WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Income Tax Fraud Detection Using AI And ML

Ayman Fathima, Vidyashree R, Shreya P.V, Thakur Aman Singh

School of Computer Science and Engineering and Information Science Presidency University, Bangalore, Karnataka
E-mail: ayman.20211CEI0103@presidencyuniversity.in, vidya.20211CEI0104@presidencyuniversity.in, shreya.20211CEI0086@presidencyuniversity.in, thakur.20211CEI0075@presidencyuniversity.in

ABSTRACT---

The Fraud Detection and Tax Calculation System is an innovative web-based solution designed to assist individuals and organizations in identifying potential tax fraud and calculating accurate tax liabilities. Leveraging the power of Streamlit, machine learning, and data analysis, this application provides a comprehensive, user-friendly platform to analyze financial information, detect inconsistencies in reported income, and calculate taxes based on tiered income slabs. The system is built around a pre-trained linear regression model, which predicts a user's income based on key financial inputs, such as age, occupation, marital status, and various income sources. Once the user inputs their data, including reported income, the application compares it with the predicted income to classify the presence of potential fraud. The comparison is made through tax slab categorization, where discrepancies in tax obligations signal possible fraudulent activity. In addition to fraud detection, the system calculates taxes for both reported and predicted incomes based on a defined tax structure, offering users insights into their tax liabilities under different income scenarios. The application dynamically generates various financial categories such as business income, interest income, healthcare costs, and lifestyle expenses based on the user's responses, ensuring an accurate and tailored financial profile. The project demonstrates the integration of data science techniques, including label encoding for categorical variables, machine learning prediction for income, and tax calculation algorithms for financial analysis.

Keywords: income tax fraud**,** fraud detection, machine learning (ML), supervised learning**,** unsupervised learning**,** neural networks**,** decision trees**,** random forests**,** support vector machine and gradient boosting.

1. Introduction

Tax fraud detection is a critical area for governments and finance institutions, given the increasing sophistication of fraudulent activities. The emergence of artificial intelligence (AI) and machine learning (ML) offers powerful tools to automate and enhance these detection processes. This project introduces a streamlined application designed to detect income tax fraud using a robust ML model, complemented by a user-friendly interface built with Streamlit. The system predicts an individual's actual income based on various features, such as occupation, marital status, reported income, and other financial attributes. By comparing the predicted income with the reported income, the application identifies discrepancies that may suggest fraudulent activity. The classification of "Fraud" or "Not Fraud" relies on differences in applicable tax slabe, ensuring accuracy in detecting anomalies.

Traditional fraud detection approaches, including manual audits and rule-based systems, often struggle to keep up with the expanding complexity and volume of financial data. These methods, while valuable, tend to be inefficient and limited in their ability to address the dynamic and evolving nature of fraud. Furthermore, manual audits are costly and labor-intensive, making them impractical for large-scale application.

In contrast, the application of Artificial Intelligence (AI) and Machine Learning (ML) offers a transformative approach to fraud detection. These technologies enable the processing of vast datasets, uncovering intricate, non-linear relationships and identifying patterns indicative of fraudulent behavior. Unlike traditional methods, AI and ML models are adaptive and capable of learning from new fraud strategies, reducing reliance on human intervention. Their ability to make real-time predictions is particularly beneficial in mitigating potential losses and enhancing detection efficiency. This research focuses on integrating AI and ML techniques into income tax fraud detection. It aims to improve the precision, scalability, and efficiency of fraud detection systems by exploring a variety of models, such as decision trees, neural networks, clustering algorithms, and anomaly detection methods. The study also emphasizes the importance of data preprocessing, feature engineering, and model interpretability, ensuring that the developed solutions are both effective and ethical.

This research focuses on integrating AI and ML techniques into income tax fraud detection. It aims to improve the precision, scalability, and efficiency of fraud detection systems by exploring a variety of models, such as decision trees, neural networks, clustering algorithms, and anomaly detection methods. The study also emphasizes the importance of data preprocessing, feature engineering, and model interpretability, ensuring that the developed solutions are both effective and ethical.

