

# AI DRIVEN INCOME TAX FRAUD DETECTION

**Project Reference No.:** 48S\_BE\_2909

*College : Nagarjuna College of Engineering and Technology, Bengaluru*  
*Branch : Computer Science and Engineering*  
*Guide(S) : Prof. Narendra N*  
*Student(S) : Ms. Radhika R*  
*Ms. Roopa S Mattur*  
*Ms. Sahana V Iyer*  
*Ms. Yashaswini G E*

## **Keywords:**

Tax Fraud Detection, Machine Learning Algorithms, Supervised and Unsupervised Learning, Random Forest & SVM, Data Preprocessing and Visualization, Streamlit Interface.

## **Introduction:**

In the contemporary landscape of financial governance, the identification and prevention of tax fraud have become pivotal challenges for revenue authorities. The "Tax Fraud Detection Using Machine Learning" project represents a proactive and technologically advanced approach to address this issue. With the escalating complexity of financial transactions, traditional methods of tax monitoring have proven insufficient in detecting subtle patterns indicative of fraudulent activities.

This project leverages machine learning algorithms to delve into the vast datasets associated with registered and unregistered taxpayers. The user interface, implemented through the Streamlit framework, introduces a bifurcated system catering to REGISTERED and UNREGISTERED users. For the former, the project utilizes Permanent Account Number (PAN) as a unique identifier, employing data filtering and visualization techniques to unearth potential fraud trends. On the other hand, for UNREGISTERED users, historical tax-related information spanning a decade serves as input for machine learning models, enabling the prediction of potential tax fraud. Through this endeavor, we seek to enhance the integrity of financial systems, fostering fair and transparent fiscal.

## **Objectives:**

1. Analyze historical tax data for inconsistencies between reported and actual taxpayments.
2. It will classify taxpayers into compliant and fraudulent groups based on their financial behavior and tax records.
3. To check whether the model used, Support Vector Machines (SVM), RandomForest, and XGBoost, would perform well in the detection of tax fraud.
4. To create an interactive platform for visualizing tax data trends and model predictions, enabling intuitive decision-making for tax administrators.

## **Methodology:**

**1.Data Preprocessing:** Collect and clean the dataset, addressing missing values, outliers, and ensuring consistency. Feature engineering to extract relevant information and create meaningful variables. Standardize and normalize numerical features for uniformity in model training.

**2.User Authentication and Data Input:** Implement a secure Streamlit login page for user access. Differentiate between REGISTERED and UNREGISTERED users. For REGISTERED users, prompt input of PAN numbers; for UNREGISTERED users, gather historical tax-related data.

**3.Exploratory Data Analysis (EDA):** Conduct a thorough EDA to understand the distribution and characteristics of the data. Visualize data patterns, correlations, and potential outliers to inform feature selection.

**4.Model Selection and Training:** Choose machine learning algorithms suitable for the task, such as RandomForest, LogisticRegression, and Gradient Boosting. Train the model on labeled data for REGISTERED users and use unsupervised learning for UNREGISTERED users. Utilize the HUNOD method as a baseline for unsupervised outlier detection.

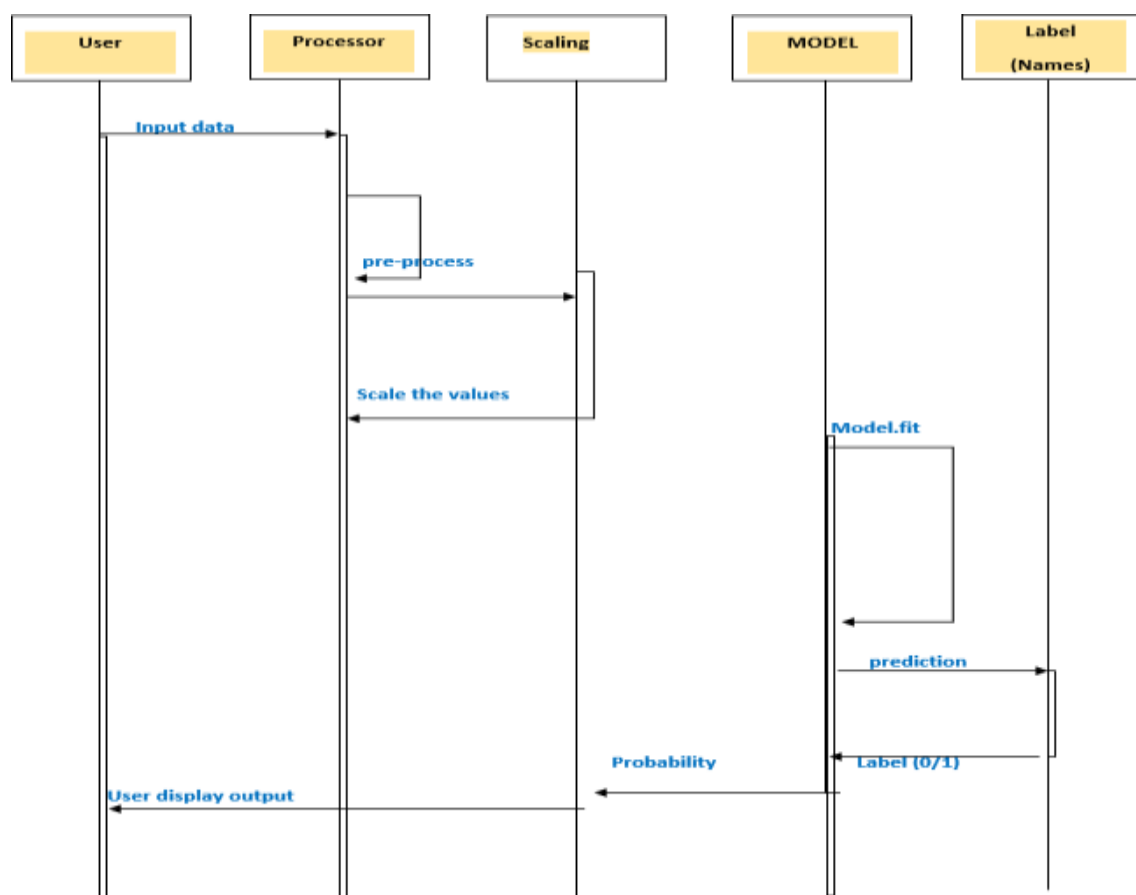
**5.Feature Importance and Explainability:** Assess feature importance using techniques like SHAP values or permutation importance. Enhance model interpretability through the integration of explainable-by-designsurrogatemodels.

