A
Project Report
On
"Credit Card Fraud Detection"

Prepared by

Rudri Bhatt (22CE008)

Under the guidance of

Dr. Mrugendra Rahevar

Submitted to

Charotar University of Science & Technology

Degree of Bachelor of Technology

in Computer Engineering

CE363: Project-III

Of 5th Semester of B.Tech

Submitted at

CHARUSAT
CHAROTAR UNIVERSITY OF SCIENCE AND TECHNOLOGY

U. & P. U PATEL DEPARTMENT OF COMPUTER ENGINEERING

Faculty of Technology & Engineering, CHARUSAT

Chandubhai S. Patel Institute of Technology

At: Changa, Dist: Anand – 388421

November 2024

A
Project Report
On
"Credit Card Fraud Detection"

(CE363 – Project-III)

Prepared by

Rudri Bhatt (22CE008)

Under the Supervision of

Dr. Mrugendra Rahevar

Submitted to

Charotar University of Science & Technology (CHARUSAT)
for the Partial Fulfillment of the Requirements for the
Degree of Bachelor of Technology (B.Tech.)
in U & P U. Patel Department of Computer Engineering (CE)
for B.Tech Semester 5

Submitted at

Accredited with Grade A by NAAC

U & P U. PATEL DEPARTMENT OF COMPUTER ENGINEERING
Chandubhai S. Patel Institute of Technology (CSPIT)
Faculty of Technology & Engineering (FTE), CHARUSAT
At: Changa, Dist: Anand, Pin: 388421.
November, 2024

CERTIFICATE

This is to certify that the report entitled "**Credit Card Fraud Detection**" is a bonafide work carried out by **Rudri Bhatt (22CE008)** under the guidance and supervision of Dr. Mrugendra Rahevar for the subject **Project - III (CE363)** of 5th Semester of Bachelor of Technology in **Computer Engineering** at Faculty of Technology & Engineering (C.S.P.I.T.) – CHARUSAT, Gujarat.

To the best of my knowledge and belief, this work embodies the work of the candidate himself, has duly been completed, and fulfills the requirement of the ordinance relating to the B.Tech. Degree of the University and is up to the standard in respect of content, presentation and language for being referred to the examiner.

Under the supervision of,

Dr. Mrugendra Rahevar
Assistant Professor
U. & P. U Patel Dept. of Computer Engg.
C.S.P.I.T., CHARUSAT-Changa.

Dr. Nikita Bhatt
Head,
U. & P. U Patel Dept. of Computer Engg.
C.S.P.I.T., CHARUSAT-Changa.

Chandubhai S Patel Institute of Technology (C.S.P.I.T.)

Faculty of Technology & Engineering, CHARUSAT

At: Changa, Ta. Petlad, Dist. Anand, PIN: 388 421. Gujarat

DECLARATION BY THE CANDIDATE

I hereby declare that the project report entitled "**Credit Card Fraud Detection**" submitted by me to Chandubhai S. Patel Institute of Technology, Changa in partial fulfillment of the requirements for the award of the degree of **B.Tech Computer Engineering**, from U & P U. Patel Department of Computer Engineering, CSPIT, FTE, is a record of bonafide CE363 Project-III (project work) carried out by me under the guidance of Dr. Mrugendra Rahevar. I further declare that the work carried out and documented in this project report has not been submitted anywhere else either in part or in full and it is the original work, for the award of any other degree or diploma in this institute or any other institute or university.

Rbhatt

(Rudri Bhatt- 22CE008)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Dr. Mrugendra Rahevar
Assistant Professor
U & P U. Patel Department of Computer Engineering,
Chandubhai S Patel Institute of Technology (CSPIT)
Faculty of Technology (FTE)
Charotar University of Science and Technology (CHARUSAT) - Changa.

