**INSURANCE FRAUD DETECTION**

**Abstract:**

Countries all over the world are working towards improving their economies, and a key goal is to prevent money laundering and fraud. One major area of concern is insurance fraud, which costs both companies and the public a significant amount of money. To tackle this issue, many countries are turning to data analysis and machine learning. These technologies can help identify and prevent fraudulent activities in insurance claims, saving companies time and money. To address this problem effectively, it's important to research existing solutions and build upon those findings. The ultimate aim is to create a model that can quickly identify suspicious claims, making the insurance industry more efficient in dealing with fraud.

**Business Need**:

In our world, where people highly value their belongings, everyone wants to protect what they have. When the COVID-19 pandemic started, many countries faced challenges in getting vaccines to protect their citizens. Just like how people rush to get a vaccine to safeguard themselves, the idea behind insurance is similar. People are willing to pay money to protect themselves from unexpected losses.

In the United States, the insurance industry is worth a massive 1.28 trillion dollars. Unfortunately, about 80 billion dollars are lost each year due to insurance fraud in the U.S. This fraud causes insurance companies to raise their policy costs, making them less competitive. Higher costs also mean people have to pay more for insurance policies.

This project aims to find the best and simplest way to fight against fraudulent insurance claims. The main issue in identifying fraud is the large number of claims processed by insurance companies. However, this large amount of data could be an advantage. If officials combine and analyze this data, they can create better models to identify suspicious claims.

The project will explore different methods used to solve similar problems, testing and comparing them to find the most effective approach. The goal is to build a simple, quick, and accurate model that can flag suspicious claims without causing stress to the system.

**Problem statement:**

This project's main goal is to create a model that can figure out if an insurance claim is a fraud or not. We'll test different methods to find the best one for this job. The idea is to present this model to insurance companies, suggesting they use a personalized version for their systems. The model should be easy to use for large sets of data but still smart enough to be quite successful in spotting fraud.

**Summary**:

We have gone through many articles for this project and below are the few articles that we feel good to go with.

Existing financial fraud detection systems:

In a 2016 article by Jarrod West, various research papers were compared to understand fraud detection systems and the different models used for fraud detection, along with their accuracy. The article also provided a detailed description of each model. Additionally, it compared different types of fraud investigation and the methods used in recent research. For example, common methods for detecting credit card fraud include support vector machines, decision trees, hybrid methods, and artificial immune systems.

In a 2019 study by Bruns et al., an evolutionary algorithm-based model was proposed to learn Complex Event Processing rules for extracting relevant information from large-scale data streams. This model considered heuristics when determining the ideal parameters for the process. The validation of this model used real-world transportation data to assess its benefits and limitations for real-world decision-making.

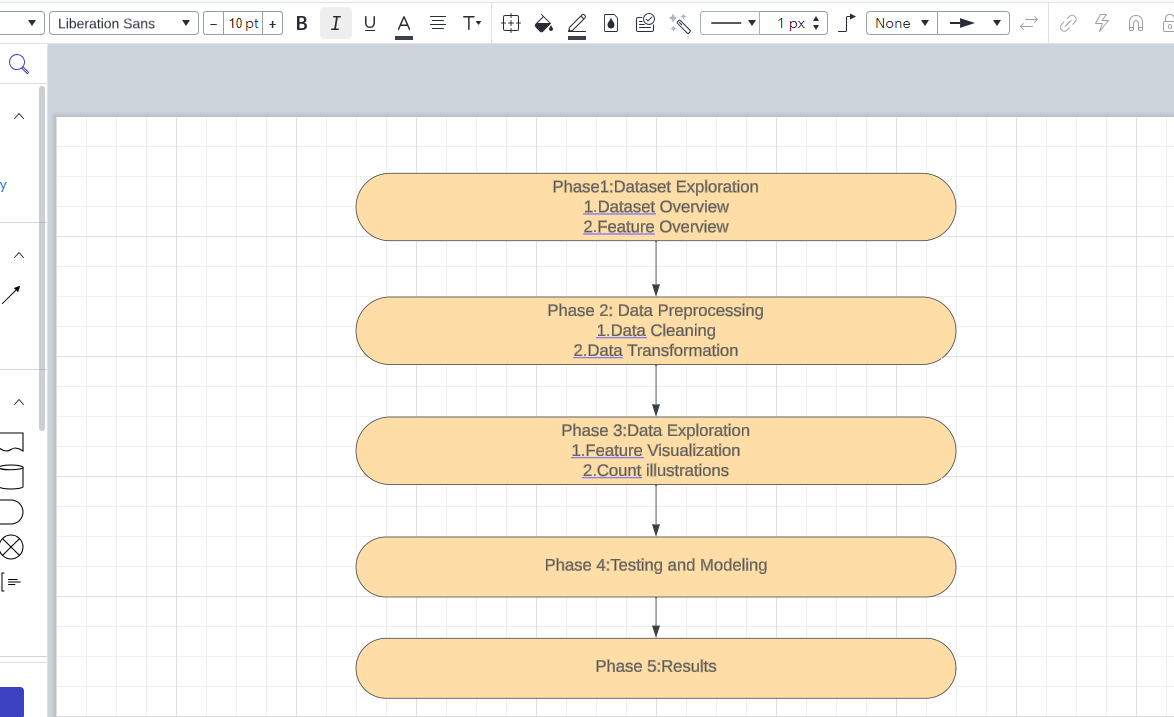
In another 2019 study by Eweoya et al., decision trees were employed to predict fraud in bank loan administration, aiming to reduce losses from loan defaults in finance. The study used real-world data from a financial institution for its fraud detection application.

