**Machine Learning based Method for Insurance Fraud Detection on Class Imbalance Datasets with Missing Values**

**ABSTRACT: -**

Insurance fraud, particularly within the automobile insurance sector, is a significant challenge faced by insurers, leading to financial losses and influencing pricing strategies. Fraud detection models are often impacted by class imbalance, where fraudulent claims are much rarer than legitimate claims, and missing data further complicates the process. This research tackles these issues by utilizing two car insurance datasets—an Egyptian real-life dataset and a standard dataset. The proposed methodology includes addressing missing data and class imbalance, and it incorporates the AdaBoost Classifier to enhance the model’s accuracy and predictive power. The results demonstrate that addressing class imbalance plays a crucial role in improving model performance, while handling missing data also contributes to more reliable predictions. The AdaBoost Classifier significantly outperforms existing techniques, improving prediction accuracy and reducing overfitting, which is often a challenge in fraud detection models. This study presents valuable insights into how improving data quality and using advanced algorithms like AdaBoost can enhance fraud detection systems, ultimately leading to more effective identification of fraudulent claims. These enhancements can significantly aid insurance companies in reducing financial losses, improving decision-making, and refining pricing models.

|  |  |
| --- | --- |
| **EXSISTING SYSTEM** | **PROPOSED SYSTEM** |
| * Existing fraud detection models in the insurance industry typically rely on conventional machine learning algorithms like Logistic Regression, Decision Trees, and Random Forests. However, they face challenges when working with imbalanced datasets, where fraudulent claims are relatively rare compared to legitimate claims. This imbalance leads to poor model performance, as the classifiers are biased towards the majority class. Furthermore, missing data and inconsistencies in the dataset contribute to unreliable predictions, affecting the overall accuracy of these models. * Despite their utility, traditional methods often suffer from overfitting, where models perform well on training data but fail to generalize to unseen data. Overfitting occurs because these models can become too complex when trying to fit noisy or incomplete data, resulting in reduced model robustness. While some approaches use sampling methods like undersampling or oversampling to address class imbalance, they do not always achieve optimal results, especially when the dataset contains many missing or noisy entries. | * The proposed system enhances insurance fraud detection by incorporating the AdaBoost Classifier, an ensemble technique that improves predictive performance by combining multiple weak classifiers to create a strong, accurate model. AdaBoost's ability to reduce overfitting makes it especially effective for handling noisy or incomplete datasets, ensuring that the model generalizes well to new data. This classifier is also integrated with techniques like SMOTE (Synthetic Minority Over-sampling Technique) to address the class imbalance problem, thereby improving the detection of fraudulent claims that are underrepresented in the dataset. * Furthermore, the proposed system employs more robust data preprocessing methods to handle missing data effectively. These preprocessing techniques ensure that the model works with cleaner, more complete datasets, improving overall prediction accuracy. The combination of AdaBoost with these advanced data handling methods makes the model more reliable, scalable, and efficient. By addressing both class imbalance and missing data, the proposed system outperforms traditional methods in terms of accuracy and robustness, providing a more effective solution for fraud detection in the insurance industry. |
| **EXISTING ALGORITHM**   * Logistic Regression, Decision Trees, Random Forest and KNN | **PROPOSED ALGORITHM: -**   * AdaBoost Classifier |
| **ALGORITHM DEFINITION: -**   * Traditional insurance fraud detection systems typically use machine learning algorithms like Logistic Regression, Decision Trees, and Random Forest. These algorithms, while effective for general classification tasks, struggle with imbalanced datasets where fraudulent claims are much fewer than legitimate ones. They are also limited by issues like overfitting, where the model learns noise from the training data, and missing values, which can degrade model performance. Despite using techniques like undersampling or oversampling to handle class imbalance, these models often fail to achieve optimal accuracy and robustness. * Additionally, many existing systems struggle with overfitting, where models memorize specific patterns in the training data, leading to poor generalization on unseen data. While Random Forest and similar algorithms are used, they still face challenges with large, imbalanced datasets, particularly when handling complex fraud patterns. The lack of advanced techniques to address both class imbalance and missing values diminishes the effectiveness of these systems, making it difficult to maintain a high fraud detection rate without sacrificing accuracy for legitimate claims. | **ALGORITHM DEFINITION: -**   * The proposed system leverages the AdaBoost Classifier, an ensemble method that combines weak learners to create a strong classifier, improving model robustness and accuracy. AdaBoost’s ability to reduce overfitting makes it especially useful for noisy and incomplete datasets. Additionally, techniques like SMOTE are used to address class imbalance, ensuring better performance in detecting fraudulent claims. The system also includes advanced data preprocessing methods to handle missing values, ensuring a more complete dataset and enhancing model performance, making it a more reliable solution for fraud detection. * In addition to AdaBoost, the proposed system employs the Synthetic Minority Over-sampling Technique (SMOTE) to address the class imbalance problem by generating synthetic samples for the minority class, which is critical for improving the fraud detection rate. The integration of advanced data preprocessing methods ensures that missing values are effectively handled, preventing data gaps from negatively impacting the model's performance. These combined strategies—AdaBoost for classification, SMOTE for data balancing, and advanced preprocessing for handling missing data—lead to a more accurate, reliable, and robust system for insurance fraud detection, significantly outperforming traditional methods in terms of both prediction accuracy and generalization. |
| **DRAWBACKS: -**   * Class Imbalance * Handling Missing Data * Overfitting * Limited Performance with Complex Fraud Patterns * Inefficient Fraud Detection | **ADVANTAGES: -**   * Enhanced Fraud Detection * Effective Missing Data Handling * Optimized Model Performance * Reduced Overfitting * Increased Accuracy in Predictions |

**SYSTEM ARCHITECTURE:**

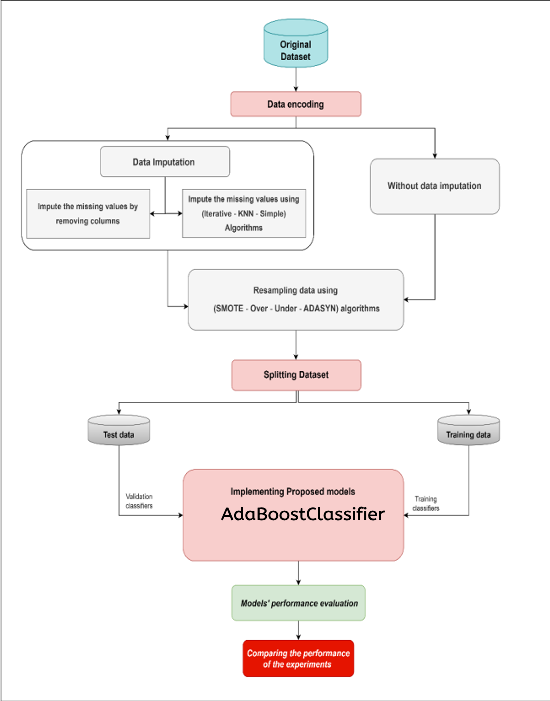


Fig:- proposed model

**MINIMUMSYSTEM REQUIREMENTS**

**HARDWARE REQUIREMENTS**

* PROCESSOR : Pentium i3 Processor
* RAM : 4GB DD RAM
* HARD DISK : 500 GB

**SOFTWARE REQUIREMENTS**

* BACK END : PYTHON
* OPERATING SYSTEM : WINDOWS 10
* IDE : Spyder3