TABLE OF CONTENTS

	(Book / Web-link / Journal /Articles/Magazine/online course/online video lectures/ References) Referencing in the thesis should be as per any one of the internationally accepted Referencing System (Harvard System/APA System/Vancouver System/Oxford/Footnote.....etc)	
Anti-Plagiarism Report		

CHAPTER 1: INTRODUCTION

1.1 Introduction

Credit card fraud is a critical concern in the financial sector, causing substantial financial losses and affecting both businesses and consumers. As digital transactions continue to grow, fraudsters are finding increasingly sophisticated methods to exploit vulnerabilities. Traditional fraud detection methods, which rely heavily on labeled data and supervised learning techniques, often struggle with identifying novel fraud patterns. Unsupervised learning methods, however, provide a promising alternative for detecting fraudulent activities without needing pre-labeled data. This research leverages unsupervised clustering algorithms to identify anomalous credit card transactions and potentially fraudulent activities in a dataset.

1.2 Research Definition

This research explores the application of unsupervised learning algorithms—DBSCAN (Density-Based Spatial Clustering of Applications with Noise), GMM (Gaussian Mixture Model), and K-means clustering—to detect credit card fraud. The aim is to evaluate and compare the performance of these algorithms in identifying fraudulent transactions, focusing on clustering behavior and the ability to differentiate between normal and fraudulent transactions in an unlabeled dataset.

1.3 Problem Description

Detecting credit card fraud presents unique challenges:

- Fraudulent transactions represent a small proportion of all transactions, making it difficult to detect without oversampling or rebalancing.
- Fraud patterns vary widely, making it challenging to use rigid rules or supervised learning.
- Manual fraud detection is labor-intensive and not feasible for real-time analysis.

This research addresses these issues by implementing and testing three clustering algorithms that do not require labeled data, making them suitable for real-time fraud detection. Each algorithm's clustering structure will be analyzed to determine its effectiveness in distinguishing between legitimate and potentially fraudulent transactions.

1.4 Motivation

The motivation for this research is driven by the need for more adaptable and scalable fraud detection systems that can operate with minimal supervision. With an increasing number of digital transactions, organizations are under pressure to employ faster, more efficient fraud detection mechanisms that adapt to evolving patterns without the need for extensive labeled data. Unsupervised learning techniques have the potential to automatically identify patterns in the data, allowing organizations to detect fraud more effectively and respond proactively to emerging threats.

1.5 Scope and Objectives

The scope of this research is limited to the exploration of unsupervised learning techniques for credit card fraud detection using three clustering algorithms:

1. **DBSCAN**: To test its effectiveness in isolating noise (fraudulent transactions) from dense regions of legitimate transactions.
2. **Gaussian Mixture Model (GMM)**: To assess its ability to model transaction clusters probabilistically and its flexibility in handling data that may not have distinct boundaries.
3. **K-means**: To evaluate its baseline clustering effectiveness in separating potential fraud clusters within transaction data.

Objectives:

- Implement DBSCAN, GMM, and K-means on a credit card transaction dataset.
- Evaluate and compare clustering performance using relevant metrics (e.g., silhouette score, anomaly detection accuracy).
- Identify the clustering model most effective for real-time or semi-supervised fraud detection use cases.

1.6 Planning

This research follows a structured plan:

1. **Data Collection and Preprocessing**: Prepare a credit card transaction dataset, focusing on feature engineering and scaling for algorithm compatibility.
2. **Algorithm Implementation**: Implement DBSCAN, GMM, and K-means, adjusting parameters to optimize clustering results.
3. **Evaluation and Comparison**: Assess each model's performance using quantitative metrics to evaluate its fraud detection capability.
4. **Analysis and Reporting**: Present findings on each algorithm's strengths and limitations and propose practical applications and future research directions.