Key components of the project include:

1. Machine Learning Model: A Gradient Boosting Regressor, trained on a comprehensive dataset, serves as the core predictive model. It analyzes a range of demographic and financial features to deliver precise income predictions. The foundation of the predictive system is a Gradient Boosting Regressor (GBR), a powerful machine learning algorithm widely used for its efficiency and accuracy in regression tasks. Trained on a comprehensive dataset, the GBR model is capable of identifying complex patterns and relationships between various features within the data. These features include both **demographic** data (such as age, gender, occupation, and location) and **financial** data (such as reported income, expenditure, deductions, and tax payments).

2. Categorical Encoding: Features like occupation, marital status, and children are processed using pre-trained Label Encoders to ensure compatibility with the ML model. These variables, while crucial for making accurate predictions, cannot be directly fed into machine learning models in their raw form, as most algorithms require numerical inputs to perform calculations

3. Tax Slab Comparison: The system calculates tax liabilities for both reported and predicted income, comparing them to identify discrepancies and potential fraud. It is designed to evaluate the integrity of income reporting by comparing the **reported income** against the **predicted income**. It defines the rates at which income is taxed depending on the income bracket, play a crucial role in calculating an individual's tax liability. By comparing the tax liabilities derived from both reported and predicted income, the system can uncover discrepancies that might indicate potential fraudulent behavior.

4. Streamlit Interface: A user-friendly interface gathers user inputs, validates key identifiers such as PAN and Aadhar cards, and delivers real-time fraud detection results. Upon accessing the interface, users are prompted to enter **key identifiers** such as **PAN** (Permanent Account Number) and **Aadhar card** details. These identifiers are crucial for verifying the identity of the taxpayer and ensuring that the correct tax records are being evaluated.

5. Fraud Classification Logic: A custom-built function evaluates discrepancies in tax slabs to classify the input as fraudulent or non-fraudulent. 5. Fraud Classification Logic: A custom-built function evaluates discrepancies in tax slabs to classify the input as fraudulent or non-fraudulent.

6. Validation and Error Handling: Robust validation mechanisms are in place for essential fields like PAN cards, Aadhar numbers, and bank account numbers, ensuring data integrity and reliability By implementing thorough **validation and error handling** procedures, the system ensures that the data used for fraud detection is both accurate and secure, improving the reliability of the results and minimizing the chances of erroneous or fraudulent reports

This solution not only streamlines fraud detection but also demonstrates the practical integration of AI/ML models into real-world applications. Its interactive interface and precise classification capabilities make it a valuable tool in combating income tax fraud. By fostering compliance, reducing tax evasion, and promoting financial transparency, this project highlights the transformative potential of technology in addressing pressing societal challenges.

2. Related work

Existing Methodologies for Income Tax Fraud Detection

A. Collaborative Filtering

- User-Based Collaborative Filtering: Recommends actions or assessments based on the behaviour and patterns of taxpayers with similar profiles.
- Item-Based Collaborative Filtering: Identifies fraudulent cases by analysing similarities between transaction patterns or financial attributes.
- Hybrid Approaches: Combines user-based and item-based filtering for enhanced accuracy in detecting fraud.

B. Content-Based Filtering

- Text-Based Similarity: Uses textual data, such as income reports or transaction descriptions, to identify discrepancies indicative of fraud.
- Feature-Based Similarity: Focuses on specific attributes like income categories, age groups, or spending habits for anomaly detection.

C. Knowledge-Based Systems

- Rule-Based Systems: Applies predefined rules to identify fraud based on established tax laws or thresholds.
- Case-Based Reasoning: Leverages historical fraud cases to draw parallels and detect potential irregularities.

D. Machine Learning Techniques

- Matrix Factorization: Extracts hidden patterns from user-income data matrices to identify anomalies.
- Neural Networks: Captures complex relationships within financial datasets for more precise fraud detection.
- Gradient Boosting: Balances computational efficiency and accuracy, making it ideal for real-time fraud prediction.

