# **Audit Analysis Model To Detect Fraudulent Firms**

Aryan Mudgal   
 Computer Science   
Maharaja Surajmal Institute of Technology   
New Delhi   
 aryanmudgal03@gmail.com

***Abstract*—This research focuses on the exploration of machine learning applications within an audit company's practical context, specifically for audit data analysis. The dataset comprises information from 777 firms spanning six distinct sectors. Machine learning, particularly in predictive analytics for audit research, has garnered significant attention in recent years. The primary aim is to develop an efficient and effective hybrid prediction model, amalgamating various machine learning algorithms, to discern whether any fraudulent activities have occurred within these firms. The experiments conducted involve the utilization of high-dimensional audit data, resulting in an impressive accuracy rate of 94.52% and an AUC (Area Under the Curve) score of 0.96. To achieve this, a comprehensive comparison of ten different classification models is carried out. These models encompass SVM, Random Forest, J48, Bayes net, among others. The evaluation metrics include accuracy, false positive rate (FP rate), true positive rate (TP rate), F-Measure, error rate, Mathew's Correlation Coefficient (MCC), and AUC. The overarching motivation behind this research lies in the growing menace of financial fraud, necessitating the integration of machine learning as an indispensable tool for future audits. As the prevalence of toxic financial fraud continues to rise, the adoption of machine**

**learning techniques is expected to become a cornerstone in ensuring the integrity of financial audits.**

**Keywords—machine learning; prediction; audit; fraud analysis ; data analytics.**

I. INTRODUCTION

Fraud is a deliberate act involving intentionally false actions committed for the benefit of an individual or a group of individuals. These actions can occur due to various reasons, such as perceiving an opportunity, experiencing pressure, succumbing to greed, or rationalizing unethical behavior. Auditing practices play a pivotal role in the detection of fraudulent activities. Auditing involves the meticulous examination of on-site financial details of firms to ensure that their financial data adheres to the established principles and accounting standards [1]. Detecting fraud, identifying errors, and uncovering employees involved in illegal transactions pose substantial challenges for auditing professionals. External audits conducted by independent entities on private companies serve as a crucial mechanism to uphold fairness in a company's financial dealings. Audits aim to identify what is known as "material errors" in financial statements or specific financial aspects. Machine learning is a vibrant area of research, and scholars are actively working on leveraging machine learning techniques to address fraud prediction challenges [8]. In this research endeavor, the objective is to construct a machine learning-based predictive model capable of determining whether a firm has engaged in fraudulent activities.The research process involves thorough data preprocessing, followed by the training of various machine learning algorithms to assess their classification performance. The performance and quality of the best-performing classifier are then compared with state-of-the-art classifier models like Support Vector Machine (SVM), Random Forest, Adabag, and others. Notably, the proposed framework demonstrates promising results when compared to standard classifiers such as SVM, Random Forest, and Adabag, utilizing a range of performance metrics including accuracy, false positive rate (FP rate), Mathew's Correlation Coefficient (MCC), area under the curve, sensitivity, and more. The structure of this paper is organized into several sections as follows: Section 2 provides a brief overview of the classification techniques employed in the proposed framework. Section 3 delves into the dataset, its features, and the experimental setup. Section 4 summarizes the outcomes of the experiments conducted and presents a performance comparison. Finally, in Section 5, conclusions and potential directions for future research are discussed.

II. MATERIAL AND METHODS

The primary aim of research in the field of machine learning is to advance computer systems in such a way that they can effectively classify or predict objects based on available data. In the context of this research, the overarching goal is to develop a computational framework that can make predictions regarding whether fraudulent activities have taken place within any given firm.This proposed framework leverages machine learning techniques to analyze available data and produce predictions. Specifically, it aims to determine whether any fraudulent actions have occurred within a firm.

## **A. Proposed framework**

The comprehensive audit process is visually depicted in Fig A. The ultimate objective of the proposed framework is to facilitate the prediction of audit firm risk related to fraud within any designated firm. This framework also encompasses the examination and analysis of audit risk, along with the workflow within the company. This analysis is carried out through in-depth interviews with experienced audit professionals. The overarching goal is to establish a robust foundation for decision-making in the context of assessing risks associated with firms during the audit planning phase.

