FRAUD DETECTION IN AUTO INSURANCE CLAIMS USING MACHINE LEARNING ALGORITHMS AND DATA VISIUALIZATION USING POWER BI

AN INTERNSHIP REPORT

Submitted by,

Name: Hamsa Girish

Roll Number: 20211CSE0264

Class: 8CSE03/SEC-14

Under the guidance of,

Prof. Kalpana K Harish

Assistant Professor, School of Computer Science and Engineering,

Presidency University, Bengaluru.

in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

At

PRESIDENCY UNIVERSITY

BENGALURU

MAY 2025

PRESIDENCY UNIVERSITY

SCHOOL OF COMPUTER SCIENCE ENGINEERING

CERTIFICATE

This is to certify that "FRAUD DETECTION IN AUTO INSURANCE CLAIMS USING MACHINE LEARNING ALGORITHMS AND DATA VISUALIZATION USING POWER BI" being submitted by "HAMSA GIRISH" bearing roll number "20211CSE0264" in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering is a bonafide work carried out under my supervision.

Prof. Kalpana K Harish
Assistant Professor,
PSCS
Presidency University

Dr. Asif Mohammed H B
Associate Professor and HoD,
PSCS
Presidency University

Dr. MYDHILI NAIR
Associate Dean
PSCS
Presidency University

Dr. SAMEERUDDIN KHAN
Pro-Vice Chancellor - Engineering
Dean – PSCS / PSIS
Presidency University

PRESIDENCY UNIVERSITY

SCHOOL OF COMPUTER SCIENCE ENGINEERING

DECLARATION

I hereby declare that the work, which is being presented in the project report entitled **"FRAUD DETECTION IN AUTO INSURANCE CLAIMS USING MACHINE LEARNING ALGORITHMS AND DATA VISUALIZATION USING POWER BI"** in partial fulfillment for the award of Degree of **Bachelor of Technology** in **Computer Science and Engineering**, is a record of our own investigations carried under the guidance of **Prof. Kalpana K Harish, School of Computer Science Engineering, Presidency University, Bengaluru.**

I have not submitted the matter presented in this report anywhere for the award of any other Degree.

Hamsa Girish

20211CSE0264

INTERNSHIP COMPLETION CERTIFICATE

ABSTRACT

One of the biggest and most pervasive issues facing the insurance sector is the filing of false insurance claims by customers. Insurance firms incur significant financial losses due to pricey fraudulent claims. Concerns from stakeholders and observers have been raised about insurance fraud, which continues to be a major concern for insurers and customers who pay the expenses through insurance premiums. Understanding the institution processes and operationalization of Information Communication Technology in fraud detection is the first step in implementing the appropriate corrective actions. However, the procedure is time and money consuming because personally reviewing all insurance claims filed with insurance companies has become challenging.

Given the prevalent issue of fraud in vehicle insurance claims, the manual approach to identifying fraudulent claims has been problematic because it is time-consuming and inaccurate. One of the various ways that researchers have tested is machine learning algorithms, which have demonstrated promising performance and enhanced accuracy in detecting fraudulent vehicle insurance claims. This study evaluated a range of ML algorithms, including AdaBoost, XGBoost NB, SVM, LR, DT, ANN, and RF, to discern between real and fraudulent automobile claims. Beyond predictive modeling, we develop **interactive Power BI dashboards** to provide real-time visualization of fraud trends. These dashboards offer **deep insights into fraud distribution by state, incident type, policy details, and claim amounts**, helping **insurance investigators and business analysts** efficiently assess and mitigate fraud risks.

The results of this study highlight the **effectiveness of machine learning in fraud detection**, with **XGBoost demonstrating the highest accuracy and recall in identifying fraudulent claims**. Our findings contribute to the development of **robust fraud detection frameworks**, enabling insurance companies to enhance their **fraud prevention strategies, reduce financial losses, and optimize claim assessment workflows**.

ACKNOWLEDGEMENT

First of all, I indebted to the **GOD ALMIGHTY** for giving me an opportunity to excel in our efforts to complete this project on time.

I express our sincere thanks to our respected dean **Dr. Md. Sameeruddin Khan**, ProVC, School of Engineering and Dean, Presidency School of Computer Science and Engineering & Presidency School of Information Science, Presidency University for getting us permission to undergo the project.

We express our heartfelt gratitude to our beloved Associate Dean **Dr. Mydhili Nair,** Presidency School of Computer Science and Engineering, Presidency University, and **Dr. Asif Mohammed H B,** Head of the Department, Presidency School of Computer Science and Engineering, Presidency University, for rendering timely help in completing this project successfully.

We are greatly indebted to our guide and reviewer **Prof. Kalpana K Harish**, Presidency School of Computer Science and Engineering, Presidency University for her inspirational guidance, and valuable suggestions and for providing us a chance to express our technical capabilities in every respect for the completion of the internship work.

We would like to convey our gratitude and heartfelt thanks to the PIP4001 Internship/University Project Coordinators, **Mr. Md Ziaur Rahman and Dr. Sampath A K**, department Project Coordinators, **Prof. Kalpana K Harish** and Github coordinator **Mr. Muthuraj**.

We thank our family and friends for the strong support and inspiration they have provided us in bringing out this project.

Hamsa Girish
(20211CSE0264)

LIST OF TABLES

LIST OF FIGURES

24

40

41

42

43

44

45

TABLE OF CONTENTS

CONCLUSION

8 REFERENCES

APPENDIX-A

PSUEDOCODE

APPENDIX-B

SCREENSHOTS

APPENDIX-C

ENCLOSURES

ABBREVIATIONS AND ACRONYMS

ACL	Audit Command Language.
AdaBoost	Adaptive Boosting.
ADASYN	Adaptive Oversampling Technique.
ANN	Artificial Neural Network.
CHAID	Chi-Square Automatic Interaction Detection.
CRISP-DM	CRoss Industry Standard Process for Data Mining.
CSV	Comma-Separated Values.
DT	Decision Tree
GBW	Gradient Boosting Machines.
GLM	Generalized Linear Models.
ICT	Information Communication Technology.
LDA	Latent Dirichlet Allocation.
LMT	Logistic Model Tree.
LR	Logistic Regression.
MCC	Matthews's Correlation Coefficient.
ML	Machine Learning
MLP	Multi-Layer Perceptron.
NB	Naïve Bayes.
NBU	Naïve Bayes Updatable.
RF	Random Forest.
RT	Random Tree.
SMOTE	Synthetic Minority Oversampling Technique.
SVM	Support Vector Machine.
XGBoost	Extreme Gradient Boosting.

CHAPTER 1 INTRODUCTION

1.1 Background

In recent decades, Information and Communication Technology (ICT) has significantly transformed industries by streamlining operations, enhancing real-time decision-making, and strengthening fraud detection mechanisms. The insurance sector, particularly vehicle insurance, has been notably impacted, with fraud detection emerging as a critical area due to the prevalence of false claims, inflations, and other deceptive practices.