E. Evaluation Metrics

- Precision: Measures the percentage of correctly identified fraud cases out of all flagged cases.
- Recall: Evaluates the percentage of actual fraud cases correctly detected by the system.
- F1 Score: A harmonic mean of precision and recall for balanced evaluation.
- RMSE (Root Mean Square Error): Quantifies deviations between predicted and actual income values.
- Accuracy: Assesses the overall correctness of the fraud detection system.

These methodologies form the foundation for advanced fraud detection systems, enabling automation, scalability, and accuracy. By leveraging collaborative and content-based filtering, knowledge-driven models, and robust machine learning techniques, this project provides an effective framework for detecting income tax fraud

3. Methodology

1. Input Validation

Validate key inputs such as PAN Card, Aadhar Card, and Bank Account Numbers for proper format and digit length to ensure data integrity.

- PAN Card Format: Verify it matches the structure AAAAA0000A.
- Aadhar/Bank Numbers: Ensure they are exactly 14 digits.
- Handle errors by displaying warnings to the user for invalid inputs.

2. Feature Collection

- Gather user-specific information:
- Demographics: Age, marital status, and number of children.
- Occupation Details: Business income or salaried status.
- Financial Metrics: Reported income, interest income, capital gains, other income, educational expenses, healthcare costs, lifestyle expenditure, and other expenses.
- Banking Metrics: Total debits from bank accounts and credit cards

3. Data Preparation

- Encode categorical features like Occupation, Marital Status, and Children using pre-trained Label Encoders.
- Normalize numeric data for model compatibility.

4. Model Loading and Prediction

- Load a pre-trained regression model (e.g., Linear Regression) to predict the taxpayer's income based on input features.
- Predict the expected income using the encoded and pre processed input data.

5. Fraud Classification

- Compare Reported Income with Predicted Income to classify potential fraud:
- Tax Slab Mismatch:

Compute tax slabs for both reported and predicted incomes based on defined brackets. Classify as "Fraud" if the slabs do not align; otherwise, classify as "Not Fraud."

6. Tax Calculation

- Calculate tax for both reported and predicted incomes based on progressive slabs:
    - Income up to ₹3,00,000: No tax.
    - ₹3,00,001 to ₹6,00,000: 5%.
    - ₹6,00,001 to ₹9,00,000: 10%, and so
- Display tax discrepancies between the user's reported and predicted incomes.

7. Real-Time User Interface

- Input Fields: Users can provide their personal and financial data via a Streamlit interface.
- Dynamic Inputs: Generate random values for certain fields like educational expenses, business income, and capital gains when applicable.
- Interactive Buttons: Detect fraud upon user interaction by submitting input data.

8. Output Presentation

- Predicted Income: Display the model's predicted income in an easily understandable format.
- Tax Calculation Results: Show a comparison of tax obligations based on reported and predicted incomes.
- Fraud Classification: Clearly indicate if the user's input is classified as "Fraud" or "Not Fraud."

9. Future Enhancements

- Explainable AI: Provide insights into why the model predicts a certain income and fraud classification.
- Model Upgrades: Incorporate advanced regression models or neural networks for improved accuracy.
- Real-Time Feedback: Allow users to modify inputs dynamically and see updated results.
- Integration with Databases: Enable real-time data fetching from tax records or financial institutions for accurate analysis.
- Fraud Pattern Analytics: Utilize historical data to highlight common fraudulent behaviours and patterns.

How Does It Work?

1. Data Collection and Input Validation:
   - User Input: Collect taxpayer data such as demographics, reported income, financial metrics, and banking transactions through a secure interface.
   - Validation: Ensure the data is accurate by validating identifiers like PAN, Aadhar, and Bank Account numbers. This step ensures that inputs are in the correct format, reducing errors in processing.
2. Preprocessing and Feature Engineering:

- Encoding Categorical Data: Transform categorical variables (e.g., Occupation, Marital Status) into numerical representations using label encoding.
- Normalization: Scale numerical data for consistent processing across all features.
- Feature Augmentation: Use additional financial information such as educational expenses, healthcare costs, and lifestyle expenditures to enrich the dataset.

