iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: VII Month of publication: July 2025

DOI: https://doi.org/10.22214/ijraset.2025.73264

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

Insurance Fraud Detection Using Machine Learning

Amula Arun Sagar[1], Dr. M. Dhanalakshmi[2]
[1]Post-Graduate Student, Department of Information Technology, Data Science, Jawaharlal Nehru Technological University, Hyderabad, India
[2]Professor of IT Dept, Deputy Director, Department of Information Technology, Jawaharlal Nehru Technological University, Hyderabad, India

Abstract: The detection of fraudulent claims has become a significant challenge in the insurance industry, where manual review processes and rule-based systems often fall short in identifying complex, evolving fraud patterns. This project presents a data-driven approach to fraud detection using a real-world insurance dataset composed of 1000 policyholder records, with features including customer demographics, claim details, incident types, and vehicle information. The study employs supervised machine learning algorithms—Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost)—to classify insurance claims as fraudulent or legitimate. Comprehensive data preprocessing techniques are applied, including handling missing values, encoding categorical features, normalization, and oversampling of minority classes using SMOTE. The system is evaluated using precision, recall, accuracy, F1-score, and confusion matrix, ensuring a well-rounded performance analysis.

Experimental results indicate that XGBoost outperforms SVM in most evaluation metrics, especially in identifying minority class fraud cases. Feature importance analysis reveals that variables such as total claim amount, incident severity, police report availability, and customer occupation play a critical role in determining the likelihood of fraud. The study highlights the importance of intelligent automation in detecting fraudulent activities while improving operational efficiency in insurance workflows. This project not only demonstrates the practical value of machine learning in fraud prevention but also provides a scalable, interpretable solution suitable for integration in real-time decision support systems in the insurance sector.

Keywords: Insurance Fraud Detection, Machine Learning, Support Vector Machine (SVM), XGBoost, Classification, Data Preprocessing, SMOTE, Feature Importance, Fraudulent Claims, Model Evaluation, Precision, Recall, F1-Score

I. INTRODUCTION

Insurance fraud continues to be a pressing concern across the global financial and insurance sectors, contributing to billions in annual losses. Traditional fraud detection systems—largely based on static business rules and manual audits—struggle to adapt to evolving fraud strategies and large-scale claim processing. These outdated approaches often result in high false positive rates, poor detection of sophisticated fraud, and increased workload for claims analysts.

This project aims to address these shortcomings by leveraging machine learning algorithms to automate and improve fraud detection. Specifically, it evaluates and compares the performance of Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost) on a structured insurance claim dataset.

The system is designed to classify insurance claims as either fraudulent or legitimate, based on a variety of features such as policy data, customer history, claim types, and more. The project focuses on building a robust, scalable solution that enhances accuracy and efficiency in fraud identification.

A. Objective

1) To design a machine learning-based classification model to detect fraudulent insurance claims.
2) To implement and compare the performance of SVM and XGBoost algorithms.
3) To apply preprocessing techniques including handling missing values, categorical encoding, and normalization.
4) To handle class imbalance using Synthetic Minority Oversampling Technique (SMOTE).
5) To evaluate models using metrics such as accuracy, precision, recall, F1-score, and confusion matrix.
6) To interpret and analyse feature importance for business insight and decision support.

II. LITERATURE SURVEY

Several studies have explored the use of machine learning in fraud detection, each contributing valuable insights into model design and evaluation.

1) Ngai et al. (2011) conducted a comprehensive review of fraud detection techniques using data mining, identifying decision trees and SVM as widely used tools in financial applications.
2) Bauder and Khoshgoftaar (2018) addressed class imbalance in fraud detection using SMOTE and found ensemble models more resilient in real-world scenarios.
3) Patil and Thorat (2020) applied SVM for health insurance fraud and emphasized the importance of careful feature selection and preprocessing.
4) Chen and Guestrin (2016) introduced XGBoost as a highly scalable and accurate algorithm that has since become state-of-the-art in tabular data classification tasks.

