

A Project Report

On

**“An Income Tax Fraud Detection Idea Using AI & ML”**

Batch Details

|  |  |  |
| --- | --- | --- |
| Sl. No. | Roll Number | Student Name |
| 1 | 20211CST0063 | Uday G |
| 2 | 20211CST0035 | Ranganath R |
| 3 | 20211CST0104 | Rohith M |
| 4 | 20211CST0099 | Abhishek V |
| 5 | 20211CST0048 | Karthik Kumar SM |

**School of Computer Science,**

**Presidency University, Bengaluru.**

Under the guidance of,

Dr. **Saira Banu Atham**

School of Computer Science,

Presidency University, Bengaluru

**CONTENTS**

1. Introduction about Project
2. Literature Review
3. Objectives

## Methodology

1. Timeline for Execution of Project
2. Expected Outcomes
3. Conclusion
4. References

**INTRODUCTION**

Tax fraud is a major issue that leads to significant revenue losses for governments. Income tax fraud occurs when individuals or businesses intentionally misreport their income, expenses, or deductions to reduce their tax liability. The complexity and volume of tax data make it challenging for authorities to detect such fraud using traditional methods like manual audits and rule-based systems, which are often inefficient and inadequate against evolving fraudulent schemes.

With advancements in artificial intelligence (AI) and machine learning (ML), tax authorities now have powerful tools to improve fraud detection. AI and ML can analyze large datasets, identify hidden patterns, and detect anomalies more effectively than traditional methods. These systems can flag suspicious activities, such as unusual deductions or income changes, while continuously learning from past fraud cases to adapt to new tactics.

Our project aims to evaluate various ML models and select the best one for tax fraud detection based on multiple performance metrics. By leveraging advanced AI/ML techniques, we aim to enhance the accuracy, efficiency, and adaptability of fraud detection systems, ensuring better protection of public funds and a more equitable tax system.

**LITERATURE REVIEW**

The paper, *"Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review,"* systematically reviews various machine learning (ML) techniques used for detecting fraud in financial statements. Below is a literature review focusing on eight machine learning techniques used in this study, their advantages, and disadvantages:

### **1. Support Vector Machines (SVM)**

**Advantages:**

* Effective in high-dimensional spaces and works well for binary classification problems, making it suitable for fraud detection.
* Can handle non-linear data efficiently by using kernel tricks like Radial Basis Function (RBF).

**Disadvantages:**

* Sensitive to the choice of kernel and tuning parameters.
* Computationally expensive, especially for large datasets.

### **2. Self-Organizing Maps (SOM) with K-Means Clustering**

**Advantages:**

* SOM is useful for visualizing high-dimensional data and is effective for clustering similar instances of financial data, making it suitable for fraud detection tasks.
* When combined with K-Means clustering, SOM can refine the cluster boundaries and produce more accurate classifications of fraudulent statements.

**Disadvantages:**

* SOM struggles with defining clear cluster boundaries on its own, which is why K-Means is needed to improve precision.
* The performance of SOM can be affected by the input order of samples and the initialization values of the nodes.

### **3. Random Forest (RF)**

**Advantages:**

* Reduces overfitting by aggregating the results of multiple decision trees.
* Provides high accuracy and robustness, even when dealing with imbalanced datasets (common in fraud detection).

**Disadvantages:**

* Less interpretable compared to single decision trees.
* Requires more computational resources due to the need to build and combine multiple trees.

### **4. Logistic Regression (LR)**

**Advantages:**

* Easy to implement and interpret as it provides probabilistic estimates of fraud.
* Works well with binary classification problems and is widely used in financial fraud detection.

**Disadvantages:**

* Assumes a linear relationship between independent and dependent variables, limiting its effectiveness with complex, non-linear data.
* Sensitive to multicollinearity among features, which may lead to misleading results.

### **5. Naive Bayes (NB)**

**Advantages:**

* Simple and fast to train, making it suitable for large datasets in fraud detection.
* Performs well with small datasets despite its simplicity.

**Disadvantages:**

* Assumes conditional independence between features, which is often not the case in real-world data, leading to lower accuracy.
* It can struggle with imbalanced datasets.

### **6. Artificial Neural Networks (ANN)**

**Advantages:**

* Capable of learning complex patterns in data, which is useful for detecting sophisticated fraud schemes.
* Can be used with both structured and unstructured data, such as text from financial reports.

**Disadvantages:**

* Requires a large amount of data and computational resources for training.
* The "black box" nature of ANNs makes them difficult to interpret and justify in decision-making contexts

### **7. K-Nearest Neighbors (KNN)**

**Advantages:**

* Simple to understand and implement and does not require explicit training.
* Performs well in situations where the decision boundary is highly non-linear.

