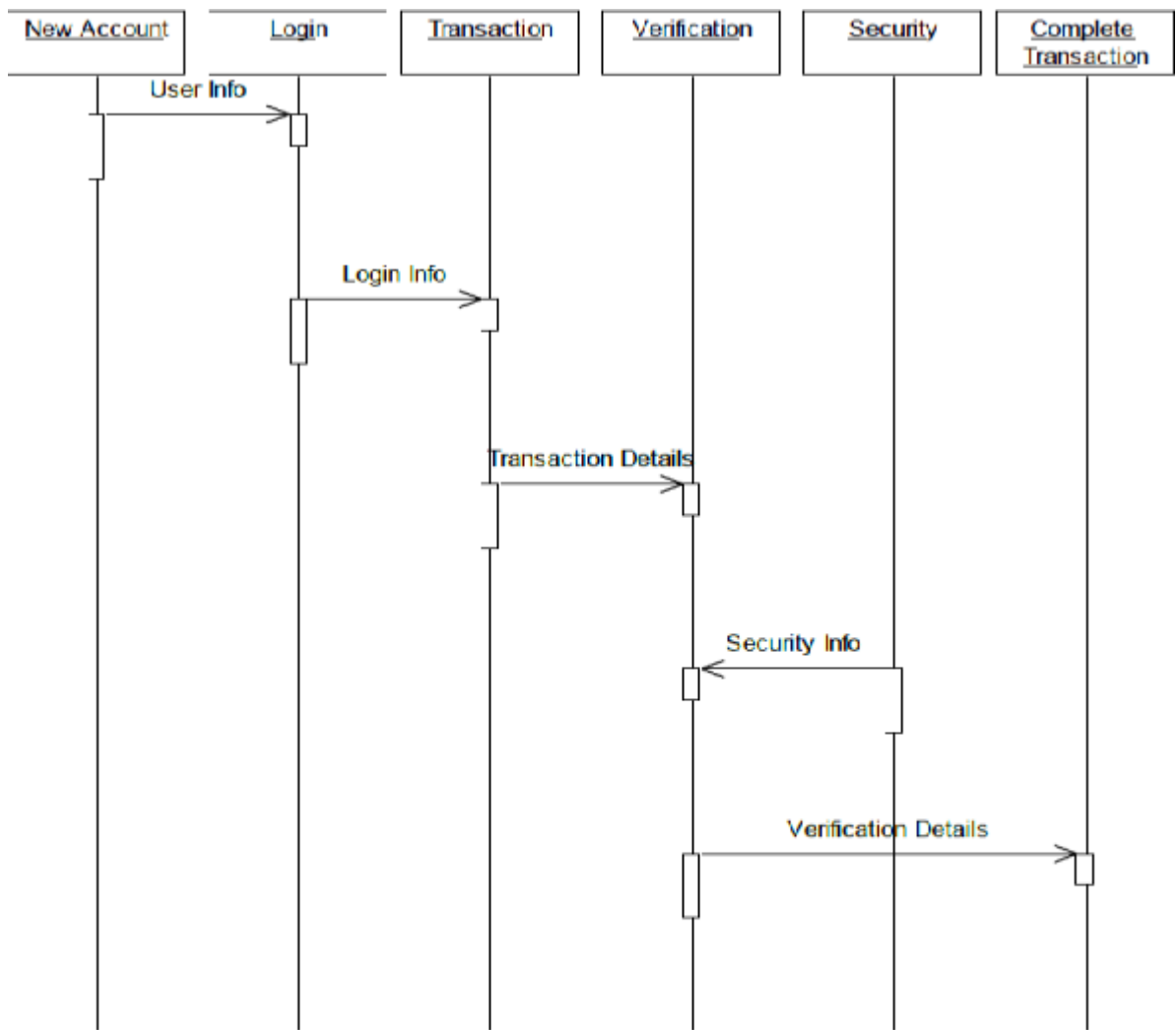


Fig no:5.5: SEQUENCE DIAGRAM

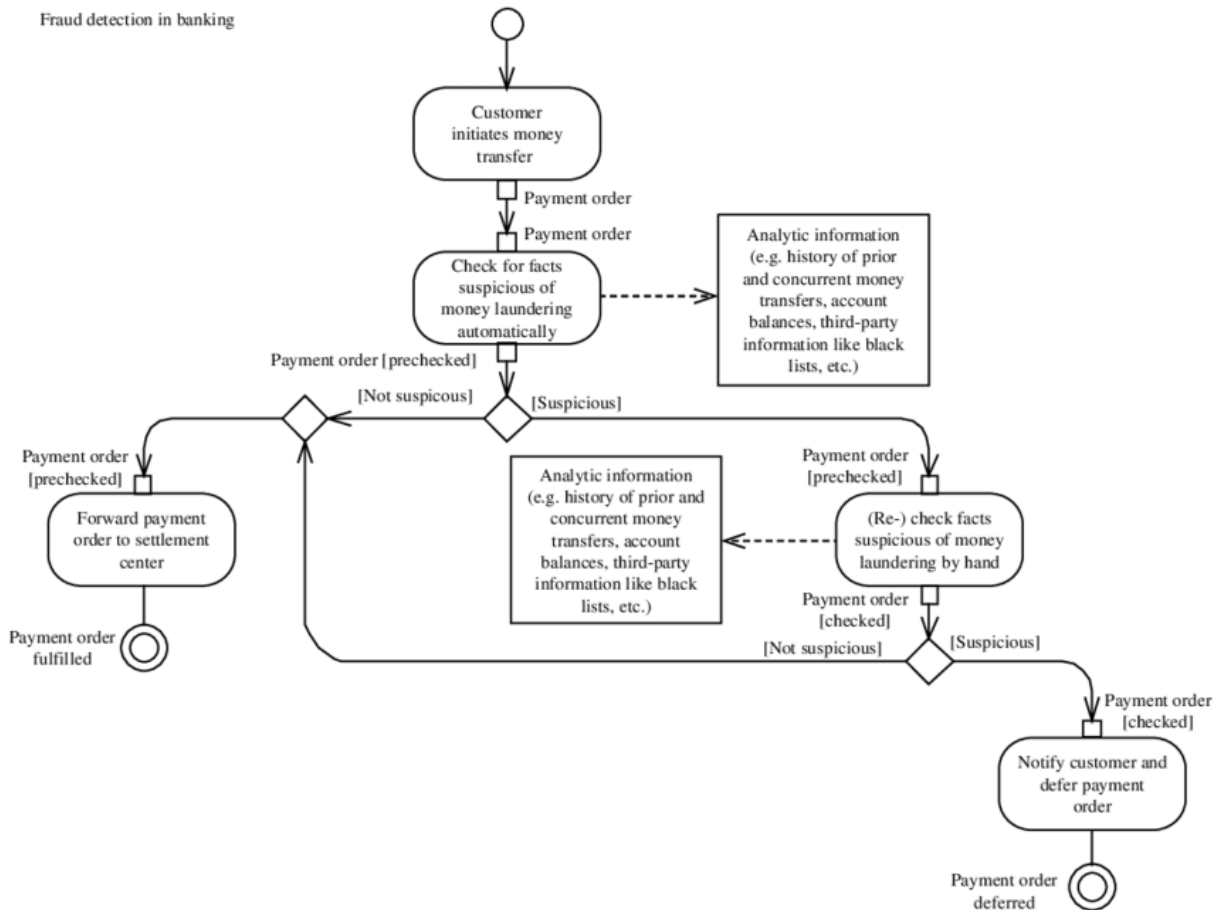


Fig no: 5.6: ACTIVITY DIAGRAM

SYSTEM DESIGN AND TESTING PLAN

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

What data should be given as input?

How the data should be arranged or coded?

The dialog to guide the operating personnel in providing input.

Methods for preparing input validations and steps to follow when error occur.

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

The output form of an information system should accomplish one or more of the following objectives.