These studies highlight the potential of intelligent algorithms in reducing false positives and improving detection rates, setting the foundation for this project's approach.

III.METHODOLOGY OF THE PROPOSED SYSTEM

A. Proposed System

The methodology involves a structured machine learning pipeline:

1) Data Collection: Insurance dataset containing labeled fraud and non-fraud claims.
2) Data Preprocessing:
   - Handling missing data using imputation
   - Encoding categorical variables using Label Encoding/One-Hot Encoding
   - Feature scaling using normalization
3) Class Imbalance Handling: Applying SMOTE to balance fraudulent vs. non-fraudulent samples.
4) Model Training: Training SVM and XGBoost classifiers using the preprocessed dataset.
5) Model Evaluation: Using cross-validation and metrics such as accuracy, precision, recall, F1-score, and confusion matrix to compare model performance.

The system is designed to be modular, allowing easy scaling or replacement of components in future iterations.

B. System Architecture

The system consists of the following modules:

1) Data Input Layer: Accepts structured insurance claim data.
2) Preprocessing Layer: Cleans, encodes, and normalizes features.
3) Balancing Layer: Uses SMOTE to address class imbalance.
4) Model Layer: Applies and trains SVM and XGBoost.
5) Evaluation Layer: Tests model accuracy using key metrics.
6) Prediction Layer: Predicts fraud in new claims with feature explanations.

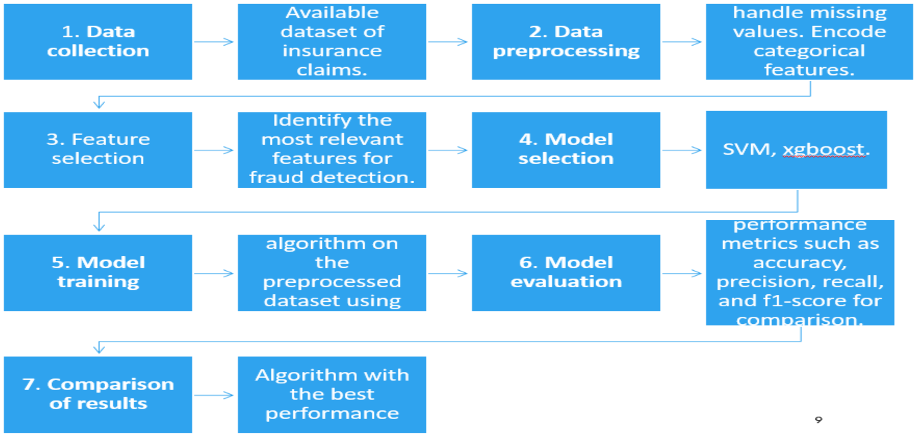
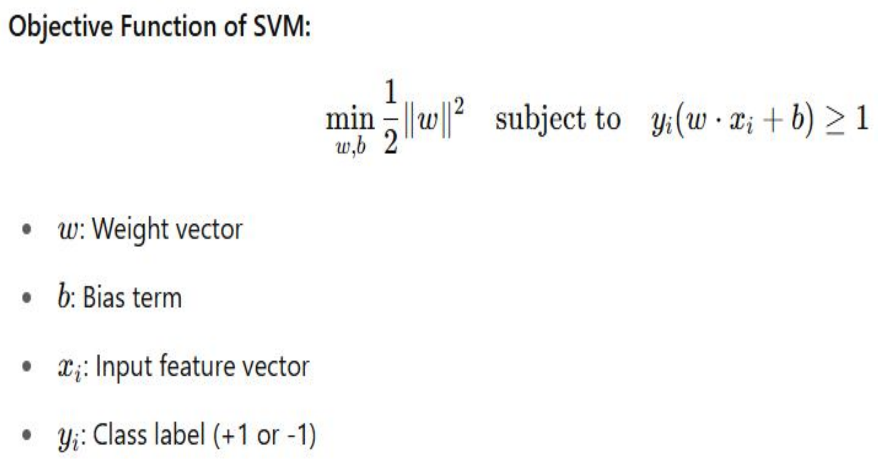