**Disadvantages:**

* Sensitive to the choice of k (number of neighbors) and the metric used to measure distance.
* Computationally expensive as it requires calculating the distance to every training example for each test instance.

### **8. Ensemble Methods (Boosting, Bagging, Stacking)**

**Advantages:**

* Combines multiple models to improve overall performance and reduce variance or bias.
* Boosting, such as AdaBoost, reduces bias, and Random Forest (a bagging method) reduces variance, making these techniques highly accurate.

**Disadvantages:**

* More complex to implement and requires greater computational resources.
* The combined models are less interpretable, making it difficult to understand the decision process.

These machine learning techniques offer a range of advantages and limitations when applied to fraud detection in financial statements. While methods like Random Forest and Ensemble Techniques offer high accuracy, they are resource-intensive and less interpretable. Simpler models like Logistic Regression and Decision Trees offer better interpretability but may not perform as well with complex or non-linear data.

**OBJECTIVES**

Based on the gaps identified in the literature review, the following objectives have been set for this project:

1. **Evaluation and Selection of Best Model for Fraud Detection Using Various Performance Metrics:** Selecting the most appropriate model for fraud detection involves evaluating multiple AI/ML techniques to determine which provides the best balance between detecting fraudulent activities and minimizing errors. This process involves assessing various models such as **decision trees, anomaly detection, and neural networks**, based on their overall performance and ability to handle the complexity of fraud data. By systematically comparing models, the goal is to identify the solution that best detects fraud with the least number of false positives and false negatives, ensuring a robust and reliable system.
2. **Improve Fraud Detection Accuracy**: The project aims to develop a model that enhances the accuracy of fraud detection by combining multiple AI/ML techniques. This will reduce both false positives and false negatives.
3. **Efficiency in Handling Large Datasets**: Given the vast amount of tax data, the model will be optimized to process and analyze large datasets quickly. Techniques like random forests and deep learning will be explored for their ability to handle big data.
4. **Enhanced Interpretability and Transparency**: While powerful, models like deep learning often act as "black boxes." One objective is to improve the interpretability of the fraud detection system, making it easier for tax authorities to understand why a tax return was flagged for potential fraud.

**EXPERIMENTAL DETAILS**

Softwares used:

1.**Data Ingestion and Processing**: Apache Kafka, Apache Spark, Python, Pandas, NumPy

2.**Machine Learning and Modeling**: Scikit-learn, TensorFlow

3.**Model Deployment and Serving**: Docker, Kubernetes, AWS Lambda/GCP Cloud Functions, REST API

4.**Data Visualization and Analysis**: Tableau/Power BI, Matplotlib/Seaborn

5.**Time Series Analysis:** A Python module that allows users to perform statistical tests and model data sets(Statsmodels).

6.**Data Storage and Management:** MongoDB, Cassandra, Elasticsearch for flexible data storage

and querying(NoSQLDatabases)

**Third-Party Tools & APIs**

•Anaconda: A distribution of Python with many popular data science packages.

•Git: For version control and collaboration.

•Jupyter Notebook: For interactive data exploration and prototyping.

**Testing Tools**

•Unit Testing: pytest, unittest

•Data Quality Testing: Apache Data Quality, Talend Data Quality

•Model Testing: TensorFlow Model Testing, Mlflow

•Performance Testing: JMeter testing

**4. METHODOLOGY**

**- DESIGN PROCEDURE**

### **1. Data Preparation**

* **Loading and Copying:** The dataset is imported into the environment and a copy is created to prevent accidental modifications to the original data.
* **Column Identification:** The dataset is examined to identify relevant columns, such as transaction details, balance changes, and a flag indicating fraudulent transactions.

### **2. Feature Selection**

* **Relevant Features:** Based on domain knowledge and exploratory analysis, key features are chosen. These features capture essential information about the transactions, such as transaction type, amounts, and balance changes.
* **Calculated Differences:** New features are created by calculating the differences between old and new balances for both the origin and destination accounts. These differences can be indicative of fraudulent activity.

### **3. Labeling**

* **Fraudulent Transactions:** A clear definition of fraudulent transactions is established based on specific criteria.
* **Target Variable:** A binary target variable (isFraud) is created to label each transaction as either fraudulent or legitimate.

### **4. Exploratory Data Analysis (EDA)**

* **Grouping and Aggregation:** Transactions are grouped by type or other relevant categories to analyze patterns and trends.
* **Visualization:** Pie charts and bar graphs are used to visualize the distribution of transaction types, amounts, and other relevant features. This helps in understanding the characteristics of fraudulent and legitimate transactions.