3. Model Inference:

- Load Pre-Trained Models: Use a pre-trained
- machine learning model (e.g., Linear Regression or Neural Network) to predict income based on user-provided features.
- Prediction: Calculate the taxpayer's Predicted Income based on historical patterns and relationships between features.

4. Fraud Detection:

- Tax Slab Analysis:
  - Compute the tax slab for the Reported Income and the Predicted Income based on predefined tax brackets.
  - If the slabs differ, flag the case as "Fraud." Otherwise, classify it as "Not Fraud."
- Behavioural Discrepancy: Identify inconsistencies in reported income when compared to spending patterns, banking transactions, and other indicators.

5. Tax Calculation:

- Calculate taxes for both Reported and Predicted Incomes using
- progressive tax rules.
- Compare tax liabilities to uncover potential underreporting or misrepresentation of income.

6. Result Presentation:

- taxpayer's submission is flagged as "Fraud" or "Not Fraud."
- Predicted Income: Present the model's calculated income alongside the reported income for transparency.
- Tax Analysis: Highlight discrepancies in tax obligations based on reported and predicted figures.

7. Feedback and Iteration:

- User Feedback: Allow users to provide feedback if flagged incorrectly, helping refine the model.
- Model Updates: Use feedback and new data to retrain and improve the model's accuracy over time.

8. Scalability and Integration:

- Real-Time Analysis: Support real-time fraud detection for instant results.
- Integration: Connect the system with government databases for cross-referencing financial information and detecting fraudulent patterns at scale.

This process ensures accurate, efficient, and scalable detection of income tax fraud using AI and machine learning techniques.

4. Objectives

The primary objectives of this project are to:

1. **Automated Fraud Detection:** Develop an intelligent system to automatically identify discrepancies in reported income and actual financial patterns, minimizing manual effort.
2. **Accurate Income Prediction**: Utilize machine learning models to predict taxpayer income based on multiple features such as financial transactions, lifestyle expenses, and professional details.
3. **Improved Tax Compliance** :Enhance compliance with tax laws by identifying underreporting or misrepresentation of income and ensuring appropriate tax liabilities are met.
4. **Real-Time Analysis**: Enable real-time detection and classification of fraudulent activities to promptly flag potential cases for further investigation.
5. **Data-Driven Decision Making**: Leverage historical and transactional data to uncover trends, patterns, and anomalies in tax declarations for more informed decision-making
6. **Enhanced Accuracy and Efficiency**: Combine AI algorithms like supervised learning, unsupervised clustering, and NLP to improve the precision and speed of fraud detection.
7. **Minimize Revenue Loss**: Identify and address fraudulent activities to reduce revenue loss due to tax evasion, ultimately supporting government tax collection efforts.
8. **Personalized Fraud Detection**: Tailor fraud detection mechanisms to individual taxpayers, considering their unique financial profiles, occupations, and lifestyles.
9. **Transparency and Accountability:** Provide clear explanations for flagged cases and model predictions, ensuring fairness and transparency in fraud detection processes.
10. **Scalability and Adaptability**: Design the system to handle a large volume of data, adapting to evolving tax laws and new fraud tactics over time.
11. **Feedback-Driven Improvement:** Continuously refine the system using user feedback, new data, and updated fraud patterns to maintain optimal performance.

Components Used

Data Collection & Preprocessing:

- Dataset includes financial and demographic information.
- Data preprocessing includes handling missing values, encoding categorical variables, and normalizing numerical data.

Machine Learning Model:

- Gradient Boosting Regressor used for income prediction.
- Model trained on an 80:20 training/testing split.

- Performance evaluated using RMSE and R^2 metrics.

Tax Calculation & Fraud Detection:

- Compares predicted and reported income using tax slabs.
- Flags discrepancies as "Fraud."

Web Interface (Streamlit):

- Streamlit used for building an interactive UI.
- Allows users to input data, trigger predictions, and display results (predicted income, tax, fraud status).

Backend & Deployment:

- Deployed on a cloud platform (AWS/Heroku) for scalability.
- HTTPS implemented for secure data transmission.