Convey information about past activities, current status or projections of the

Future.

Signal important events, opportunities, problems, or warnings.

Trigger an action.

Confirm an action.

SYSTEM STUDY

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost

estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

REQUIREMENT ANALYSIS

Requirement analysis, also called requirement engineering, is the process of determining user expectations for a new modified product. It encompasses the tasks that determine the need for analysing, documenting, validating and managing software or system requirements. The requirements should be documentable, actionable, measurable, testable and traceable related to identified business needs or opportunities and define to a level of detail, sufficient for system design.

FUNCTIONAL REQUIREMENTS

It is a technical specification requirement for the software products. It is the first step in the requirement analysis process which lists the requirements of particular software systems including functional, performance and security requirements. The function of the system depends mainly on the quality hardware used to run the software with given functionality.

Usability

It specifies how easy the system must be use. It is easy to ask queries in any format which is short or long, porter stemming algorithm stimulates the desired response for user.

Robustness

It refers to a program that performs well not only under ordinary conditions

but also under unusual conditions. It is the ability of the user to cope with errors for irrelevant queries during execution.

Security

The state of providing protected access to resource is security. The system provides good security and unauthorized users cannot access the system there by providing high security.

Reliability

It is the probability of how often the software fails. The measurement is often expressed in MTBF (Mean Time Between Failures). The requirement is needed in order to ensure that the processes work correctly and completely without being aborted. It can handle any load and survive and survive and even capable of working around any failure.

Compatibility

It is supported by version above all web browsers. Using any web servers like localhost makes the system real-time experience.

Flexibility

The flexibility of the project is provided in such a way that is has the ability to run on different environments being executed by different users.

Safety

Safety is a measure taken to prevent trouble. Every query is processed in a secured manner without letting others to know one's personal information.

NON- FUNCTIONAL REQUIREMENTS

Portability:

It is the usability of the same software in different environments. The project can be run in any operating system.

Performance:

These requirements determine the resources required, time interval, throughput and everything that deals with the performance of the system.

Accuracy:

The result of the requesting query is very accurate and high speed of retrieving information. The degree of security provided by the system is high and effective.

Maintainability:

Project is simple as further updates can be easily done without affecting its stability. Maintainability basically defines that how easy it is to maintain the system. It means that how easy it is to maintain the system, analyse, change and test the application. Maintainability of this project is simple as further updates can be easily done without affecting its stability.

SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the

Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the

testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

- Valid Input : identified classes of valid input must be accepted.
- Invalid Input : identified classes of invalid input must be rejected.
- Functions : identified functions must be exercised.
- Output : identified classes of application outputs must be exercised.
- Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration-oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing:

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Unit Testing:

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.

- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

CHAPTER - 6

CONCLUSION

Machine Learning is a technique used to extract vital information from existing huge amount of data and enable better decision-making for the banking and retail industries. They use data warehousing to combine various data from databases into an acceptable format so that the data can be mined. The data is then analyzed and the information that is captured is used throughout the organization to support decision-making. Data Mining techniques are very useful to the banking sector for better targeting and acquiring new customers, most valuable customer retention, automatic credit approval which is used for fraud prevention, fraud detection in real time, providing segment based products, analysis of the customers, transaction patterns over time for better retention and relationship, risk management and marketing.