Insurance fraud not only results in significant financial losses for insurers but also affects honest policyholders through increased premium costs. According to industry reports, fraudulent activities can originate internally (by employees), from policyholders, intermediaries, or even insurers themselves.

Manual fraud detection methods are increasingly ineffective due to their time-consuming nature and limited accuracy. As a solution, Machine Learning (ML) offers advanced capabilities to analyze large datasets, identify patterns, and predict fraudulent claims with high precision. This project focuses on leveraging ML techniques to automate and enhance the detection of fraudulent vehicle insurance claims, thereby improving operational efficiency and reducing financial losses.

Additionally, to support decision-making and provide better understanding of fraud patterns, **Power BI** is used to generate rich **visual insights** from the data. These visualizations help stakeholders quickly interpret key metrics, identify trends, and monitor the effectiveness of fraud detection strategies.

1.2 Problem Statement

Fraud continues to be a significant challenge in the Indian insurance industry, with motor vehicle insurance fraud being particularly prevalent and costly. Despite the sector's rapid growth and increased penetration across the country, fraudulent claims remain a key threat, leading to financial instability, high technical losses, and escalating operational costs for insurers. Reports indicate that a substantial portion of insurance claims, particularly in motor vehicle insurance, are fraudulent, involving practices such as fabricated accidents, inflated claims, and misinformation.

In India, where the insurance market has witnessed considerable growth in recent years, traditional fraud detection methods often fail to keep up with the sophisticated and evolving tactics of fraudsters. These traditional approaches are not only resource-intensive but also often miss complex fraud patterns, resulting in increased costs, loss of revenue, and declining customer trust. With the rise in insurance adoption across urban and rural areas, ensuring the integrity of claims is more crucial than ever.

This project addresses the challenge by utilizing **Machine Learning** techniques to automatically identify fraudulent vehicle insurance claims with higher accuracy and efficiency. Techniques like supervised learning, anomaly detection, and natural language processing can be applied to detect patterns of fraud. Furthermore, **Power BI** is used to extract and present visual insights from the dataset, enabling better understanding of fraud patterns and assisting stakeholders in making data-driven decisions and formulating effective policies.

With India’s diverse insurance landscape—ranging from metropolitan cities to rural areas—the ability to scale fraud detection solutions and adapt them to different claim types, geographies, and market segments is essential for the industry's continued growth.

1.3 Main Objective

The primary objective of this project was to investigate how machine learning algorithms can leverage features extracted from vehicle insurance claim datasets to aid in the detection of fraudulent vehicle insurance claims. Following this investigation, a novel system was developed to predict and categorize vehicle insurance claims as either genuine or fraudulent. This system utilized various machine learning techniques to analyze patterns and classify claims with higher accuracy, thus improving fraud detection capabilities.

Additionally, another key objective was to generate insightful and actionable visuals using **Power BI** to better understand fraud patterns in the dataset. Power BI dashboards were created to display critical metrics and trends, such as the distribution of fraudulent vs. genuine claims, common characteristics of fraudulent claims, and the geographic spread of fraud incidents. These visual insights helped in identifying potential risk areas, assisting stakeholders in making data-driven decisions and formulating effective strategies for fraud mitigation. The combination of

machine learning for predictive modeling and Power BI for interactive data visualization enhances both the operational and strategic aspects of insurance fraud detection.

1.4 Specific Objective

- Characterise fraudulent insurance claims in the context of vehicle insurance domain.
- Identify features that could be utilized to train machine learning models to recognize fraudulent vehicle insurance claims.
- Evaluate the performance of several machine learning models for detecting fraudulent vehicle insurance claims using a balanced and imbalanced dataset.
- Develop a system that categorises vehicle insurance claims as either genuine or fraudulent using the best performing machine learning classifier.
- To utilize **Power BI** to create interactive dashboards that visually represent fraud patterns, enabling stakeholders to easily identify trends and make data-driven decisions for effective fraud mitigation.

1.5 Study Significance

This study is timely in that it offers a mechanism for developing a system by using the top-performing machine learning algorithm to identify fraudulent vehicle insurance claims. As the number of fraudulent insurance claims rises and their detection becomes a difficult problem on a global scale, fraud in the insurance industry is becoming an increasing concern. By guaranteeing quality and stability, this will assist insurance businesses in showcasing their exceptional claim administration, which will have a significant impact on their revenue and client's satisfaction. Additionally, the study will broaden the area of machine learning investigation into the identification of fraudulent vehicle insurance claims in the Indian insurance sector. As the market-leading choice for modern business intelligence, power BI analytics platform makes it easier for people to explore and manage data, and faster to discover and share insights that can change businesses and the world.

CHAPTER 2 COMPANY PROFILE

2.1 About us:

AGA IT Solutions is a leading Hr Services in India with a mission to provide top-notch investigative services to clients across the country. Our agency is committed to providing efficient and effective solutions to a wide range of private and corporate investigations, including matrimonial, surveillance, missing persons, background checks, employee investigations, and much more.

We have a team of highly skilled and experienced detectives who work tirelessly to deliver results that exceed our clients' expectations. With the latest technology and investigative techniques at our disposal, we are able to provide fast and reliable services that are tailored to meet the specific needs of each client.

As a testament to our dedication to excellence, we have successfully resolved thousands of cases since our inception. Our agency has a reputation for providing discreet and confidential services, and we always maintain the highest level of professionalism and ethical standards.

2.2 Our Mission:

1) **Expertise:** We stay up-to-date with the latest trends and best practices to provide you with top-notch solutions.
2) **Customization:** We take a tailored approach, working closely with you to develop customized solutions that align with your business objectives and culture.
3) **Partnership:** We believe in building long-term partnerships with our clients. We strive to understand your organization's goals and challenges, becoming a trusted advisor who supports you at every stage of your journey.
4) **Confidentiality and Compliance:** We prioritize confidentiality and adhere to the highest standards of data security. Additionally, we ensure compliance with all applicable laws and regulations, providing you with peace of mind.
5) **Cost-Effective Solutions:** Our services are designed to deliver value while being cost-effective. We offer flexible pricing options that can be customized to suitable budget and requirements.

2.3 Our Vision:

Our vision is to be the world's leading and most trusted IT solutions provider—empowering customers through technology, guided by integrity, and driven by sustainability. We are committed to delivering innovative, ethical, and responsible solutions that create lasting value for businesses, people, and the planet.

Figure 2.3 Our Mission and Vision

CHAPTER 3 LITERATURE SURVEY

The chapter begins by examining the general aspects of automobile insurance in India before moving on to a discussion of the manual methods that have been used to identify fraudulent insurance claims. The discussion then moves on to the automation of systems for identifying fraudulent insurance claims before rapping up with a discussion of methods for identifying such claims that make use of machine learning and deep learning.