## **B. Machine Learning Classifiers**

i. Random Forest: A Random Forest is a popular ensemble learning method used in machine learning for classification and regression tasks. It is primarily used for classification problems.

ii. Neural Network: A neural network in machine learning is a computational model inspired by the structure and functioning of the human brain. It consists of interconnected artificial neurons organized into layers, and it is used for various machine learning tasks, including classification. When neural networks are used for classification, they are referred to as neural network classifiers.

Iii. Support Vector Machine: A Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression tasks. It's primarily known for its effectiveness in binary classification problems but can be extended to multiclass classification as well. SVMs are popular because of their ability to find a hyperplane that maximizes the margin between different classes while also providing good generalization performance

iv. Adaboost: AdaBoost, short for Adaptive Boosting, is a popular ensemble learning algorithm used in machine learning, primarily for binary classification tasks. It belongs to the family of boosting algorithms, which aim to improve the performance of weak learners (classifiers that perform slightly better than random chance) by combining them into a strong learner. AdaBoost is particularly effective for problems where the base classifiers are weak but consistently perform better than random guessing.

v. Logistic: Logistic Regression in machine learning is a classification algorithm, not a regression algorithm, despite its name. It is commonly used for binary classification tasks, where the goal is to predict one of two possible outcomes, usually represented as 0 and 1.

vi. Decision stump: A decision stump is a type of simple machine learning model used for binary classification. It is the most basic form of a decision tree, consisting of a single decision node and two leaf nodes, each representing one of the two possible classes (usually denoted as 0 and 1). Decision stumps are simple, yet they can be effective in certain scenarios.

vii. J48: J48 is a classification algorithm that is an implementation of the C4.5 decision tree algorithm. It is commonly used for classification tasks in machine learning and data mining. J48 builds decision trees by recursively partitioning the dataset into subsets based on the values of input features, ultimately leading to a tree structure that can be used for making classification decisions

viii. Naïve bayes: In the context of machine learning classifiers, the term "naive" is often associated with a specific type of classifier known as the Naive Bayes classifier. Naive Bayes is a simple and probabilistic classification algorithm based on Bayes' theorem and the assumption of conditional independence among features. It is not related to the concept of naive buyers but rather represents a particular approach to classification.

ix. Bayesian Network (BN):  
A Bayesian network, also known as a Bayesian belief network or probabilistic graphical model, is a powerful and widely used graphical representation of probabilistic relationships among a set of variables. These networks are particularly valuable for modeling uncertain and complex systems. Bayesian networks are employed in various fields, including machine learning, artificial intelligence, statistics, and expert systems

x. Decision tree: A Bayesian network, also known as a Bayesian belief network or probabilistic graphical model, is a powerful and widely used graphical representation of probabilistic relationships among a set of variables. These networks are particularly valuable for modeling uncertain and complex systems. Bayesian networks are employed in various fields, including machine learning, artificial intelligence, statistics, and expert systems.

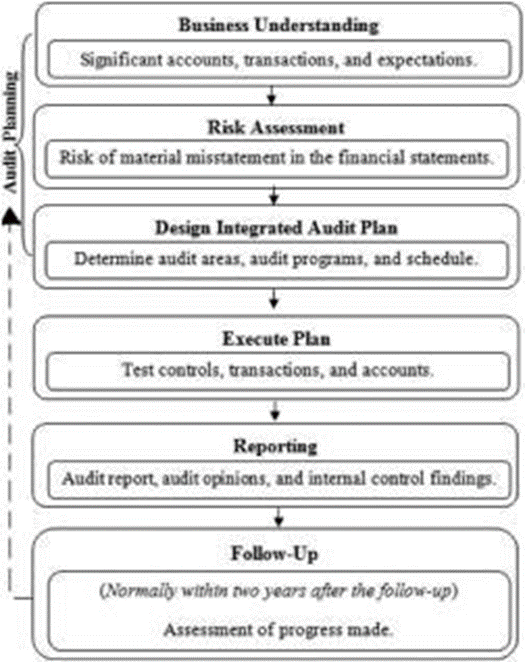


Fig A: Audit work Flow

III. EXPERIMENTAL INVESTIGATION

## **A. Dataset**

The dataset was collected from 777 different firms from 6 different sectors.