1.7 Organization of Thesis

This thesis is organized as follows:

- **Chapter 1: Introduction** – Outlines the background, problem, and motivation for fraud detection using unsupervised learning.
- **Chapter 2: Literature Review** – Reviews previous work on fraud detection in finance, focusing on unsupervised learning techniques and clustering approaches.
- **Chapter 3: Methodology** – Describes the dataset, preprocessing steps, and the implementation of DBSCAN, GMM, and K-means algorithms.
- **Chapter 4: Experimental Results and Analysis** – Presents results of the clustering experiments, comparing each algorithm's performance on fraud detection metrics.
- **Chapter 5: Conclusion and Future Work** – Summarizes key findings, discusses limitations, and suggests future research directions.

CHAPTER 2: Literature Review & Comparative Study

2.1 Evolution of the Topic

The study of fraud detection, particularly in credit card transactions, has evolved significantly due to the growth in digital transactions and the rise in cybercrime. Early approaches relied heavily on rule-based and supervised learning techniques. Rule-based systems used predefined thresholds to flag suspicious activities, while supervised machine learning models trained on labeled transaction data helped identify fraudulent patterns. However, these methods often struggled with generalizability, especially as fraudsters adapted their strategies to evade detection.

In recent years, unsupervised learning methods have gained traction in fraud detection due to their adaptability and ability to identify anomalies in data without requiring labels. Techniques such as clustering have become popular, as they can isolate outliers or small groups of suspicious transactions within large datasets. This shift has enabled financial institutions to detect novel or evolving fraud patterns more effectively, paving the way for the application of algorithms like DBSCAN, Gaussian Mixture Model (GMM), and K-means in fraud detection.

2.2 Background Study (Theoretical and Mathematical Background)

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN is a density-based clustering algorithm that groups points closely packed together, designating points in low-density areas as noise or outliers. It relies on two parameters: `epsilon` (ε), the radius within which points are considered neighbors, and `minPts`, the minimum number of points required to form a dense region. DBSCAN's ability to label noise makes it a useful algorithm for fraud detection, as fraudulent transactions often appear as isolated points in a high-density dataset.

Mathematically, DBSCAN identifies clusters by iteratively searching for points within the `ε` neighborhood of a given point. If the number of points within this radius meets `minPts`, a new cluster is formed; otherwise, the point is marked as noise. This algorithm is non-deterministic, and its performance can be sensitive to the choice of `ε` and `minPts`, making parameter selection crucial in practical applications.

Gaussian Mixture Model (GMM)

GMM is a probabilistic model that assumes the data is generated from a mixture of several Gaussian distributions with unknown parameters. Each Gaussian component represents a cluster, with the probability density function calculated based on the mean and covariance matrix of each distribution. GMM's flexibility allows it to model clusters of various shapes, making it useful for fraud detection where fraudulent transactions may not follow a specific pattern.

GMM optimizes the likelihood of the data belonging to each Gaussian component, and the Expectation-Maximization (EM) algorithm is often used to estimate parameters. The EM algorithm alternates between assigning data points to clusters based on current parameter estimates (E-step) and updating parameters to maximize data likelihood (M-step).

K-means Clustering

K-means is a partitioning algorithm that divides the dataset into `K` clusters by minimizing the variance within each cluster. It iteratively assigns each point to the nearest cluster centroid, recalculates centroids, and repeats until convergence. K-means is efficient but sensitive to the initial placement of centroids, and it assumes clusters are spherical and of similar size, which may limit its accuracy in detecting fraud in non-uniform data.

Mathematically, K-means minimizes the sum of squared distances between data points and their respective cluster centroids. However, its requirement for specifying the number of clusters (`K`) in advance and its sensitivity to initialization make it less adaptable in cases of highly variable or unstructured data, such as in fraud detection scenarios.

2.3 Review of Previous Research Findings

Several studies have investigated unsupervised learning methods for fraud detection:

- **DBSCAN in Fraud Detection**: Studies have shown DBSCAN's effectiveness in detecting anomalies in finance by isolating low-density areas. For instance, research by Xu and Tian (2015) demonstrated DBSCAN's success in identifying outliers in credit card data, noting its robustness in distinguishing dense legitimate transactions from sparse fraudulent activities. However, DBSCAN's reliance on manually set parameters (ε and minPts) can limit its adaptability across datasets (Xu & Tian, 2015).