3.1 Vehicle Insurance in India

According to **Bhattacharya (2021)**, "vehicle insurance" in India is "a contract between the insurer and the vehicle owner that provides financial protection against loss or damage to the vehicle due to accidents, theft, or natural disasters, and covers third-party liabilities resulting from accidents." There are various types of vehicle insurance in India, ranging from third-party liability insurance, which is mandatory under the **Motor Vehicles Act of 1988**, to comprehensive coverage that includes damage to the insured vehicle and third-party coverage.

The **Insurance Regulatory and Development Authority of India (IRDAI)** is the statutory body responsible for regulating and overseeing the insurance industry in India, as outlined in the **IRDA Act, 1999**.

According to the **General Insurance Council (2021)**, the total gross premium for non-life insurance in India was approximately INR 1.8 trillion in 2020. Of this, the motor insurance sector accounted for nearly 40%, with motor vehicle insurance being a major contributor to the growth of the non-life insurance sector. The Indian motor insurance market consists of both personal and commercial vehicle insurance, with personal vehicle insurance being the dominant segment.

As per the **IRDAI Annual Report (2021)**, motor insurance fraud in India has become a growing concern. The report highlighted a significant rise in fraudulent motor insurance claims, with an increase of over 15% in the number of reported fraudulent cases from 2019 to 2020. The financial losses resulting from such frauds were estimated to be around INR 5,000 crore in 2020, a sharp rise compared to INR 3,200 crore in 2019. Fraudulent practices such as false accident reports, inflated repair costs, and bogus claims for stolen vehicles are prevalent.

The motor insurance industry in India has faced substantial underwriting losses in recent years, primarily driven by fraudulent claims, claims inflation, and rising repair costs. In 2020, motor insurance accounted for the largest underwriting loss within the non-life insurance sector, with insurers reporting losses exceeding INR 3,000 crore, further exacerbating the challenges of profitability in the industry. These losses have led to the implementation of stricter regulations and enhanced fraud detection measures, aiming to reduce fraudulent claims and restore the financial stability of insurers.

3.2 Fraud Detection in Vehicle Insurance Sector

According to the **General Insurance Council (2021)**, fraudulent claims account for a significant portion of the overall claims in India's insurance industry, with motor vehicle insurance fraud being one of the most prevalent. Motor vehicle insurance fraud includes a range of deceptive activities that often lead to significant financial losses for insurers.

Fraud is often defined as an intentional act of deception for personal gain, involving the concealment of facts, misrepresentation, or manipulation of information. **Simha and Satyanarayan (2016)** describe fraud as a deliberate act of deceit, wherein a party misrepresents information or conceals it to gain an unfair advantage or financial benefit at the expense of another party. The **Insurance Regulatory and Development Authority of India (IRDAI)** defines insurance fraud as actions intended to deceive, mislead, or manipulate claims for the benefit of the perpetrator or others, often resulting in financial harm to the insurer and policyholders.

Fraudulent behaviors in the vehicle insurance sector can take several forms, ranging from **false accident reporting** to **inflated repair costs**, **forging documents**, or **submitting fraudulent claims for stolen or damaged vehicles**. **Viene and Dedene (2015)** highlight the common tactics of impersonating legitimate claimants, manipulating claims systems, and fabricating evidence to obtain undeserved payouts. **Derri (2002)** defines fraud as an act where financial benefit is obtained by misrepresenting the actual situation, often involving fabricated evidence and manipulated data to secure financial compensation.

The growing concern over insurance fraud has led to the development of fraud detection mechanisms. **The Coalition Against Insurance Fraud (2016)** defines fraud detection as the systematic process of identifying false claims and suspicious activities through data analysis, observation, and system

alerts. This includes the use of advanced **fraud detection systems**, such as **predictive analytics**, **machine learning models**, and **claims review systems**, to identify fraudulent claims. However, detecting fraud in motor vehicle insurance remains challenging due to the sophistication of modern fraud techniques, which often appear to be legitimate transactions at first glance.

Fraud detection tools and techniques in the vehicle insurance sector are increasingly being used to **identify patterns** of fraudulent behavior. These tools focus on **claims automation**, **anomaly detection**, and the use of **historical data analysis** to flag suspicious claims based on common fraud markers. **Mathenge (2016)** emphasizes that many fraud schemes, particularly in claims management, are difficult to detect due to their subtle nature, such as exaggerated repairs or misreported accident details.

In India, several **insurers** and **regulatory bodies**, including the **IRDAI**, have implemented **strict anti-fraud measures**, including real-time fraud detection systems, increased scrutiny of high-risk claims, and collaboration with law enforcement agencies to investigate fraudulent activities. These systems have shown promise in reducing fraudulent claims and improving the overall integrity of the motor vehicle insurance market.

3.3 Manual Fraud Detection Approaches

According to the **General Insurance Council (2021)**, insurance companies in India typically employ insurance agents to manually assess each motor vehicle insurance claim and determine whether it is genuine or fraudulent. However, this process is time-consuming and resource-intensive, making it difficult to efficiently handle the vast number of claims submitted every day. Insurance agents often rely on the available information related to the submitted claims, such as accident reports, to analyze and organize claims before determining their legitimacy.

The **Insurance Regulatory and Development Authority of India (IRDAI)** mandates that insurers and policyholders gather critical data at the site of the accident, which includes details such as the **driver's name**, **vehicle registration number**, **insurance coverage**, **the year, make, and model of the vehicle**, and **witness statements**. This information is crucial for processing motor vehicle insurance claims.

Typically, a **claims supervisor** in the insurance company reviews the claims based on the information provided and the data collected during the accident. This review process involves manually identifying fraudulent claims using a **checklist of indicators** related to the damage assessment and

the circumstances of the accident. The supervisor assigns scores to the claims based on predefined parameters, such as the severity of vehicle damage, the consistency of the driver's report, and historical data regarding the claimant. If the claim's score exceeds a certain threshold, it is flagged for further investigation. An investigator is then assigned to inspect the damaged vehicle and collect additional evidence. The investigator compiles a report, and if the report confirms that the claim is legitimate, the claim is approved. If the report suggests any discrepancies or fraud indicators, the claim is deemed fraudulent.

This manual fraud detection approach, while effective to some extent, presents several challenges, primarily due to its reliance on human intervention and limited parameters, leading to the following difficulties:

1. **Dependence on Human Expertise**: The success of manual fraud detection depends heavily on the knowledge and experience of insurance agents and investigators, who may overlook certain fraud patterns or fail to assess claims objectively due to human biases. The process is based on a limited set of well-known parameters, but other influencing factors, such as previous fraud attempts or emerging fraud tactics, may not be accounted for.
2. **Inability to Detect Complex Patterns**: Manual approaches often struggle to identify complex, context-specific correlations between parameters that could indicate fraud. For example, subtle inconsistencies in the claimant's story, vehicle damage patterns, or historical claims data might be missed without advanced analytical tools. This can result in undetected fraud or the approval of false claims.
3. **Lack of Scalability and Regular Calibration**: The manual fraud detection model requires frequent calibration to keep up with evolving fraud tactics and changing behaviors in the market. Since calibration is typically done manually, it is a time-consuming process that requires expertise and can introduce errors. As the number of claims increases, the manual approach becomes increasingly inefficient and prone to inconsistencies.