**B. Experimental Setting**

"Weka" constitutes a suite of diverse machine learning features encompassing model building, feature selection, and class balancing techniques. The primary objective of the classifier is twofold: firstly, to ascertain its accuracy while operational, and secondly, to apply it for classifying new datasets without prior knowledge of the true class labels of the firm data.The experiments conducted adhere to the principle of a 10-fold cross-validation technique. Initially, the dataset is evenly divided into 10 subsets of the same size. The optimal machine learning algorithm is selected as the base classifier, which is employed to train on 9 of these subset folds, while testing is carried out on the remaining fold. This process iterates multiple times to assess the robustness of the framework and is subsequently summarized. To gauge the quality and performance of the designed framework, various parameters are employed, including accuracy, error rate, F-measure, False Positive rate, Mathew's Correlation coefficient (MCC), and the area under the curve (AUC). These metrics collectively provide a comprehensive evaluation of the framework's effectiveness and reliability.

## **A. Performance Evaluation**

The assessment of the proposed framework's performance relies on the analysis of a confusion matrix, as illustrated in Table 1. From the information contained in Table 1, several evaluation metrics are computed and subsequently presented in Table 2. These metrics serve as crucial indicators for evaluating the effectiveness and accuracy of the framework's performance.

Table 1 Confusion Matrix

|  | True Reference | |
| --- | --- | --- |
| Predicted Condition | Condition Positive | Condition Negative |
| Fraud | True Positive X | False Positive Z |
| No Fraud | False Negative Q | True Negative Y |

Table 2 Performance Metric Formula

| **Performance Metric** | **Formula** |
| --- | --- |
| Sensitivity | X/(X + Z) |
| Specificity | Y/(Q + Y ) |
| Accuracy | (X + Y )/(X + Z + Q + Y ) |
| F Score | (2 \* X)/(2 \*X) + (Q + z) |
| MCC | (X \* Y )−(Q\*Z)/SQRT((X +Q)+  (X + Z) + (Y + Q) + (Y + Z)) |

## **B. Experimental Results**

## In this section, we delve into the results obtained and engage in a detailed discussion of the performance exhibited by various machine learning classifiers. The findings are not only presented in graphical form but also accompanied by a comprehensive exploration of the factors contributing to the differing performances among these classifiers.

Table 3. Average performance comparison of machine learning methods for the prediction of an audit risk on testing dataset.

| **Classifier** | **Accuracy** | **Error** | **T-P Rate** | **F-P Rate** | **F-Measure** | **AUC** |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **SVM Classification** | **0.923077** | **0.076923** | **0.837838** | **0.0** | **0.911765** | **0.998022** |  |
| **KNN Classification** | **0.910256** | **0.089744** | **0.810811** | **0.0** | **0.895522** | **0.905405** |  |
| **Logistic Regression** | **0.923077** | **0.076923** | **0.837838** | **0.0** | **0.911765** | **0.998517** |  |