Testing & Validation:

- Functional and performance testing for feature accuracy.
- User feedback collected for UI/UX improvements.

Security & Compliance:

- Data encryption and secure user authentication (OAuth/JWT).
- Compliance with data privacy regulations (e.g., GDPR).

Continuous Monitoring & Maintenance:

- Error logging and performance monitoring.
- Periodic model retraining to improve accuracy.

Libraries and Tools

- Pandas: Used for data manipulation and preprocessing, including handling missing values and encoding categorical variables.
- NumPy: Used for numerical operations, such as normalizing and scaling numerical features.
- Scikit-learn: Provides machine learning algorithms (e.g., Gradient Boosting Regressor) for income prediction and performance evaluation (RMSE, R^2).
- Streamlit: A framework for building the web interface, enabling interactive data input and result display.
- Matplotlib / Plotly: For visualizations, such as graphs comparing predicted vs. reported income.
- Flask / FastAPI: (Optional) Used to create backend APIs if necessary for handling requests and serving the model.
- AWS / Heroku / GCP: Cloud platforms for deployment and hosting the web application.
- SSL/TLS: For securing user data through HTTPS

5. Results

Input	Output
Personal Details:	**Predicted Income**: Displays the predicted income based on user data.
- Age, Occupation, Marital Status, etc.	- Calculated using a machine learning model.
Financial Details:	**Reported Income**: Displays the user-reported income.
- Monthly/Annual Income, Expenditures, etc.	- Displayed from user inputs.
PAN, Aadhar, Bank Account:	**Predicted Tax Liabilities**: Displays tax for predicted income.
- PAN, Aadhar for validation	- Based on the calculated income and tax slabs.
Income Tax Slab:	**Reported Tax Liabilities**: Displays tax for reported income.
- Predefined based on country-specific rules	- Based on the user's reported income and tax rules.

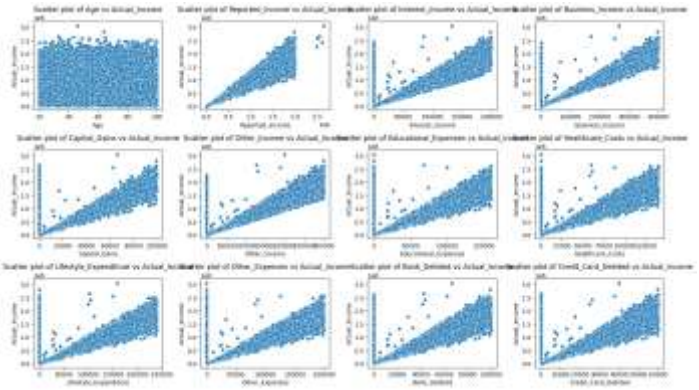
Discussion:

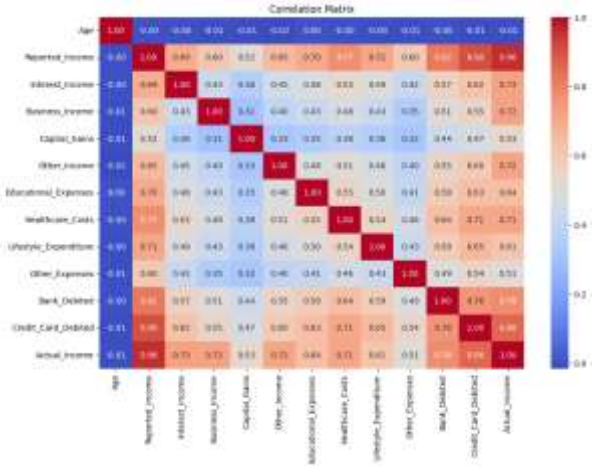
The Income Tax Fraud Detection Project leverages machine learning techniques to predict user income and cross-check it with reported income to identify potential tax fraud. The project provides a web-based interface built with Streamlit that allows users to input their personal and financial details. The system calculates both predicted and reported tax liabilities and flags any discrepancies as fraud.