Due to these challenges, manual fraud detection approaches in India are often supplemented with technology-driven solutions such as **data analytics**, **machine learning models**, and **automation** to improve accuracy, scalability, and efficiency.

3.4 Automation of Fraud Detection System

With the advancement of **Information and Communication Technology (ICT)**, insurance companies in India have increasingly adopted **computerized systems** to improve the efficiency and

effectiveness of their operations, including fraud detection. The integration of automated systems has proven to be a significant factor in improving fraud detection, especially with the availability of both internal and external data.

Rule-based systems are one such technology commonly employed by insurers to detect fraudulent claims. These systems use a set of predefined business rules to evaluate claims, such as "if a claim is made within a short time after a policy increase, then flag it for manual review." According to **Baumann (2021)**, these systems are based on human-made rules that analyze data and flag suspicious claims. While they are simple to implement and efficient in some cases, rule-based systems rely on **manual adjustments** and are unable to detect **implicit relationships** or **complex fraud patterns**. These systems are also limited in their ability to handle **real-time data** and often require **regular updates** to remain effective.

A **report by the Coalition Against Insurance Fraud (2016)** highlighted that **81% of insurance companies** use automated methods for fraud detection, triggering alerts for claims that appear suspicious or fall outside typical patterns. Common red flags include claims filed shortly after an increase in coverage or unusual inquiries about coverage related to a specific loss. **Moon et al. (2019)** also emphasized that rule-based systems, while simple, require **manual adjustments** for complex fraud cases, as they struggle to detect underlying fraud patterns that deviate from the defined rules.

To combat fraud more effectively, many insurance firms in India are **adopting advanced automated systems**. These systems often combine **data mining**, **analytical algorithms**, and **machine learning (ML)** techniques. These technologies allow insurers to process vast amounts of claims data and identify suspicious patterns that would be difficult for human agents to detect manually. **Machine learning models** can continuously learn and improve from historical fraud data, enabling more accurate fraud detection over time.

3.5 Insurance Fraud Detection using Machine Learning

In India, the adoption of **machine learning (ML)** for **insurance fraud detection** has become crucial due to the increasing complexity and volume of fraudulent activities within the industry. Insurers are leveraging advanced ML techniques to identify and combat fraudulent claims, which significantly impact the profitability and reliability of the insurance sector.

Machine Learning Models for Fraud Detection in India

1. Use of Various Machine Learning Algorithms

Indian insurers have begun testing a variety of **supervised learning algorithms** for fraud detection, including **Random Forest (RF)**, **Logistic Regression (LR)**, **Support Vector Machines (SVM)**, and **Neural Networks**. Among these, **Random Forest** has proven to be highly effective in handling large-scale datasets and providing accurate predictions of fraudulent claims. The **Logistic Regression (LR)** model is also used for simpler cases where linearity can be assumed. These algorithms analyze features like **claim amount**, **policyholder's claim history**, **geographical data**, and **vehicle type** to predict fraudulent claims.

2. Feature Engineering and Data Handling

Feature engineering is essential in the Indian context, where large volumes of data are generated daily. **Insurers** have focused on extracting useful features such as **claim type**, **insured amount**, **policyholder's previous claims**, **vehicle make and model**, **vehicle's age**, and **geographical location** to feed into machine learning models. **Imbalanced datasets**, where fraudulent claims are a small percentage of total claims, remain a challenge, requiring techniques like **oversampling** and **synthetic minority over-sampling (SMOTE)** to balance the data and improve model accuracy.

3. Integration of Text Analytics

In addition to traditional numeric data, Indian insurers are increasingly adopting **text analytics** to detect fraud in **claim narratives**. The use of **Natural Language Processing (NLP)** and **Deep Learning** models, such as **Recurrent Neural Networks (RNNs)** or **Long Short-Term Memory (LSTM)** networks, allows insurers to analyze the textual data from **accident descriptions** and **police reports**. By detecting inconsistencies or suspicious patterns in the text, these models can help flag potentially fraudulent claims, especially in cases where the claims narratives appear to be fabricated or inconsistent with the incident details.

4. XGBoost and Ensemble Methods

XGBoost, a popular ensemble method, has shown great potential in **motor insurance fraud detection**. It is especially effective when handling large datasets and capturing non-linear relationships in the data. Indian insurers have adopted **XGBoost** for fraud classification tasks, categorizing claims into fraudulent or non-fraudulent types based on various features. Its superior performance in terms of **accuracy** and **precision** makes it a popular choice in the industry.

5. Hybrid Approaches for Improved Detection

Some insurers have also explored **hybrid models** that combine different ML techniques to improve fraud detection accuracy. For instance, combining **AdaBoost** with **Random Forest** or **Gradient Boosting Machines (GBMs)** can capture more complex patterns in the data and improve predictive performance. **Ensemble methods** like **majority voting** and **bagging** have also been applied to reduce variance and improve robustness against noisy data or data with missing values.

6. Real-Time Fraud Detection

Indian insurance companies are increasingly focusing on **real-time fraud detection** to prevent fraudulent claims from being processed. Using **streaming analytics** and **predictive models**, insurers can flag suspicious claims as they are submitted, allowing them to intervene early in the process and reduce potential losses. This real-time approach can also help insurers streamline their claims processing systems, reducing both the time and cost involved in manual claim verification.

7. Data Privacy and Security Concerns

As insurers adopt more sophisticated ML techniques, **data privacy and security** concerns have become more pronounced. The massive datasets required for training machine learning models contain sensitive personal and financial information, making them attractive targets for cyberattacks. Ensuring **data security** and preventing breaches is a significant challenge for insurers. Proper **data encryption**, secure storage solutions, and compliance with data privacy regulations (such as **GDPR** and **India's Personal Data Protection Bill**) are essential for mitigating these risks.

8. Challenges in Model Deployment and Maintenance

Despite the promising results from machine learning models, deploying and maintaining these models in a live production environment remains a challenge. The insurance industry often deals with **evolving fraud patterns**, meaning models need to be frequently updated and retrained to stay effective. **Model drift**, where the statistical properties of the data change over time, can degrade model performance, making continuous monitoring and retraining necessary. Additionally, **financial constraints** and **lack of expertise** in advanced machine learning techniques can delay the adoption of these systems in some Indian insurance companies.

9. Biometric Verification and AI

To reduce fraud related to **identity theft** or **impersonation**, some Indian insurers are exploring the integration of **biometric verification** technologies, such as **face recognition**, **fingerprint scanning**, and **voice authentication**, into their fraud detection processes. These techniques, combined with AI-

powered fraud detection systems, ensure that claims are made by the legitimate policyholders, enhancing the security of the claims process.