### **5. Machine Learning Model**

* **Data Splitting:** The dataset is divided into training and testing sets to evaluate the model's performance on unseen data.
* **Pipeline Creation:** A pipeline is constructed using PySpark to streamline the data processing and model training process.
* **Feature Engineering:** Categorical variables are encoded using StringIndexer and numerical features are combined into a single vector using VectorAssembler.
* **Model Training:** A DecisionTreeClassifier is chosen to build the machine learning model. This model can effectively handle the decision-making process involved in fraud detection.

### **6. Model Evaluation**

* **Performance Metrics:** The model's performance is evaluated using metrics like AUC-PR and AUC-ROC, which are suitable for imbalanced datasets (where fraudulent transactions may be rare).
* **Cross-Validation:** Cross-validation is employed to assess the model's generalization ability and to optimize its hyperparameters.

### **7. Results Interpretation**

* **Model Effectiveness:** The evaluation metrics provide insights into the model's ability to distinguish between fraudulent and legitimate transactions.
* **Feature Importance:** The analysis can help identify which features are most influential in predicting fraud, providing valuable information for future improvements.
* **Model Refinement:** If necessary, the model can be further refined by exploring different algorithms, tuning hyperparameters, or incorporating additional features.

**5. OUTCOMES**

**Data Exploration**

* The dataset, comprising 6,362,620 financial transactions, was thoroughly analyzed to understand its structure and content.
* This analysis provided valuable insights into the types of transactions, their frequencies, and other relevant characteristics.

### **Transaction Type Analysis**

* The distribution of transaction types was visualized using pie charts and bar charts.
* This visualization helped identify the most common transaction types and their relative proportions within the dataset.

### **Fraud Detection**

* Rule-based methods were applied to flag potential fraud cases based on specific conditions related to balances and transaction types.
* These rules were designed to capture common patterns associated with fraudulent activity.
* A total of 255,640 transactions were flagged as potential fraud cases, representing 4.02% of the total dataset.

### **Machine Learning Preparation**

* The code outlines the steps necessary to prepare the data for training a machine learning model for fraud detection using PySpark.
* This includes tasks such as feature engineering, data cleaning, and splitting the data into training and testing sets.