Dealing with huge data sets:

In a 2016 article by Alejandro Correa Bahnsen, the author discusses how to handle a large dataset. The dataset initially had over 100 million records, each with 27 attributes. One of these attributes identifies fraudulent transactions. However, only 40,000 transactions were labeled as fraud, making up just 0.025% of the total dataset.

Training a model on such a massive dataset would be challenging for the application and describing findings with so few fraud cases would be difficult. The suggested solution was to reduce the dataset to 236,735 transactions, but with a higher fraud ratio of 1.5%. From this modified dataset, three subsets were created: a training dataset with 50%, a validation dataset with 25%, and a testing dataset with 25%. The primary use of the validation dataset was to adjust and improve the model before testing it again.

High Level Architecture:



**Team Identification**:

We are a group of three people and the following are the group members.

Thirumalesh Pathuri, Srinivas Kollu, Ravi Teja

**Gantt Chart**:

| Phase 1 | 12/18/2023 |
| --- | --- |
| Phase 2 | 1/2/2024 |
| Phase 3 | 1/8/2024 |
| Phase 4 | 1/15/2024 |
| Phase 5 | 1/22/2024 |

**References:**

1.Alejandro Correa Bahnsen, D. A. (2016). Feature engineering strategies for credit card fraud detection. Expert Systems With Applications.

2.Jarrod West, M. B. (2016). Intelligent financial fraud detection: A comprehensive review. ScienceDirect.

3. Javad Forough, S. M. (2021). Ensemble of deep sequential models for credit card fraud detection. Applied Soft Computing Journal.

4. Johannes Jurgovskya, M. G.-E.-G. (2018). Sequence classification for credit-card fraud detection. Expert Systems With Applications.

5. Mieke Jans, ,. J. (2011). A business process mining application for internal transaction fraud mitigation. Exert Systems with Applications .

6. Siddhartha Bhattacharyya, S. J. (2011). Data mining for credit card fraud: A comparative study. Decision Support Systems.

7. X. Liu, J. W. (2009). Exploratory undersampling for class-imbalance learning. Systems, Man, and Cybernetics.

8. Severino, M.K. and Peng, Y., 2021. Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata. Machine Learning with Applications, 5.

9. Hitesh Kumar Reddy Toppi, R., Bhavna, S., & Ginika, M. (2018). Crime Prediction & Monitoring Framework Based on Spatial Analysis. International Conference on Computational Intelligence and Data Science (ICCIDS 2018). Elsevier

10. Andrea Dal Pozzolo, G. B. (2018). Credit Card Fraud Detection: A Realistic Modeling. IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, vol 29, NO 8.

11. Andrea Dal Pozzolo, O. C.-A. (2014). Learned lessons in credit card fraud detection from a practitioner. Expert systems with applications .

12. Bart Baesens, S. H. (2021). Data engineering for fraud detection . Decision Support Systems .

13. E.W.T. Ngai, Y. H. (2011). The application of data mining techniques in financial fraud detection: A classification. Decision Support Systems, 559-569.

14. Eunji Kim, J. L.-k.-a.-i. (2019). Champion-challenger analysis for credit card fraud detection: Hybrid. Expert Systems With Applications, 214 - 224.