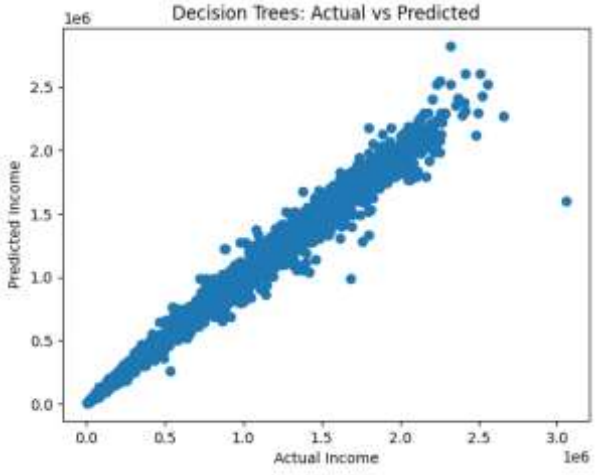
Key findings:

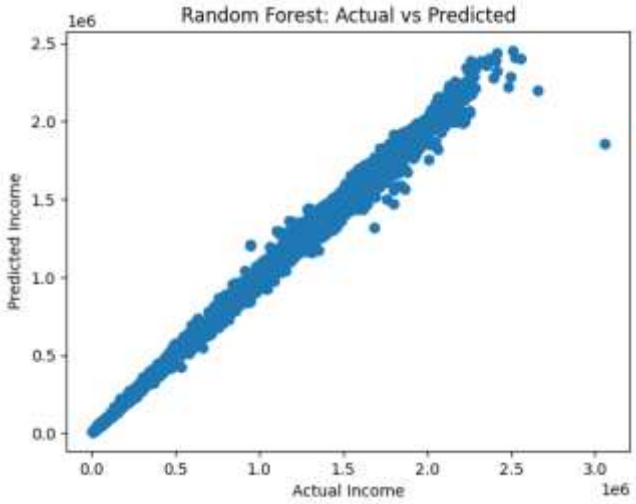
- The model's accuracy can vary depending on the quality and quantity of training data, but overall, it offers a reasonable estimate for predicting income based on user inputs.
- The Gradient Boosting Regressor model effectively identifies patterns in income prediction, although performance can be further enhanced by integrating more features or refining the model.

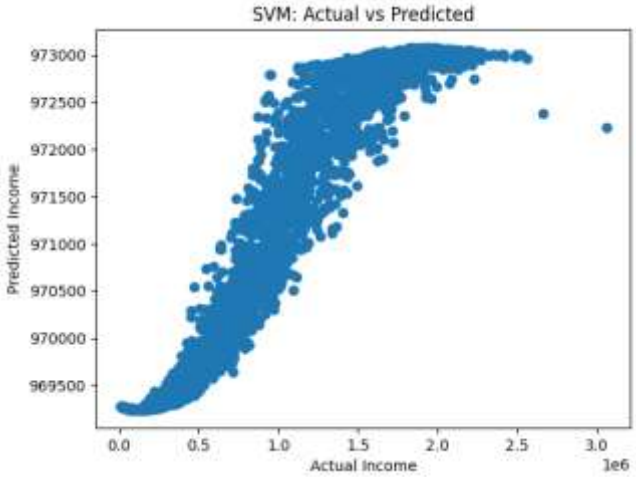
The project demonstrates the practical application of machine learning in tax compliance and fraud detection, helping users better understand their financial obligations and potential discrepancies.