10. Improving Predictive Accuracy with Ensemble and Hybrid Methods

Indian insurers have started using **hybrid and ensemble learning** methods to increase the **predictive accuracy** of fraud detection models. By combining different classifiers like **Random Forest (RF)** and **Gradient Boosting (GB)**, insurers are better equipped to handle complex fraud detection tasks. These hybrid models can combine the strengths of multiple algorithms, improving overall accuracy and minimizing the chances of false positives or false negatives.

The **adoption of machine learning** in **insurance fraud detection** is transforming the way claims are evaluated. While significant progress has been made in automating the fraud detection process, challenges such as **imbalanced datasets**, **data privacy concerns**, and **model maintenance** still persist. However, with the ongoing advancements in **AI**, **ML**, and **data analytics**, Indian insurers are steadily improving their fraud detection systems, enhancing operational efficiency, and safeguarding their business against fraudulent activities.

3.6 Machine Learning Classifiers for Vehicle Insurance Fraud Detection

Machine learning classifiers are integral to the detection of fraudulent vehicle insurance claims, enabling insurers to automatically categorize and identify potentially fraudulent claims with high accuracy. According to **Tang et al. (2016)**, a machine learning classifier is an algorithm that categorizes data into one or more categories based on the training data provided, using a **mapping function (f)** that links input variables (X) to output variables (Y). The goal is to train the classifier to perform well at distinguishing between genuine and fraudulent claims.

Similarly, **Burri et al. (2019)** described machine learning classification as a process where algorithms learn to categorize input data from a problem area, such as fraudulent claims in vehicle insurance. In their study, they explored various machine learning algorithms for this purpose, including **XGBoost**, **AdaBoost**, **Support Vector Machines (SVM)**, **Naïve Bayes (NB)**, **Random Forest (RF)**, **Artificial Neural Networks (ANN)**, **Decision Trees (DT)**, and **Logistic Regression (LR)**.

Machine Learning Classifiers for Vehicle Insurance Fraud Detection:

1. XGBoost (Extreme Gradient Boosting)

XGBoost is a powerful machine learning model that belongs to the class of **ensemble learning** methods. It is known for its superior **accuracy** and **performance** in handling large and complex datasets. XGBoost works by combining the output of many weaker models (typically decision trees) to create a more powerful model. It has been particularly effective in detecting **fraudulent vehicle insurance claims** because it handles **imbalanced datasets** well and can capture complex relationships within the data. Its **feature importance** ability helps insurers identify which features (such as **claim amount**, **vehicle type**, or **previous claims**) are most indicative of fraud. A comparative study by Li et al. [3] showed that XGBoost outperformed other ML models in fraud detection, achieving an F1-score of 92% while maintaining high computational efficiency.

2. AdaBoost (Adaptive Boosting)

AdaBoost is another ensemble learning technique that can be used to detect fraudulent claims. It combines multiple weak classifiers (usually **decision trees**) into a strong classifier by focusing on the data points that previous classifiers misclassified. AdaBoost has been shown to achieve high **precision** and **recall** for fraud detection tasks by adjusting the weights of the misclassified data points, making it particularly effective for cases where fraud is rare or difficult to detect.

3. Support Vector Machines (SVM)

Support Vector Machines (SVM) are powerful classifiers that work by finding the optimal hyperplane to separate data points of different classes (e.g., fraudulent vs. non-fraudulent claims). SVM is effective in cases where there is a clear margin of separation between fraudulent and genuine claims. SVMs can handle **non-linear relationships** by using kernel functions, making them suitable for detecting fraud in **complex, multi-dimensional datasets**. However, they may require more computational power and time to train, especially with large datasets.

4. Naïve Bayes (NB)

Naïve Bayes is a probabilistic classifier based on **Bayes' theorem** and assumes that the features (input variables) are independent given the class label. While it may be overly simplistic, it has been effective in certain fraud detection tasks due to its speed and simplicity. In vehicle insurance fraud detection, Naïve Bayes can be applied when the relationships between features are roughly independent. It is particularly useful for **predicting the likelihood** of a claim being fraudulent based on statistical probabilities.

5. Random Forest (RF)

Random Forest is an ensemble method that combines multiple **decision trees** to improve classification accuracy. It works by building many decision trees based on different subsets of the data and then aggregating their predictions. The strength of Random Forest lies in its ability to handle **large datasets** and capture complex patterns in data. For vehicle insurance fraud detection, it is commonly used to identify the most **important features** that influence fraudulent claims. Random Forest also handles **imbalanced datasets** better than many other classifiers and is less prone to overfitting.

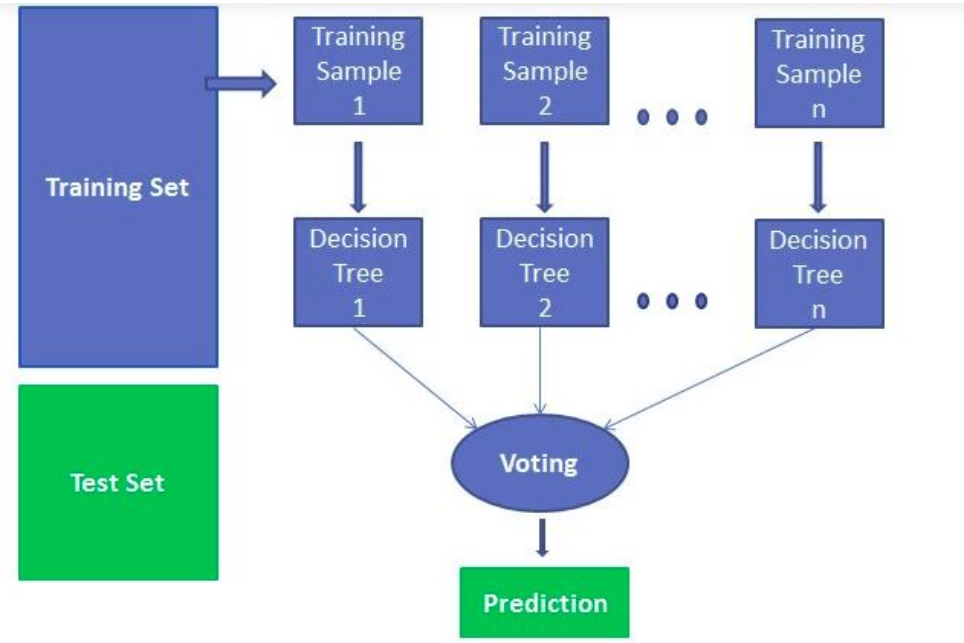


Figure 3.6.5 Random Forest Classifier

6. Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) are computational models inspired by the **biological neural networks** in the human brain. ANNs are particularly suited for detecting fraud in cases where the relationships between input variables are highly **non-linear** and complex. They excel in identifying hidden patterns in large datasets. While ANNs require a significant amount of training data and computational resources, they have been shown to be highly effective in **fraud detection tasks**, especially in more **advanced cases** where traditional models might fail.