**6.Performance Evaluation:** Employ appropriate evaluation metrics such as precision, recall, F1-score, andaccuracy.Assess the model's robustness using cross-validation techniques. Validate the model'spredictions against known cases of tax fraud.

**7.Optimization and Iterative Improvement:** Fine-tune hyperparameters based on model performance. Consider ensemble methodstoimprove overall prediction accuracy. Iterate on the methodology based on feedbackandemerging fraud patterns.

**8.Documentation and Reporting:** Document the entire process, including data preprocessing steps, model configurations, and evaluation results Generate.

Figure1:Sequence Diagram



## Result and Conclusion:

In conclusion, this comprehensive tax fraud detection initiative stands as a testament to the potential of leveraging cutting-edge technologies, data analysis, and user-friendly interfaces to strengthen financial monitoring systems. Its multifaceted approach, encompassing registered and unregistered users, historical data, machine learning algorithms, and performance metrics, positions it as a pioneering effort in the realm of tax compliance and fraud prevention.

The "Tax Fraud Detection Using Machine Learning" project signifies a significant advancement in fortifying tax compliance and ensuring revenue integrity. Employing sophisticated machine learning techniques, the system caters adeptly to both REGISTERED and UNREGISTERED users, delivering tailored solutions for fraud detection. The incorporation of supervised and unsupervised learning methods, notably the HUNOD method, highlights the project's adaptability to emerging fraud patterns.

The screenshot displays a web application for tax fraud detection. The interface is split into two columns. The left column contains the Income Tax Department logo and a 'Welcome Analyst' message. The right column contains a 'Citizen Details' table, an 'Income and Exempted Amount' section, and 'XGBOOST DIAGNOSIS' and 'XGBOOST PARAMETERS' sections with various numerical values.

| Citizen Details |                                 |
|-----------------|---------------------------------|
| citizen_id      | 123456                          |
| citizen_name    | John Doe                        |
| citizen_age     | 35                              |
| citizen_gender  | Male                            |
| citizen_address | 123 Main St, New York, NY 10001 |
| citizen_phone   | 9876543210                      |
| citizen_email   | john.doe@example.com            |

Income and Exempted Amount

XGBOOST DIAGNOSIS

NO FRAUD

XGBOOST PARAMETERS

agboost accuracy: 0.985

agboost precision: 0.97123456789012

agboost recall: 0.98901234567890

agboost f1 score: 0.98012345678901

### **Future Scope:**

1. Expansion to Other Tax Domains: Extend the system to detect fraud in other areas like GST, property tax, and sales tax, not just income tax.
2. Integration with Government Databases: Seamlessly connect the model with real-time government and financial databases for more accurate and up-to-date fraud detection.
3. Incorporation of Deep Learning Models: Introduce advanced deep learning techniques (like LSTM or CNN) to capture complex fraud patterns in large and unstructured datasets.
4. Cross-Border Fraud Detection: Adapt the system for international use, supporting fraud detection in cross-border taxation and multi-country tax systems.
5. Real-Time Fraud Alerts: Enhance the platform to provide real-time fraud alerts and automated reports to tax authorities for quicker response and enforcement.
6. Public and Private Sector Adoption: Offer the system as a service to financial institutions, auditing firms, and private companies to ensure compliance and prevent tax-related risks.