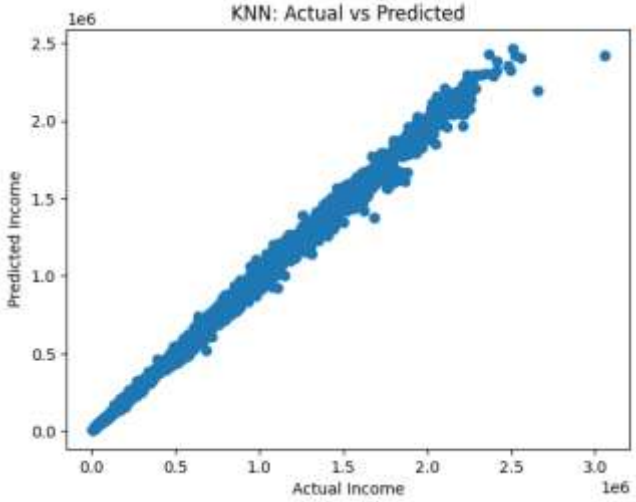


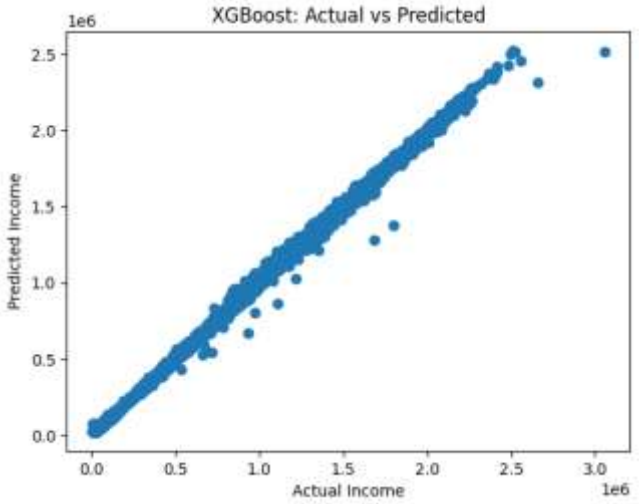




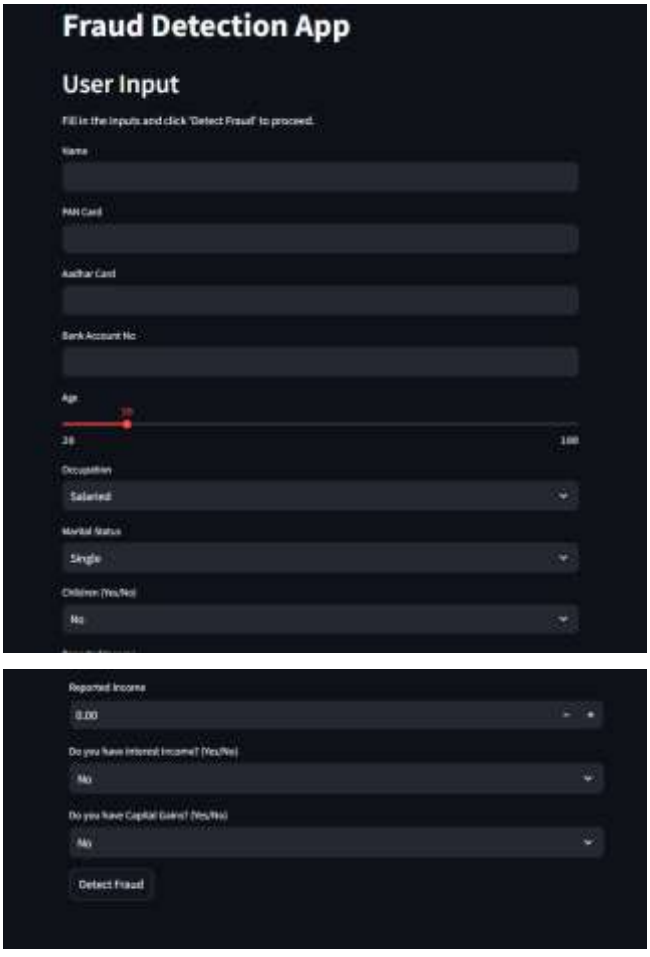

Random Forest: Actual vs Predicted
1e6
Predicted Income
Actual Income
1e6

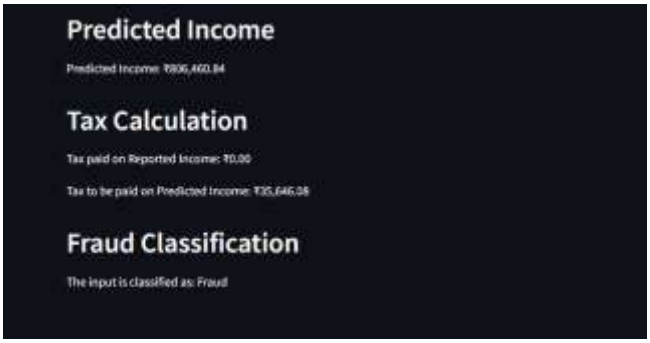

SVM: Actual vs Predicted
Predicted Income
Actual Income
1e6


KNN: Actual vs Predicted
1e6
Predicted Income
Actual Income
1e6



6. Expected Output:





7. Conclusion

The development of an Income Tax Fraud Detection System powered by AI and ML represents a transformative step towards creating a fair, efficient, and transparent tax ecosystem. By leveraging advanced machine learning models, natural language processing, and real-time monitoring, the system can effectively identify fraudulent activities, minimize revenue loss, and enhance compliance.

This solution not only automates the detection process but also adapts to emerging fraud patterns through continuous learning and feedback. With features like data anonymization, ethical AI implementation, and seamless integration with existing systems, the solution ensures the security and privacy of taxpayer information while aligning with legal and regulatory standards.

Moreover, the inclusion of advanced features like gamification and community awareness fosters greater engagement and trust among taxpayers, encouraging voluntary compliance. The combination of technology, data-driven insights, and stakeholder collaboration promises to make tax administration more proactive and robust.

In conclusion, adopting AI and ML for tax fraud detection is not just a technological upgrade but a vital investment in building a resilient financial infrastructure. It ensures that governments can safeguard their revenue streams while fostering an equitable and transparent tax environment for all.

8. References