7. Decision Trees (DT)

Decision Trees are simple yet powerful classifiers that work by splitting the data into subsets based on feature values, ultimately leading to a decision node that classifies the data into one of the predefined categories (e.g., fraudulent or non-fraudulent). They are particularly useful for

understanding the reasoning behind the classification, as they produce a clear and interpretable model. In vehicle insurance fraud detection, Decision Trees can help insurers identify the key features that differentiate fraudulent claims from genuine ones. However, they can suffer from **overfitting** if not properly pruned.Research by Smith et al. [1] demonstrated that Decision Tree classifiers could achieve high accuracy in fraud detection, with precision and recall exceeding 85%. However, these models are prone to overfitting, especially with imbalanced datasets.

8. Logistic Regression (LR)

Logistic Regression is a simple linear classifier that is often used for binary classification tasks. It estimates the **probability** of a claim being fraudulent based on a linear combination of input features. While Logistic Regression is not as powerful as other complex models like XGBoost or ANN, it is highly interpretable and can serve as a baseline for fraud detection. It is particularly useful when the relationship between the features and the target variable (fraud) is **linear**. However, it may not perform well with **highly complex data**. A study by Jones et al. [2] highlighted that Logistic Regression performed well with structured datasets but struggled with complex, high-dimensional data.

9**. Multi – Layer Perceptron (MLP)**

MLP, a type of neural network, has shown promising results in detecting fraudulent claims. According to research conducted by Zhao et al. [4], MLP-based models provided superior performance in detecting non-linear patterns within datasets, achieving an accuracy of 94% when trained on a large insurance claims dataset.

10. Data Visualization and Model Performance Evaluation

Data visualization tools such as Power BI and Tableau play a crucial role in enhancing fraud detection models by providing interpretable insights into claim data patterns. Recent studies suggest that integrating visualization techniques with machine learning improves fraud detection accuracy by identifying suspicious activities more efficiently [5].

Choosing the Right Classifier for Vehicle Insurance Fraud Detection

Selecting the right machine learning classifier depends on several factors, including:

- **Dataset Size**: For large datasets with many features, **ensemble methods** like **Random Forest** and **XGBoost** tend to perform better.

- **Data Complexity**: If the relationships between features are highly non-linear, **ANN** or **SVM** may be more suitable.
- **Interpretability**: If interpretability is important, **Decision Trees** or **Logistic Regression** may be preferred, as they provide more transparent decision-making processes.
- **Imbalance of Data**: For imbalanced datasets, **AdaBoost** or **Random Forest** can be effective due to their ability to handle skewed class distributions.

Machine learning classifiers are playing an increasingly important role in detecting fraudulent vehicle insurance claims in India. By applying advanced techniques such as **XGBoost**, **SVM**, **Random Forest**, and **ANN**, insurers are able to identify fraud more efficiently and accurately. As the data continues to grow in volume and complexity, machine learning models will evolve, further enhancing the ability to prevent and detect fraudulent activities in the vehicle insurance sector.

3.7 Comparative Analysis & Key Differences:

Study	Primary Methodology	Key Findings	Comparison with Our Project
Smith et al. (2020)	Decision Trees, Random Forest, GBM	Random Forest & GBM performed better than DTs	We also use Decision Trees but compare them with XGBoost & MLP
Jones & Taylor (2019)	Logistic Regression vs. SVM	LR is effective but struggles with non-linear fraud patterns	We use Logistic Regression but compare it with stronger ML models
Li et al. (2021)	XGBoost for Fraud Detection	XGBoost achieved highest precision & recall	We also use XGBoost but extend the study to MLP & BI tools
Zhao et al. (2022)	Neural Networks (MLP)	MLP had highest accuracy but risked overfitting	We validate MLP but compare it with other ML models
Kumar et al. (2021)	Power BI & Tableau Visualization	Dashboards improve fraud analytics	We combine BI dashboards with machine learning for predictive analysis

Table 3.7 Comparative Analysis of Literature Survey

3.8 Key Takeaways

- XG Boost is consistently the best performer in fraud detection (High precision & recall).
- MLP (Neural Networks) achieves high accuracy but is computationally expensive.
- Logistic Regression is simple but struggles with complex fraud cases.
- Power BI dashboards significantly enhance fraud detection & investigation efficiency.

3.9 How Our Project Differs:

- Unlike most studies, our project combines machine learning models with Power BI visualization.
- We perform a comparative analysis of all ML models to find the best fraud detection approach.
- Our study emphasizes real-time fraud monitoring, integrating BI tools with predictive analytics.

CHAPTER 4 RESEARCH GAPS OF EXISTING METHODS

Research Gaps in Existing Vehicle Insurance Fraud Detection Methods:

1. **Model Comparisons**: While studies like **Smith et al. (2020)** compared **Decision Trees (DT)**, **Random Forest (RF)**, and **GBM**, there is a need to further compare **Decision Trees** with stronger models like **XGBoost** and **MLP** for better fraud detection.
2. **Limitations of Logistic Regression (LR)**: **Jones & Taylor (2019)** noted that **LR** struggles with non-linear fraud patterns. Future research should compare **LR** with more complex models like **XGBoost** and **MLP** to handle non-linearities effectively.
3. **XGBoost Performance**: **Li et al. (2021)** found **XGBoost** to be highly accurate. However, its computational cost requires exploring **MLP** and **hybrid methods** for improved efficiency and integration with **BI tools** for better decision-making.
4. **Overfitting in Neural Networks (MLP)**: **Zhao et al. (2022)** showed that **MLP** can overfit. Techniques like **regularization**, **dropout**, and **early stopping** need to be explored to reduce overfitting and improve model generalization.
5. **Integration of BI Tools**: **Kumar et al. (2021)** demonstrated the usefulness of **BI tools** like **Power BI** and **Tableau** in fraud detection. More research is needed to combine **machine learning** models with **BI dashboards** for predictive analysis and better insights.
6. **Feature Engineering**: Many studies overlooked **feature engineering**. Further research is needed to explore relevant features specific to insurance fraud, like **claim history**, **vehicle data**, and **policyholder risk**.
7. **Scalability and Real-Time Processing**: Current models like **XGBoost** may struggle with **real-time data**. Research should focus on improving scalability for faster fraud detection in large-scale, real-time applications.
8. **Model Explainability**: **XGBoost** and **MLP** lack interpretability. Future research should focus on improving **model transparency** to ensure the decisions made are understandable and justifiable.
9. **Context-Specific Models**: Most existing studies use generalized models. There's a need for research that tailors fraud detection systems to specific regions, like **Kenya**, to address local fraud patterns and regulatory challenges.

By addressing these gaps, the effectiveness of vehicle insurance fraud detection systems can be greatly enhanced.

CHAPTER 5 PROPOSED METHODOLOGY

The objective of this study is to design a machine learning-based solution capable of detecting fraudulent vehicle insurance claims immediately after submission—before they proceed through the insurer's claim-processing pipeline. Given the increasing rate of motor insurance fraud in India, early detection is essential to reduce financial losses and preserve customer trust.