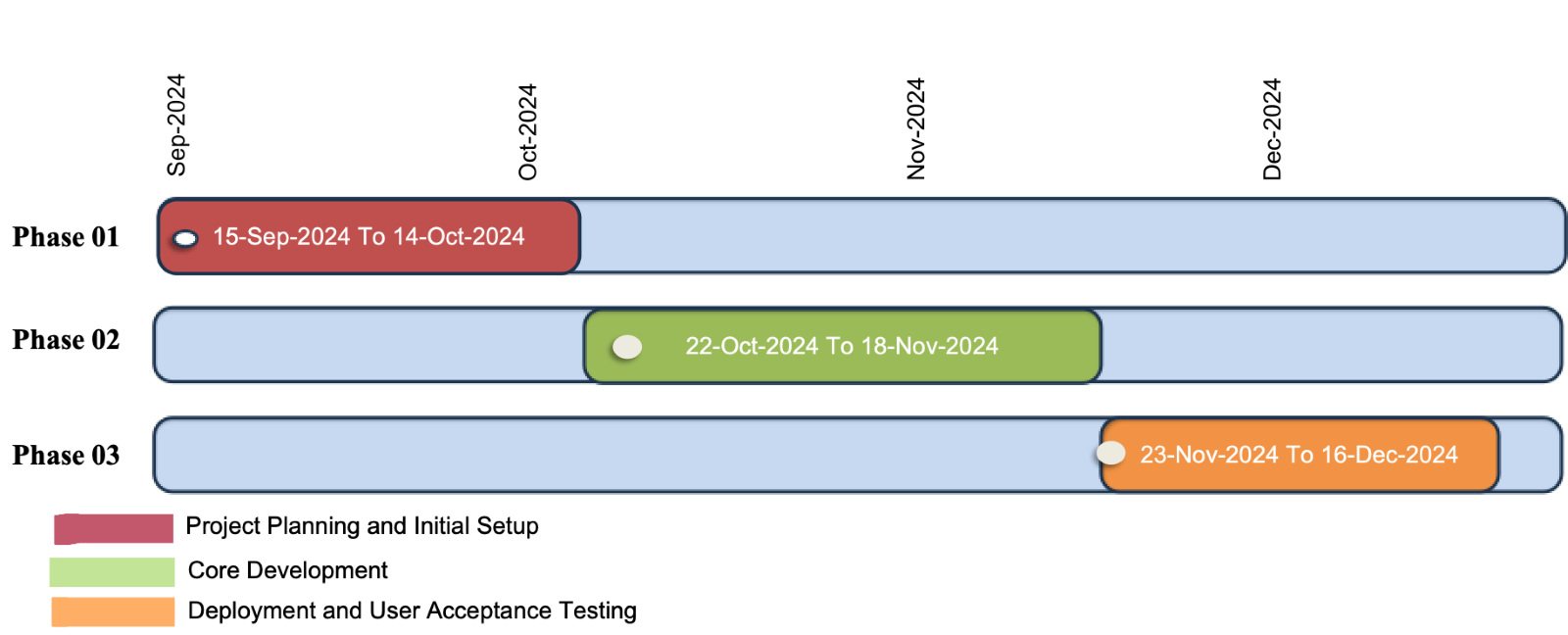
### **Model Evaluation**

* The performance of the fraud detection model was evaluated using relevant metrics, such as Area Under Precision-Recall Curve (PR) and Area Under Receiver Operating Characteristic Curve (AUC).
* These metrics provide insights into the model's ability to accurately identify fraudulent transactions and distinguish them from legitimate ones.

### **Future Enhancements**

* The analysis suggests potential areas for improvement in the fraud detection model and analysis.
* This includes exploring more sophisticated deep learning algorithms, incorporating additional features, and refining the rule-based methods.
* By addressing these areas, the fraud detection system can be further enhanced to improve its accuracy and effectiveness.

**6. TIMELINE OF THE PROJECT/ PROJECT EXECUTION PLAN**



**7. CONCLUSION**

The analysis of the dataset, comprising over 6 million transactions, revealed that different transaction types exhibit varying levels of fraud risk. This insight suggests the need for targeted fraud prevention strategies. Rule-based methods effectively identified a significant number of potential fraud cases, highlighting the importance of further investigation through manual review or automated systems.

Visualizations proved invaluable in understanding complex data patterns and anomalies. They provided an intuitive way to identify trends that were not immediately apparent in raw data. The implementation of a decision tree classifier demonstrated the effectiveness of machine learning in automating fraud detection, offering improved efficiency and accuracy compared to traditional methods.

In further studies, many more machine learning and deep learning models will be included, and the best one will be selected based on performance metrics. Exploring more sophisticated techniques, such as cross-validation, will enhance fraud detection capabilities and help adapt to evolving fraudulent activities.

**REFERENCES**

**Academic Papers**

* Matin N. Ashtiani, Bijan Raahemi (2021). "Intelligent Fraud Detection in Financial Statements Using Machine Learning and Data Mining: A Systematic Literature Review". DOI: 10.1109/ACCESS.2021.3096799

* Rohan Kumar C L, Ali Mohammed Zain, Sanjay Kumar H P, Prajwal A V, Dr. Sudarshan R (2022). "Comparative Study of Machine Learning Algorithms for Fraud Detection in Blockchain".

DOI: 10.48175/IJARSCT-5474

* Belle Fille Murorunkwere, Origene Tuyishimire, Dominique Haughton, Joseph Nzabanita (2022). “Fraud Detection Using Neural Networks: A Case Study of Income Tax”. DOI: 10.3390/fi14060168
* Dr RM Rani, Amrit Anand, Pratham Agarwal, Ayush Srivastava (2024). “Enhanced Income Tax Fraud Detection System Using Machine Learning”. DOI: 10.13140/RG.2.2.25755.68648
* N. Alsadhan (2022). “A Multi-Module Machine Learning Approach to Detect Tax Fraud”. DOI: 10.32604/csse.2023.033375
* Daniel de Roux, Boris Perez, Andrés Moreno, Maria del Pilar Villamil, César Figueroa (2018). “Tax Fraud Detection for Under-Reporting Declarations Using an Unsupervised Machine Learning Approach”. DOI: 10.1145/3219819.3219878
* Abzetdin Z. Adamov (2019). “Machine Learning and Advanced Analytics in Tax Fraud Detection”. DOI: 10.1109/AICT47866.2019.8981758
* César Pérez López, María Jesús Delgado Rodríguez, Sonia de Lucas Santos (2019). “Tax Fraud Detection through Neural Networks: An Application Using a Sample of Personal Income Taxpayers”. DOI: 10.3390/fi11040086
* Changro Lee (2022). “Deep learning-based detection of tax frauds: an application to property acquisition tax”. DOI: 10.1108/DTA-06-2021-0134

**Government Reports and Publications**

* U.S Government Accountability Office (GAO)."Artificial Intelligence May Help IRS Close the Tax Gap"