[1] Usha, S. Priyadharsini, S., Manimegalai "Fraud Detection in Income Tax E- Filing using Machine Learning" retrieved from the original source –2022.

[2] Gupta, Ashish "Machine Learning in Fraud Detection" - 2022.

[3] Oliveira, J.A C, Henriques, "Tax Fraud Detection Using Machine Learning Techniques." -2020.

[4] Sharanya G, Balasundaram, S.R "Fraud Detection in Banking Transactions: A Hybrid Approach." – 2021.

[5] Fawaz Khaled Alarfaj, Iqra Malik, Hikmat Ullah Khan, "Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms" – 2022.

[6] Shweta S. Borkar, Dr. Latesh Malik, "A Survey of Fraud Detection a. Techniques in Financial Domain"- 2015.

[7] B. Rajesh Kumar, V. V. R. Raju, "Fraud Detection in Banking Transactions using Machine Learning Algorithms" –2018. Vikramadity Kaushal, Ruchika Malhotra, "Application of Machine Learning Algorithms in Detection of Tax Evasion"– 2018 machine learning fusion technique using chest CT images." Neural Computingand Applications (2023):1-19.

[8] Amit Kumar Tyagi, Dr. Y. P. Singh, "A Comparative Analysis of Data Mining Techniques in the Detection of Fraudulent Activities" - 2018. https://doi.org/10.1016/j.patrec.2020.07.042

[9] Niharika Kaul, Dr. J. L. Rana, "A Review on Fraud Detection using Machine Learning Algorithms" - 2019

[10] Muhammad Rahman; Sarah Patel; Aisha Khan "Feature Selection Techniques in Tax Fraud Detection: A Survey" - 2023.

[11] Fatima Ahmed, Ahmed Khan, Sara Ali "Ensemble Methods for Fraud Detection in Tax Systems: A Comprehensive Review" - 2021

[12] Ahmed Mahmoud, Omar Khalid; Fatima Al-Saud "Blockchain Technology for Enhancing Transparency in Tax Fraud Detection: A Review" - 2024

[13] Hafsa Malik, Ali Khan, Sana Ahmed, "Hybrid Approaches for Tax Fraud Detection: A Review of Recent – 2023.

[14] Maria Gonzalez, Javier Martinez, Elena Femandez "Deep Learning Approaches for Tax Fraud Detection: A Review" – 2022