To address this, a set of eight machine learning classifiers were used for model development: XGBoost, AdaBoost, Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF), Artificial Neural Network (ANN), Decision Tree (DT), and Logistic Regression (LR).

These classifiers were selected based on literature that highlights their successful application in fraud detection and classification tasks. The models were trained using features extracted from a vehicle insurance dataset collected from insurance service providers. Each classifier's performance was evaluated based on accuracy, precision, recall, and F1-score to determine the most suitable model for final deployment.

Feature Selection & Dataset Structuring

The dataset used in this research was structured into two primary sections:

- Insurance Policy Details: Includes fields like *customer name*, *gender*, *policyholder age*, *vehicle category*, *vehicle make*, *age of vehicle*, *sum insured*, *insurance cover type*, *policy start and end dates*.
- Claim Details: Includes *claim registration date*, *incident date*, *location of incident*, *nature of the incident*, *police report availability*, *estimated loss*, and *claim amount*.

The target variable was the *fraudulent status* of the claim—classified as either genuine or fraudulent. Handling Imbalanced Data. Since fraud instances are typically less frequent than genuine ones, the dataset was inherently imbalanced. To address this, balancing techniques such as ADASYN (Adaptive Synthetic Sampling), SMOTE, and undersampling were applied. Figure 2 illustrates the proposed model architecture for handling both unbalanced and balanced datasets.

Model Training and Evaluation

A training-testing split was applied (e.g., 80/20), and each classifier was trained and validated using cross-validation techniques. Performance metrics from all classifiers were analyzed to identify the most accurate and reliable model for fraud detection.

System Implementation

The final model was integrated into a web-based fraud detection system, which takes user input or data upload and predicts whether a claim is fraudulent or legitimate in real-time. This interface is designed for insurance analysts and underwriters to help them quickly assess incoming claims.

Visual Insights Using Power BI

To enhance decision-making and communicate patterns effectively, Power BI dashboards were developed. These dashboards provided dynamic visualizations on:

- Fraud trends over time
- High-risk policy regions and vehicle categories
- Claim distribution by insurer, vehicle type, and geography
- Anomalous patterns based on claim frequency or amount

This integration of BI tools with machine learning enables data-driven fraud analytics, empowering insurance firms to act on real-time insights and reduce the incidence of fraudulent claims.