- **Gaussian Mixture Model for Anomaly Detection**: GMM has been used in fraud detection to leverage its probabilistic clustering, allowing for flexible modeling of complex data distributions. Research by Sahin and Duman (2017) applied GMM to fraud detection, where the model's probabilistic nature helped identify fraudulent patterns that were less well-defined. Although GMM can adapt to various data shapes, studies indicate it may struggle with distinguishing clear fraud boundaries when fraudulent transactions are sporadic.
- **K-means in Fraud Detection**: K-means has been widely used as a baseline for clustering in fraud detection studies due to its simplicity and efficiency. Research by Bhattacharyya et al. (2011) showed that K-means could effectively cluster transactions in cases where fraud patterns were moderately distinct. However, K-means often underperforms in unbalanced datasets or those with overlapping clusters, as it may not effectively capture fraudulent outliers, leading to high false negatives (Bhattacharyya et al., 2011).

2.4 Comparative Study

A comparative analysis of DBSCAN, GMM, and K-means highlights their respective strengths and weaknesses:

Table 2.4.1 Comparison between DBSCAN, GMM, and K-means

Algorithm	Strengths	Weaknesses	Suitability for Fraud Detection
DBSCAN	Excellent at detecting noise; does not require the number of clusters.	Sensitive to parameter selection; struggles with varying density clusters.	High, especially for sparse fraud data or anomalies.
GMM	Flexible modeling with probabilistic clusters; good for overlapping clusters.	Computationally intensive; may have difficulty with clear fraud boundaries.	Moderate, useful for complex fraud patterns.

K-means	Simple and computationally efficient; widely applicable as a baseline.	Requires predefined K; struggles with non-spherical clusters.	Moderate to low, often used as a benchmark.

DBSCAN is particularly well-suited to datasets where fraudulent transactions are rare and appear as isolated outliers, as it can handle noise effectively. GMM’s probabilistic framework makes it a good choice for more complex fraud detection cases, though its computational complexity can be a drawback. K-means is efficient and widely used but may underperform in identifying anomalies due to its assumptions about cluster shape and size.

CHAPTER 3: Experimentation / Simulation / Lab Setup

3.1 Dataset Selection and Preprocessing

The dataset for this study is a publicly available credit card transaction dataset, containing a total of 284,807 transactions with 31 columns. It includes both **legitimate** and **fraudulent transactions** labeled in the Class column (where 0 denotes legitimate and 1 denotes fraud), with only 0.17% of transactions marked as fraud. Key attributes include:

- **Features**: Columns V1 through V28 are features derived from Principal Component Analysis (PCA), which preserves transaction data privacy.
- **Transaction-Specific Columns**: The Time column reflects the transaction timestamp, and Amount shows the transaction amount.
- **Target Column**: The Class column serves as the target label, indicating fraudulent transactions.

Data Preprocessing Steps

1. **Handling Class Imbalance**: Fraudulent transactions are notably rare, comprising only 0.17% of the dataset. This high imbalance is particularly challenging for clustering algorithms. We address this by selecting appropriate evaluation metrics that account for the low fraud prevalence.
2. **Feature Scaling**: While `V1` to `V28` are already standardized from PCA transformation, `Time` and `Amount` were further standardized using z-scores to align their scales with the other features.
3. **Dimensionality Reduction**: PCA was applied again to reduce noise and computational complexity, focusing on retaining essential components representing the majority of variance in the data.

3.2 Model Implementation

Each unsupervised clustering algorithm—DBSCAN, Gaussian Mixture Model (GMM), and K-means—was implemented to assess their clustering performance on the preprocessed dataset.

