# Requirement Engineering Document

# B. Tech (VII Semester) Project Work-1

**------------------------------------------------------------------------------------**

**Credit Card Fraud Detection Website**

A SRS for

Project Work-1

**BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE & ENGINEERING**

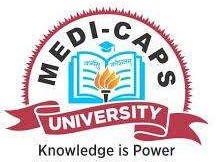
BY

## Aaditya Anand Abhay Singh Chauhan Advait Patidar

**EN19CS301004 EN19CS301011 EN19CS301027**

Under the Guidance of

**Dr. Harsh Pratap Singh**



## Department of Computer Science & Engineering Faculty of Engineering

**MEDI-CAPS UNIVERSITY, INDORE- 453331**

## NOVEMBER 2022

***Introduction***

As the payment method is simplified by the combination of the financial industry and IT technology, the payment method of the consumers is changing from cash payment to electronic payment using a credit card, mobile micropayment, and app card. As a result, the number of cases in which anomalous transactions are attempted by abusing e-banking has increased and financial companies started establishing a Fraud Detection System (FDS) to protect consumers from abnormal transactions. The abnormal transaction detection system aims to identify abnormal transactions with high accuracy by analyzing user information and payment information in real-time. Although FDS has shown good results in reducing fraud, most cases being flagged by this system are False Positives that result in substantial investigation costs and cardholder inconvenience. The possibilities of enhancing the current operation constitute the objective of this research. Based on variations and combinations of testing and training class distributions, experiments were performed to explore the influence of these parameters. In this project, we will investigate the trend of abnormal transaction detection using payment log analysis and data mining and summarize the data mining algorithm used for abnormal credit card transaction detection. We will use python programming with Apache spark for advanced processing of data and high accuracy.

**Credit – card-based purchases can be categorized into two types:**

* Physical
* Virtual

In a physical – card-based purchase, the cardholder presents his card physically to a merchant for making a payment. To carry out fraudulent transactions in this kind of purchase, an attacker must steal the credit card. If the cardholder does not realize the loss of card, it can lead to a substantial financial loss to the credit card company. In this kind of purchase, only some important information about the card is required to make the payment. Such purchases are normally done on the Internet or over the phone. To commit fraud in these types of purchases, a fraudster simply needs the card details.

Most of the time, the genuine cardholder is not aware of that someone else has seen or stolen his card information. The only way to detect this kind of fraud is to analyze the spending patterns on every card and to figure out any inconsistency with respect to the “usual” spending patterns. Fraud detection based on the analysis of existing purchase data of cardholder is a promising way to reduce the rate of successful credit card frauds. Since humans tend to exhibit specific behavioristic profiles, every cardholder can be represented by a set of patterns containing information about the typical purchase category, the time since the last purchase, the amount of money spent, etc. Deviation from such patterns is a potential threat to the system.

***Objectives***

The main objective of the Project on Credit Card Fraud Detection System is to manage the details of Credit Card, Transactions, Datasets, Files, Prediction. It manages all the information about Credit Card, Customer, Prediction, Credit Card. The project is totally built at administrative end and thus only the administrator is guaranteed the access. The purpose of the project is to build an application program to reduce the manual work for managing the Credit Card and its transaction.

**Functionalities provided by Credit Card Fraud Detection System are as follows:**

* Provides searching facilities based on factors such as credit card, data sets, files.
* It will also manage the customer details online for the credit card.
* It will track the transaction and customer information.
* Manages credit card information.
* Editing, adding, and updating records of database which results in proper resource management of Credit Card Database.

***Feasibility Study***

Feasibility study includes consideration of all the possible ways to provide a solution to the given problem. The proposed solution should satisfy all the user requirements and should be flexible enough so that future changes can be easily done based on the future upcoming requirements.

1. **Economic Feasibility**

* The technology used in the project is decided based on the minimum possible cost factor.
* All hardware and software cost has to be borne by the organization.
* Overall, we have estimated that the benefits the organization is going to receive from the proposed system will surely overcome the initial costs and the running cost for system.