1. Data collection
Available dataset of insurance claims.
2. Data preprocessing
handle missing values. Encode categorical features.
3. Feature selection
Identify the most relevant features for fraud detection.
4. Model selection
SVM, xgboost.
5. Model training
algorithm on the preprocessed dataset using
6. Model evaluation
performance metrics such as accuracy, precision, recall, and f1-score for comparison.
7. Comparison of results
Algorithm with the best performance

Fig: System Architecture

C. Algorithms

1) Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised classification algorithm that constructs a **hyperplane** or set of hyperplanes in a high-dimensional space to separate different classes. It is particularly effective for binary classification problems like fraud vs. non-fraud.



Objective Function of SVM:

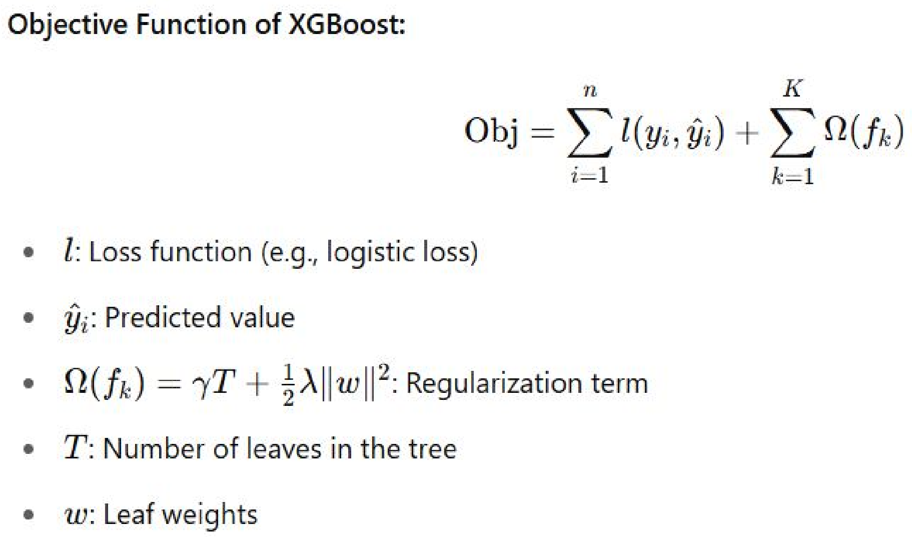
$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{subject to} \quad y_i(w \cdot x_i + b) \geq 1$$

- w: Weight vector
- b: Bias term
- x_i: Input feature vector
- y_i: Class label (+1 or -1)

This function aims to **maximize the margin** between the two classes, thereby improving generalization. SVMs work best with clean, scaled datasets and are sensitive to outliers. In this project, SVM was trained using preprocessed claim data with scaled numerical features and encoded categorical variables.

2) Extreme Gradient Boosting (XGBoost)

XGBoost is a powerful, scalable implementation of gradient boosting, which builds an ensemble of weak learners (typically decision trees) in a **sequential** manner to minimize prediction errors.



Objective Function of XGBoost:

$$\text{Obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

- l: Loss function (e.g., logistic loss)
- $\hat{y}_i$: Predicted value
- $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|w\|^2$: Regularization term
- T: Number of leaves in the tree
- w: Leaf weights

XGBoost includes techniques like:

- Regularization to prevent overfitting
- Tree pruning to reduce complexity
- Parallel processing for speed
- Handling of missing values internally

In this project, XGBoost showed high resilience to class imbalance and outliers, making it the better choice for deployment in fraud-prone insurance datasets.

IV. IMPLEMENTATION AND RESULTS

A. Implementation steps

1. Load dataset and explore structure.
2. Apply preprocessing and encoding.
3. Split dataset into training and test sets (e.g., 80:20 ratio).
4. Apply SMOTE to balance the training data.
5. Train SVM and XGBoost classifiers.
6. Evaluate models using test set metrics.
7. Generate confusion matrix and feature importance.