OUTPUT SCREENSHOTS

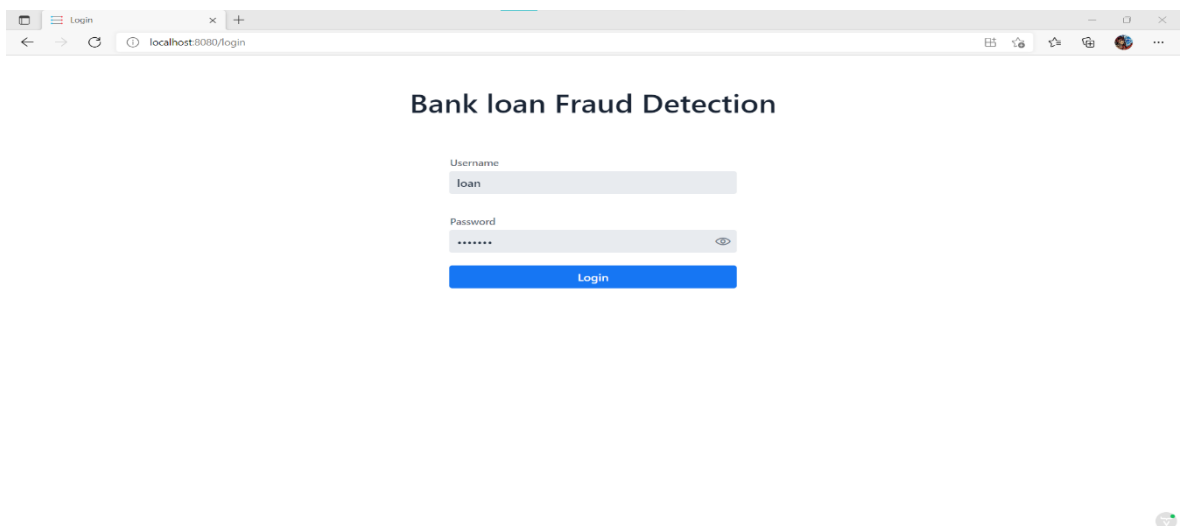


Fig no:6.1: USER LOGIN PAGE

Apply for Loan

Bank Loan

Apply for loan

Login as Loan Approver

Fill the below loan details

Aadhar Id 615098743258

Amount 50000

Send Cancel

Fig no:6.2: ADMINISTRATOR LOGIN PAGE

Bank loan Fraud Detection

Username loan

Password *****

Login

Wrong Credentials

Fig no: 6.3: Login Details Verification

Bank Loan

Loan requests

Id	Aadhar Id	Amount	Status
5	123412341236	12345	SUBMITTED
7	615032843799	100000	SUBMITTED
8	956626187668	1953269	SUBMITTED
39	619640060322	105000	SUBMITTED
72	956626187665	75000	SUBMITTED
103	123456789012	500000	SUBMITTED
104	956626187665	400000	SUBMITTED
136	9654123578529856	50000	SUBMITTED
137	619640060987	564000	SUBMITTED
138	12345789011	100000	SUBMITTED
139	123412341234	10	SUBMITTED
140	965478523654	505000	SUBMITTED
141	965478523654	9000	SUBMITTED

Id:

Aadhar Id:

Amount:

Status:

Save **Cancel**

Fig no:6.4: Loan Requests

Bank Loan

PastLoansDetails

Id	Aadhar Id	Amount	Status
6	123412341234	12345	LOAN_FRAUD
7	615032843799	100000	APPROVED
40	619640060987	50000	LOAN_FRAUD
41	609640060963	75000	APPROVED
71	956626187667	85000	APPROVED
75	965478523654	10000	REJECTED
103	123456789012	500000	APPROVED
136	9654123578529856	50000	REJECTED
139	123412341234	10	REJECTED
140	965478523654	505000	APPROVED
142	619640060987	85000	LOAN_FRAUD
168	615032843799	5000	APPROVED
200	98653553532	65000	APPROVED
202	987632545698	95000	LOAN_FRAUD
203	956626187665	50000	LOAN_FRAUD

Id:

Aadhar Id:

Amount:

Status:

Save **Cancel**

Fig no:6.5: Past-Loan Details

Bank Loan

Loan Requests
Past Loan Details
Person Details
Logout

ID	Aadhar Id	First Name	Last Name	Email	Phone	Father
1	123412341234	Father First name	Father Last name	demo@demo.com	9798798897	11
2	123412341235	Mother First name	Mother Last name	demo1@demo1.com		
3	123412341236	Son First name	Son Last name	demo2@demo2.com		1
4	123412341237	Wife First name	Wife Last name	demo3@demo3.com		
42	619640060322	srinivas	naidu	srinivas987@gmail.com	9696989878	40
73	956626187665	PUTLA	SUBHASH CHANDRA	psubhashputla99@gmail.com	9513578521	71
74	956626187665	PUTLA	SUBHASH CHANDRA	psubhashputla99@gmail.com	9513578521	71
76	956626187665	PUTLA	SUBHASH CHANDRA	psubhashputla99@gmail.com	9513578521	71
204	615032843799	PRASANNA	RAYAPATI	rayapatisai07@gmail.com	8956523589	203
206	919640060322	nani	mullepati	mullepati07@gmail.com	9940656358	202
207	956626187668	Preethi	lavangam	lavangam09@gmail.com	9998884445	103
208	619640060987	Vandan Raj	Bhupathi	vandanga95@gmail.com	9513574562	202
209	956626187665	Prem	Sana	Sana.spk999@gmail.com	9874521365	71
210	619640060987	Sandeep	chitralla	chitralasandeep987@gmail.com	6983589512	98
211	83	Bapi Reddy		BhimiReddy	9863256535	98
212	593452129874	Sudhakar	Thameneni	thamenenisidhu97@gmail.com	9876662258	98

First Name
Last Name
Email
Phone
Father Id
Mother Id
Wife or Husband Id
Date Of Birth

Save Cancel

Fig no:6.6: Person Deatils

Bank Loan

Loan Requests
Past Loan Details
Person Details
Logout

ID	Aadhar Id	Amount	Status
5	123412341236	12345	SUBMITTED
7	615032843799	100000	SUBMITTED
8	956626187668	1953269	SUBMITTED
39	619640060322	105000	SUBMITTED
72	956626187665	75000	SUBMITTED
103	123456789012	500000	SUBMITTED
104	956626187665	400000	SUBMITTED
136	9654123578529856	50000	SUBMITTED
137	619640060987	564000	SUBMITTED
138	12345789011	100000	SUBMITTED
139	123412341234	10	SUBMITTED
140	965478523654	505000	SUBMITTED
141	965478523654	9000	SUBMITTED

Id
Aadhar Id
Amount
Status
Father has a history of Loan Fraud

Save Cancel

Fig no:6.7: Fraud History

REFERENCES:

- [1] S. M. Darwish, "An intelligent credit card fraud detection approach based on semantic fusion of two classifiers," (in English), *Soft Computing*, Article vol. 24, no. 2, pp. 1243-1253, Jan 2020.
- [2] A. Eshghi and M. Kargari, "Introducing a new method for the fusion of fraud evidence in banking transactions with regards to uncertainty," (in English), *Expert Systems with Applications*, Article vol. 121, pp. 382-392, May 2019.
- [3] S. Hossain, A. Abtahee, I. Kashem, M. M. Hoque, and I. H. Sarker, "Crime Prediction Using Spatio-Temporal Data," in *Computing Science, Communication and Security*, Singapore, 2020, pp. 277-289: Springer Singapore.
- [4] M. Zamini and S. M. H. Hasheminejad, "A comprehensive survey of anomaly detection in banking, wireless sensor networks, social networks, and healthcare," (in English), *Intelligent Decision Technologies-Netherlands*, Article vol. 13, no. 2, pp. 229-270, 2019.
- [5] I. Gonzalez-Carrasco, J. L. Jimenez-Marquez, J. L. Lopez-Cuadrado, and B. Ruiz-Mezcua, "Automatic detection of relationships between banking operations

using machine learning," (in English), Information Sciences, Article vol. 485, pp. 319-346, Jun 2019.

[6] M. Pohoretskyi, D. Serhieieva, and Z. Toporetska, "The proof of the event of a financial resources fraud in the banking sector: problematic issues," (in English), Financial and Credit Activity-Problems of Theory and Practice, Article vol. 1, no. 28, pp. 36-45, 2019.

[7] K. Noor et al., "Performance analysis of a surveillance system to detect and track vehicles using Haar cascaded classifiers and optical flow method," 2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA), Siem Reap, 2017, pp. 258-263.