1. **Technical Feasibility**

This includes the study of function, performance and constraints that may affect the ability to achieve an acceptable system. For this feasibility study, we study complete functionality to be provided in the system, as described in the SRS, and checked if everything was possible using different types of frontend and backend platforms.

1. **Operational Feasibility**

The proposed system is fully GUI based and user friendly. A proper training has been conducted to let the employees know of the essence of the system so that they feel comfortable with the system.

***Software Requirements Specification***

##### **GENERAL DESCRIPTION OF THE SYSTEM**

A proper and thorough literature survey concludes that there are various methods that can be used to detect credit card fraud detection. Some of these approaches are:

* + 1. Artificial Neural Network
    2. Bayesian Network
    3. Neural Network
    4. Hidden Markov Method
    5. Genetic Algorithm

##### **Overview of Functional requirements**

**Preprocess Data**

Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. Therefore, certain steps are executed to convert the data into a small clean data set. This technique is performed before the execution of Iterative Analysis. The set of steps is known as Data Preprocessing.

Data Preprocessing is necessary because of the presence of unformatted real-world data. Mostly real- world data is composed of –

**Inaccurate data (missing data)** - There are many reasons for missing data such as data is not continuously collected, a mistake in data entry, technical problems with biometrics and much more.

**Inconsistent data -** The presence of inconsistencies is due to the reasons such that existence of duplication within data, human data entry, containing mistakes in codes or names, i.e., violation of data constraints and much more.

**Train and Test Data Creation**

The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model’s prediction on this subset.

**Model Creation**

The process of training an ML model involves providing an ML algorithm (that is, the learning algorithm) with training data to learn from. The term ML model refers to the model artifact that is created by the training process.

The training data must contain the correct answer, which is known as a target or target attribute. The learning algorithm finds patterns in the training data that map the input data attributes to the target (the answer that you want to predict), and it outputs an ML model that captures these patterns.

You can use the ML model to get predictions on new data for which you do not know the target. For example, let's say that you want to train an ML model to predict if an email is spam or not spam. You would provide training data that contains emails for which you know the target (that is, a label that tells whether an email is spam or not spam). Machine would train an ML model by using this data, resulting in a model that attempts to predict whether new email will be spam or not spam.

In our project we are using Random Forest Algorithm to build our Model on Credit Card Fraud Dataset.

**OVERVIEW OF DATA REQIREMENTS**

Creating a training data set that will allow our algorithms to pick up the specific characteristics that make a transaction more or less likely to be fraudulent. Using the original data set would not prove to be a good idea for a very simple reason: Since over 99% of our transactions are non-fraudulent, an algorithm that always predicts that the transaction is non-fraudulent would achieve accuracy higher than 99%. Nevertheless, that is the opposite of what we want. We do not want a 99% accuracy that is achieved by never labeling a transaction as fraudulent, we want to detect fraudulent transactions and label them as such.

There are two key points to focus on to help us solve this. First, we are going to utilize **random under- sampling** to create a training dataset with a balanced class distribution that will force the algorithms to detect fraudulent transactions as such to achieve high performance. Speaking of performance, we are not going to rely on accuracy. Instead, we are going to make use of the Receiver Operating Characteristics- Area Under the Curve or ROC-AUC performance measure (I have linked further reading below this article). Essentially, the ROC-AUC outputs a value between zero and one, whereby one is a perfect score and zero the worst. If an algorithm has a ROC-AUC score of above 0.5, it is achieving a higher performance than random guessing.

To create our balanced training data set, I took all the fraudulent transactions in our data set and counted them. Then, I randomly selected the same number of non-fraudulent transactions and concatenated the two. After shuffling this newly created data set, I decided to output the class distributions once more to visualize the difference.

**Technical Requirements of the System**

**Hardware Requirements**

System Processor : Core i3 / i5 / i7

Hard Disk : 500 GB.

Ram : 4 GB.

*Any desktop / Laptop system with above configuration or higher level.*

**Software Requirements**

Operating system : Window 10

Programming Language : Python Framework : Anaconda

IDE : Jupyter Notebook

DL Libraries : Numpy, Pandas