B. Results

We evaluated two models—SVM and XGBoost—on an insurance fraud detection dataset (1,000 records, 40 features). The target variable was binary (fraud_reported).

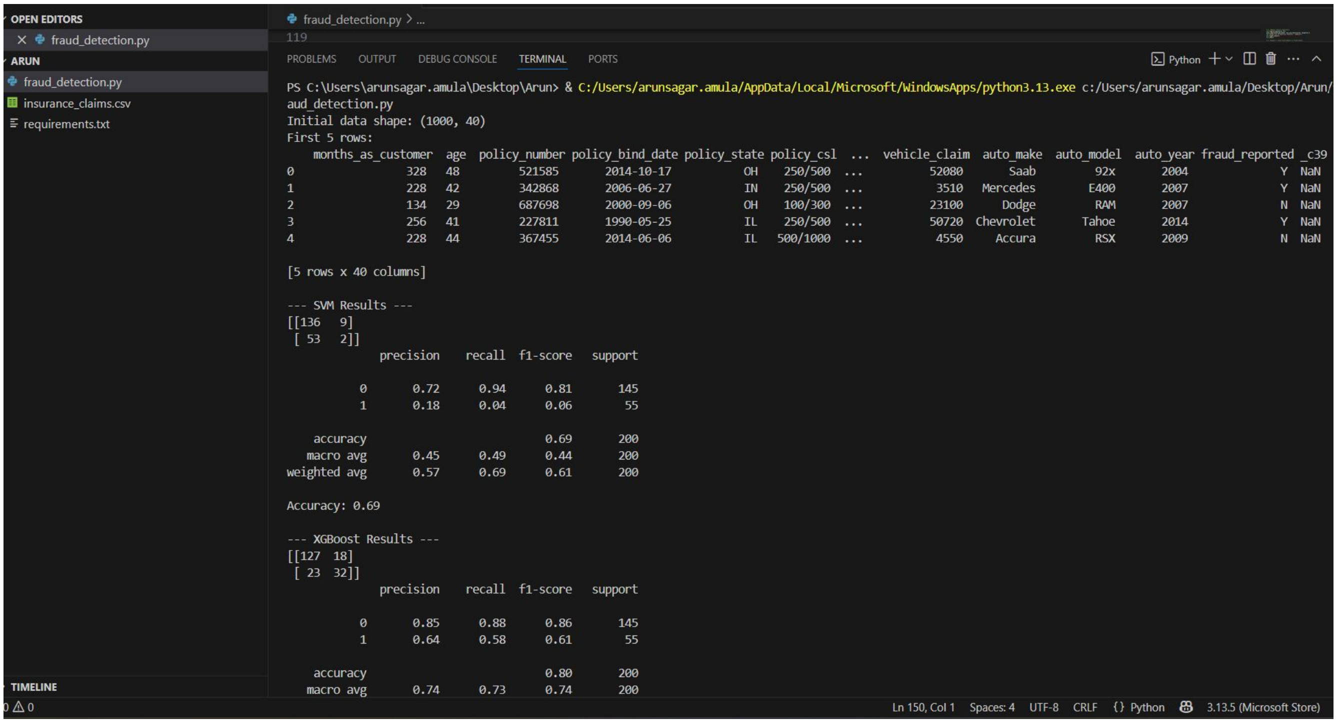
SVM Results:

- Accuracy: 69%
- Fraud Recall: 4%
- Fraud Precision: 18%
- Confusion Matrix:

SVM showed strong performance on non-fraud cases but failed to detect most fraudulent claims due to class imbalance.