1. **DBSCAN Implementation**: DBSCAN, a density-based clustering algorithm, was tested with various configurations of epsilon (ε) and minPts. A grid search method was applied to optimize these parameters, which is crucial due to DBSCAN's sensitivity to parameter selection. Outliers or sparse fraudulent points were expected to be isolated as noise by DBSCAN.
2. **GMM Implementation**: For GMM, which fits the data to multiple Gaussian distributions, the number of Gaussian components was adjusted to optimize clustering effectiveness. The Expectation-Maximization (EM) algorithm was used iteratively to estimate the parameters that maximize the likelihood of the observed data.
3. **K-means Implementation**: K-means clustering was implemented with the number of clusters (K) determined using the elbow method, identifying the optimal K where adding clusters no longer significantly reduces within-cluster variance. The k-means++ initialization method was applied to improve cluster convergence.

3.3 Evaluation Metrics

To evaluate and compare model performance, we used the following metrics:

- **Silhouette Score**: Assesses the cohesion and separation quality of clusters, offering insights into the clustering structure.
- **Cluster Purity and Precision**: Measures how accurately fraudulent transactions are clustered separately from legitimate ones.
- **Detection Rate (Recall)**: Quantifies the model's ability to correctly identify fraud.
- **F1-Score**: Combines precision and recall, providing a balanced metric for performance in an imbalanced dataset.

Each model's results were stored for comparative analysis and visualization in Chapter 4.

CHAPTER 4: Results and Discussion on Results

4.1 Introduction

This chapter presents the outcomes of applying three clustering techniques—K-Means, DBSCAN, and Gaussian Mixture Model (GMM)—to the credit card transaction dataset. Each method was evaluated based on its ability to cluster fraudulent and non-fraudulent transactions using metrics such as Silhouette Score, Adjusted Rand Index (ARI), and Davies-Bouldin Index (DBI). Cluster visualizations are included to analyze the separation and compactness of clusters.

4.2 K-Means Results

K-Means clustering was performed after preprocessing the dataset, including UMAP-based dimensionality reduction and Z-score-based outlier removal. Two clusters were hypothesized, corresponding to fraudulent and non-fraudulent transactions.

- **Evaluation Metrics:** Silhouette Score: **0.313**

The moderate Silhouette Score indicates that the clusters were somewhat distinct but not perfectly well-defined.

- **Visualization:**
  The clusters were visualized in the reduced UMAP space, revealing partial overlap between the two clusters. The scatter plot demonstrated reasonable separation but lacked clear boundaries, particularly in regions where fraudulent and non-fraudulent transactions appeared similar.
- **Discussion:**
  K-Means performed moderately well in creating distinct clusters but struggled with data points near the boundaries. Its sensitivity to outliers and noise, coupled with its reliance on centroid-based assumptions, limited its suitability for this dataset. While it demonstrated reasonable separability, the lack of noise handling detracts from its performance in detecting fraud.

4.3 DBSCAN Results

DBSCAN, a density-based clustering algorithm, was applied using an optimal epsilon (eps) value determined from a K-distance graph. A minimum sample size of 10 was used to define clusters.

- **Evaluation Metrics:**
    - Silhouette Score: **0.892**
    - Adjusted Rand Index (ARI): **-0.000**
    - Davies-Bouldin Index (DBI): **0.970**
    - Number of Noise Points: **2**

The high Silhouette Score signifies that DBSCAN formed well-defined clusters with low inter-cluster similarity. The low DBI further reinforces the distinctiveness of the clusters. However, the ARI value suggests a lack of agreement between DBSCAN's results and the true labels.

- **Visualization:**
  The scatter plot of DBSCAN clusters using PCA-reduced components illustrated clear cluster boundaries and successfully identified noise points. The algorithm effectively isolated dense regions while labeling sparse regions as noise.
- **Discussion:**
  DBSCAN outperformed K-Means in terms of cluster quality and the ability to handle noise. The density-based approach excelled in identifying well-defined clusters and managing outliers, making it a strong candidate for fraud detection. However, the mismatch with true labels (low ARI) indicates a need for additional feature engineering or parameter tuning to align better with the dataset's structure.