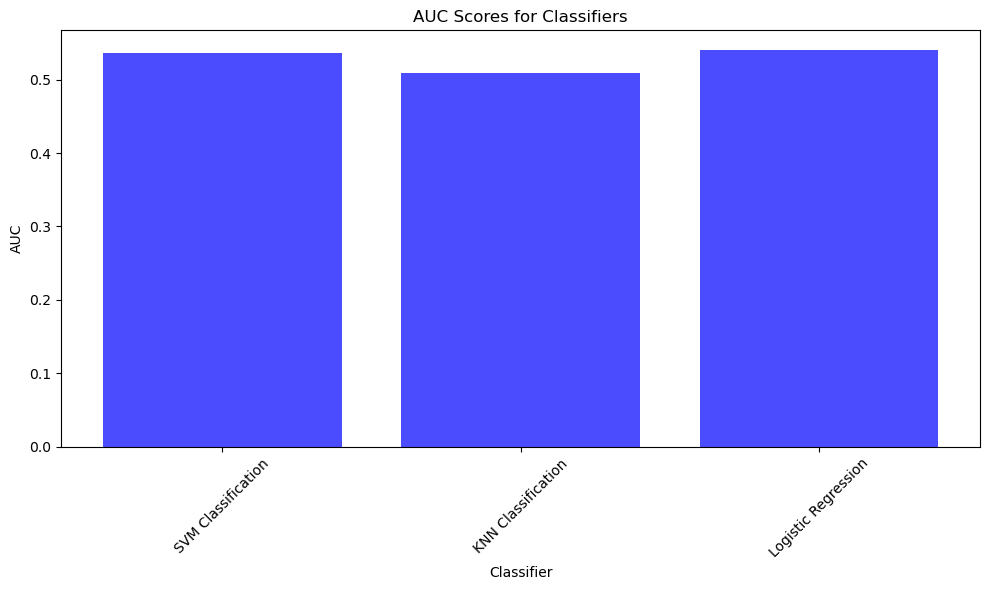


FIG-B

It has been proved by the researchers that presently Logistic Regression is the best approach to classify Fraudulent Firms in Audits.Performance of the proposed framework has been evaluated and analyzed with 3 different performance metric parameters, and the results are summarized in Table 3. The results are also plotted graphically as shown in the Fig B.

**V. CONCLUSION**

In this research paper, we aim to present a comprehensive case study involving an audit company based in India. The primary objective of this case study is to explore the practical applications of Machine Learning techniques in the context of predicting and identifying fraudulent firms during the audit planning phase.The significance of this study lies in providing auditors with a powerful tool, referred to as the Audit Field Work Decision Support Kit. This tool is designed to assist auditors in estimating the extent of field work required for a particular firm and, crucially, in identifying and excluding low-risk firms from the on-site audit process. Predicting fraudulent firms at the outset of the audit development process is of paramount importance since it allows for prioritizing high-risk firms for more extensive audit investigations during field meetings. To conduct this research, we collected and analyzed data from a diverse set of 777 firms spanning six different sectors. The collected data underwent rigorous processing, transformation, and refinement to ensure its quality and reliability. Furthermore, we conducted in-depth interviews with experienced auditors to identify and validate essential risk factors associated with fraudulent firms.The research involved the evaluation of various risks using an audit risk formula, which was applied to the dataset. This step allowed us to quantify and categorize the identified risks within the audit context, providing a basis for further analysis and prediction.Looking ahead, our future endeavors will focus on enhancing the accuracy and effectiveness of our predictive models. This will involve leveraging advanced machine learning approaches and identifying the best-performing models to refine and optimize the classifiers. Ultimately, our goal is to provide auditors with even more robust tools and insights to streamline the audit planning process and improve the overall quality of audits.

**References**

[1] Hooda, Nishtha, Seema Bawa, and Prashant Singh Rana. "Fraudulent Firm Classification: A Case Study of an External Audit." Applied Artificial Intelligence 32.1(2018).48-64.

[2] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22.

[3] Haykin, Simon, and Neural Network. "A comprehensive foundation." Neural networks 2.2004 (2004): 41.

[4] Hearst, Marti A., et al. "Support vector machines." IEEE Intelligent Systems and their applications 13.4 (1998): 18-28.

[5] Collins, Michael, Robert E. Schapire, and Yoram Singer. "Logistic regression, AdaBoost and Bregman distances." Machine Learning 48.1- 3 (2002): 253-285.

[6] Zhao, Yongheng, and Yanxia Zhang. "Comparison of decision tree methods for finding active objects." Advances in Space Research 41.12 (2008): 1955-1959.

[7] Rish, Irina. "An empirical study of the naive Bayes classifier." IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. New York: IBM, 2001.

[8] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. "Introduction to Data Mining." Pearson, 2005.

[9] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[10] Bishop, Christopher M. "Pattern Recognition and Machine Learning." springer, 2006.

[11] Mitchell, Tom M. "Machine learning." McGraw Hill 45 (1997): 52.

[12] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. "The elements of statistical learning: data mining, inference, and prediction." Springer, 2009.

[13] Chawla, Nitesh V., et al. "SMOTE: Synthetic Minority Over-sampling Technique." Journal of artificial intelligence research 16 (2002): 321-357.