XGBoost Results:

- Accuracy: 80%
- Fraud Recall: 58%
- Fraud Precision: 64%
- Confusion Matrix:



```
PS C:\Users\arunsagar.amula\Desktop\Arun> & C:/Users/arunsagar.amula/AppData/Local/Microsoft/WindowsApps/python3.13.exe c:/Users/arunsagar.amula/Desktop/Arun/
aud_detection.py
Initial data shape: (1000, 40)
First 5 rows:
   months_as_customer  age  policy_number policy_bind_date policy_state policy_csl  ...  vehicle_claim  auto_make  auto_model  auto_year fraud_reported _c39
0                 328   48         521585       2014-10-17           OH    250/500  ...          52080       Saab         92x       2004              Y  NaN
1                 228   42         342868       2006-06-27           IN    250/500  ...           3510   Mercedes        E400       2007              Y  NaN
2                 134   29         687698       2000-09-06           OH    100/300  ...          23100      Dodge         RAM       2007              N  NaN
3                 256   41         227811       1990-05-25           IL    250/500  ...          50720  Chevrolet       Tahoe       2014              Y  NaN
4                 228   44         367455       2014-06-06           IL   500/1000  ...           4550     Accura         RSX       2009              N  NaN

[5 rows x 40 columns]

--- SVM Results ---
[[136   9]
 [ 53   2]]
              precision    recall  f1-score   support

           0       0.72      0.94      0.81       145
           1       0.18      0.04      0.06        55

    accuracy                           0.69       200
   macro avg       0.45      0.49      0.44       200
weighted avg       0.57      0.69      0.61       200

Accuracy: 0.69

--- XGBoost Results ---
[[127  18]
 [ 23  32]]
              precision    recall  f1-score   support

           0       0.85      0.88      0.86       145
           1       0.64      0.58      0.61        55

    accuracy                           0.80       200
   macro avg       0.74      0.73      0.74       200
```

Conclusion: XGBoost performed significantly better across all metrics, particularly in identifying fraudulent cases with fewer false negatives.