4.4 Gaussian Mixture Model (GMM) Results

The GMM algorithm was applied after reducing dimensionality using PCA to retain 95% of the variance. Two Gaussian components were specified, corresponding to the expected clusters of fraudulent and non-fraudulent transactions.

- **Evaluation Metrics:**
    - Silhouette Score: **0.046**

- Adjusted Rand Index (ARI): **-0.000**
- Davies-Bouldin Index (DBI): **4.587**

The low Silhouette Score highlights that the clusters lacked cohesion and separation. A near-zero ARI indicates poor alignment with the ground truth labels. The high DBI underscores substantial inter-cluster similarity.

- **Visualization:**
  Cluster visualization using PCA showed considerable overlap between clusters, reflecting GMM's inability to effectively differentiate between fraudulent and non-fraudulent transactions.
- **Discussion:**
  GMM was the least effective clustering technique for this dataset. The overlapping nature of the clusters and the absence of well-defined Gaussian distributions contributed to its poor performance. This result underscores the limitations of GMM when applied to datasets with complex distributions or high noise levels.

4.4.1 Comparative Analysis of Clustering Methods

Metric	K-Means	DBSCAN	GMM
Silhouette Score	0.313	**0.892**	0.046
Adjusted Rand Index (ARI)	N/A	-0.000	-0.000
Davies-Bouldin Index (DBI)	N/A	**0.970**	4.587
Noise Handling	None	**Good**	Poor

The results highlight key differences between the methods:

- **K-Means**: Performed moderately well but struggled with noise and outliers.
- **DBSCAN**: Achieved the best results with well-defined clusters and effective noise handling.
- **GMM**: Was the least effective due to poor separation and high overlap in the clusters.

4.6 Conclusion

Among the three clustering methods, **DBSCAN** demonstrated the best performance in clustering quality and noise handling, making it the most suitable choice for the credit card transaction dataset. **K-Means** performed adequately but was limited by its sensitivity to noise and outliers. **GMM**, while a flexible algorithm, failed to provide meaningful clustering for this dataset due to overlapping distributions and lack of distinct Gaussian components. Future research can explore hybrid methods or advanced preprocessing techniques to enhance clustering performance further.

CHAPTER 5: Conclusion

This study applied three clustering algorithms—DBSCAN, GMM, and K-means—to detect credit card fraud in an unsupervised setting. DBSCAN proved to be the most effective for identifying fraud in sparse, unbalanced datasets, as it distinguished fraudulent transactions as noise or outliers. GMM offered adaptability for complex fraud patterns, while K-means was a reliable but limited baseline due to its assumptions.

In conclusion, unsupervised clustering offers a viable solution for fraud detection without needing labeled data, increasing adaptability to emerging fraud patterns and reducing the dependence on supervised approaches.

CHAPTER 6: Future Extensions

7.1 Enhanced Parameter Selection and Hybrid Models

Future work could explore hybrid models combining DBSCAN's noise-handling strengths with GMM's adaptability to complex patterns. Automated or adaptive parameter selection methods for DBSCAN, such as adaptive grid searches, could further enhance fraud detection accuracy.

7.2 Real-Time and Adaptive Fraud Detection

Real-time fraud detection is essential as fraud patterns evolve quickly. Future research could focus on optimizing DBSCAN and GMM for incremental data processing, enabling real-time clustering and faster detection responses.

7.3 Developing Custom Evaluation Metrics

Creating new metrics tailored for fraud detection could yield more reliable evaluations, capturing the nuances of anomaly detection where traditional clustering metrics may fall short.

7.4 Exploring Advanced Anomaly Detection

Advanced anomaly detection algorithms, such as Isolation Forests or autoencoders, could complement clustering algorithms in fraud detection systems, enhancing accuracy and flexibility to capture complex patterns and evolving fraud tactics.