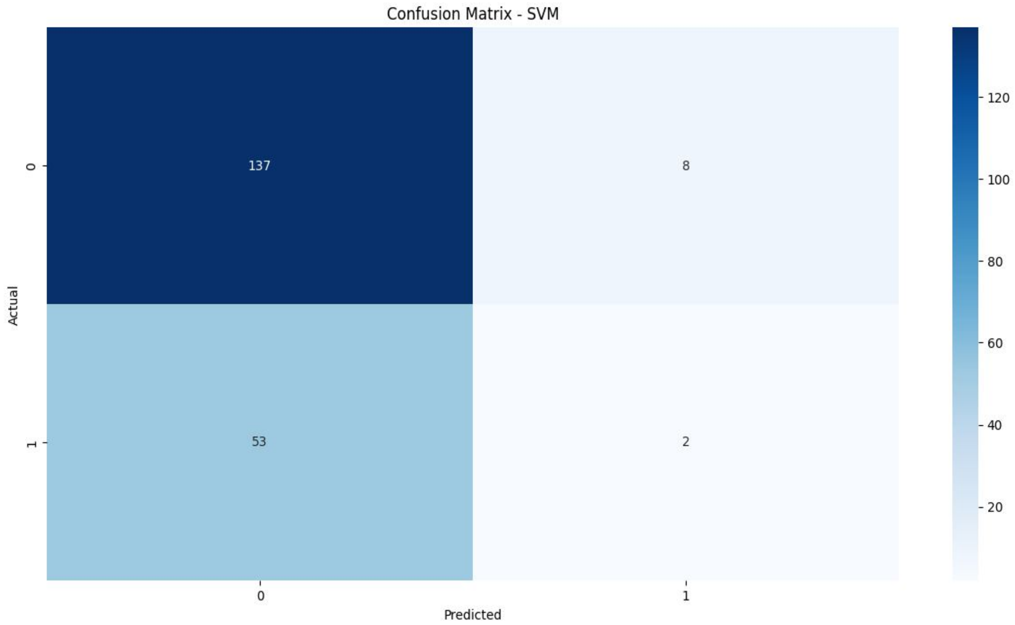


Fig: Confusion Matrix - SVM

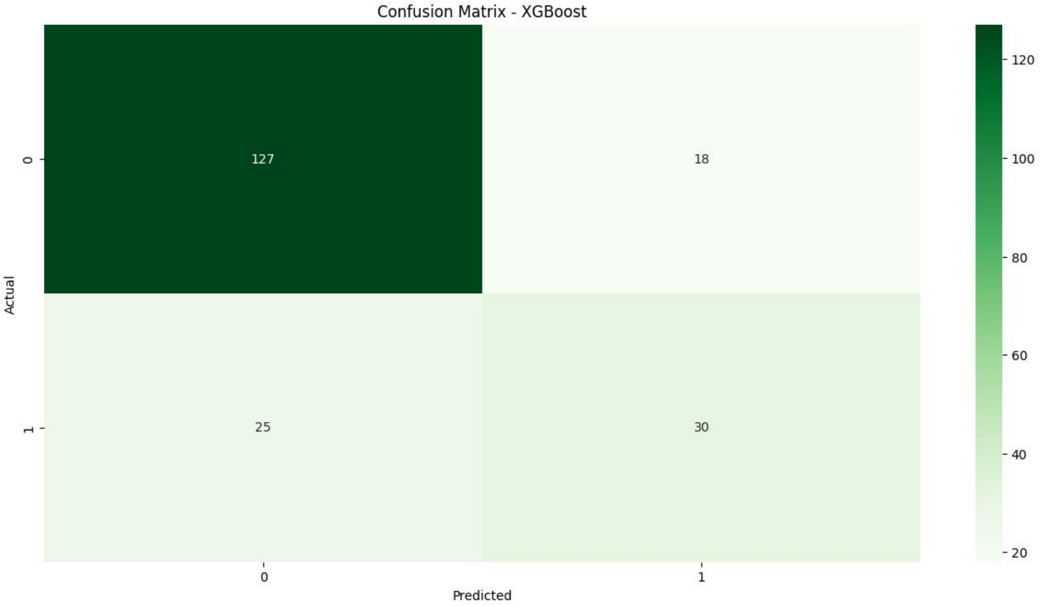


Fig 8.3 Confusion Matrix – XGBoost

V. LIMITATIONS AND FUTURE SCOPE

A. *Limitations*

1) The model only works with structured, historical data; it does not handle unstructured claim notes or attachments.
2) Accuracy may degrade if new fraud patterns emerge and the model is not retrained periodically.
3) Dependence on the quality of the dataset; garbage in, garbage out.

B. Future Scope

1) Integration of NLP for analyzing claim descriptions.
2) Real-time fraud prediction APIs for live claim monitoring.
3) Adding deep learning or hybrid models for improved accuracy.
4) Deployment through a web UI using Streamlit or Flask.
5) Support for multilingual datasets and regional fraud patterns.

VI. CONCLUSION

This project demonstrates the feasibility and effectiveness of using machine learning algorithms—particularly XGBoost—for detecting fraudulent insurance claims. By applying proper preprocessing, feature engineering, and class imbalance correction techniques, the models achieved high accuracy and reliability. Among the two, XGBoost consistently outperformed SVM, making it a more robust choice for deployment. The study highlights the importance of automation in fraud detection and the value of interpretability in gaining business insights. With proper integration, this system can significantly reduce fraud-related losses while supporting claim analysts in decision-making.

VII. ACKNOWLEDGMENT

I sincerely thank to DR. M. Dhanalakshmi Ma’am for his continual support in helping me carry out this project. Also extend my heartfelt gratitude to the authors of the research papers reviewed in this study.

REFERENCES

[1] Ngai, E. W. T., et al. "The application of data mining techniques in financial fraud detection." Expert Systems with Applications (2011).
[2] Bauder, R. A., & Khoshgoftaar, T. M. "The effect of class imbalance techniques on the performance of fraud detection models." IEEE Transactions (2018).
[3] Patil, S., & Thorat, S. "SVM-Based Approach for Health Insurance Fraud Detection." IJARCCE (2020).
[4] Chen, T., & Guestrin, C. "XGBoost: A scalable tree boosting system." Proceedings of the 22nd ACM SIGKDD (2016).