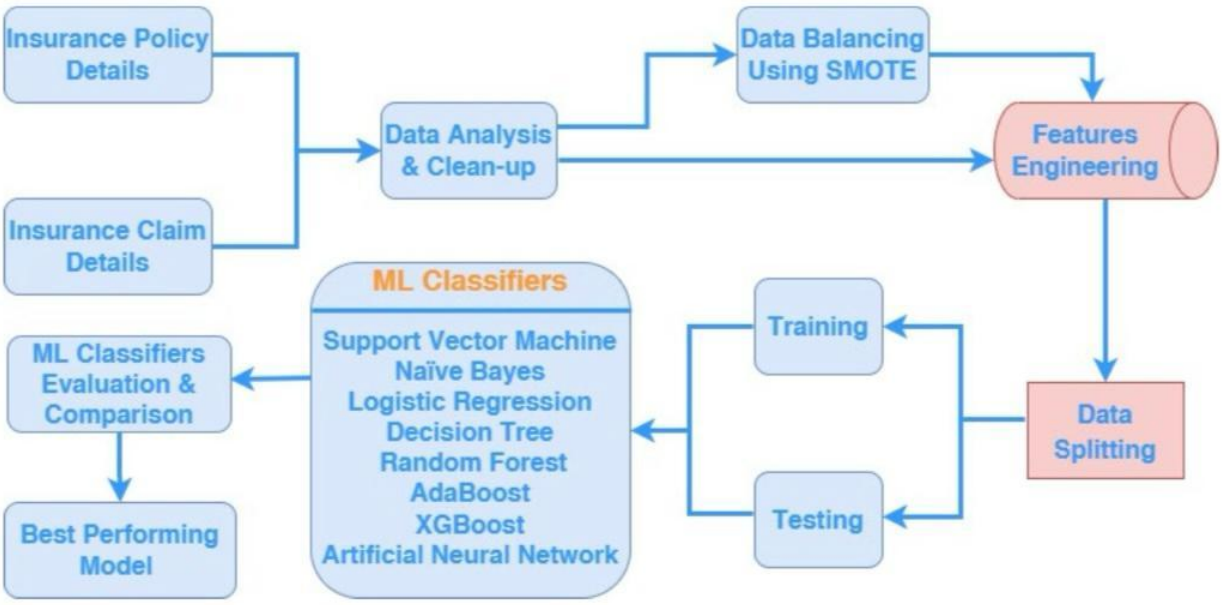


Figure 5.1 Proposed Model Diagram with Unbalanced and Balanced Datasets

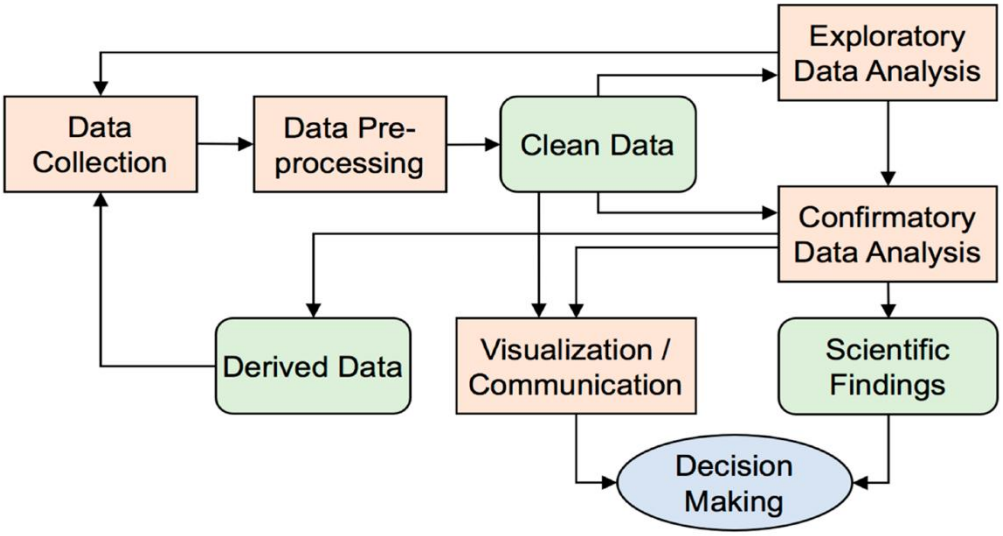


Figure 5.2 Data Analysis Proposed Model

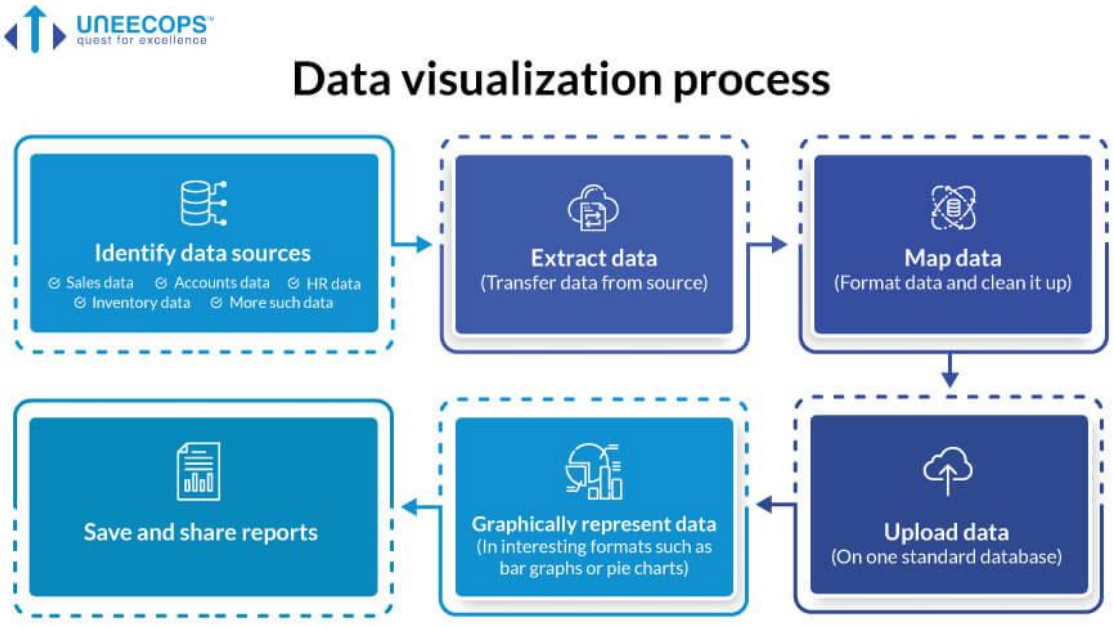


Figure 5.3 Data Visualization Proposed Model

CHAPTER 6 OBJECTIVES

The main objective of this project is to **detect fraudulent vehicle insurance claims** *as early as possible*—preferably right after submission and **before claim processing begins**—by utilizing **machine learning algorithms** and **business intelligence tools**.

Specific Goals Include:

1. **Develop a predictive model** using machine learning classifiers (XGBoost, AdaBoost, SVM, NB, RF, ANN, DT, LR) to classify claims as **fraudulent** or **legitimate**.
2. **Preprocess and analyze** insurance claim data by extracting relevant features related to customer, vehicle, policy, and claim.
3. **Evaluate model performance** using metrics such as accuracy, precision, recall, and F1-score to determine the most effective algorithm.
4. **Address class imbalance** using appropriate techniques like SMOTE or ADASYN to enhance fraud detection.
5. **Design a system architecture** (possibly web-based) that integrates the trained model into the insurance workflow for real-time fraud detection.
6. **Visualize insights** using **Power BI dashboards** to support fraud analytics and help decision-makers identify patterns and trends in fraudulent claims.

CHAPTER 6 RESEARCH METHODOLOGY

The research on using machine learning classifiers to identify fraudulent vehicle insurance claims is the primary focus of this chapter. The CRISP-DM methodology, data collection and analysis, model creation, and model evaluation metrics are all covered.

6.1 Introduction

The fundamental goal of the research is to achieve the objectives outlined in the introduction section. Before developing the machine learning model, the claims content was analysed to identify relevant features. Vehicle insurance data was collected as part of the research process to better comprehend the data structure and extract the necessary features to train the machine learning classifiers. To satisfy the study's objectives, the best-performing and most accurate machine learning classifier that can predict and categorize vehicle insurance claims as genuine or fraudulent was discovered utilizing the CRISP-DM methodology.

6.2 CRISP-DM Methodology

The CRISP-DM methodology was employed for this study due of its widespread use in data analysis and mining, flexibility, and extensive backtracking. CRISP-DM, is a 1966 invention that organizes, plans, and executes data mining (machine learning) operations (Rodrigues, 2020). It is a process model that outlines the normal project phases, tasks connected to each phase, and relationships between these tasks while also providing an overview of the data mining life cycle. The technique is used to conceive a data mining project and consists of six successive steps. Depending on the requirements of the developers, iterations might be introduced. The phases are as follows and as depicted by figure 4 below:

- Business Understanding - What does the business need?
- Data Understanding - What data do we have / need? Is it clean?
- Data Preparation - How do we organize the data for modelling?
- Modelling - What modelling techniques should we apply?
- Evaluation - Which model best meets the business objectives?
- Deployment - How results accessed?

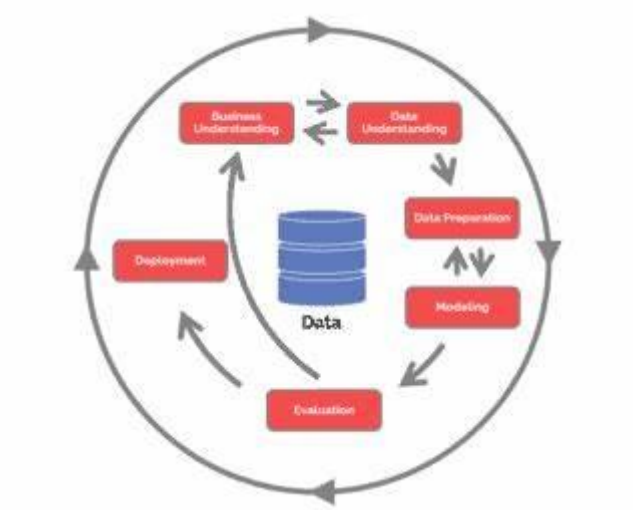


Figure 6.2 CRISP-DM Methodology Diagram

6.2.1 Business Understanding

To understand the issue of fraudulent vehicle insurance claims, both primary and secondary sources were utilized. Secondary sources included books, journals, and global online research focused on machine learning for fraud detection in insurance. The study aligned with the objective of detecting fraudulent vehicle claims and helped guide the selection of appropriate research methods.

For primary data, motor vehicle insurance proposal and claim forms were collected from Britam Insurance Company (see Appendices 3 and 4) to analyze the critical data fields used by insurers. Given the increasing number of fraudulent claims and the resulting financial losses in the vehicle insurance sector, there is a strong need for a real-time fraud detection system.

6.2.2 Data Understanding

In this step, we started by obtaining relevant data for the study, then familiarized ourselves with the data, assessed its quality, gained a fundamental understanding of the data, and extracted variables from the dataset to aid in the model construction. We obtained an online CSV file on vehicle insurance claims dataset from Kaggle (2018).

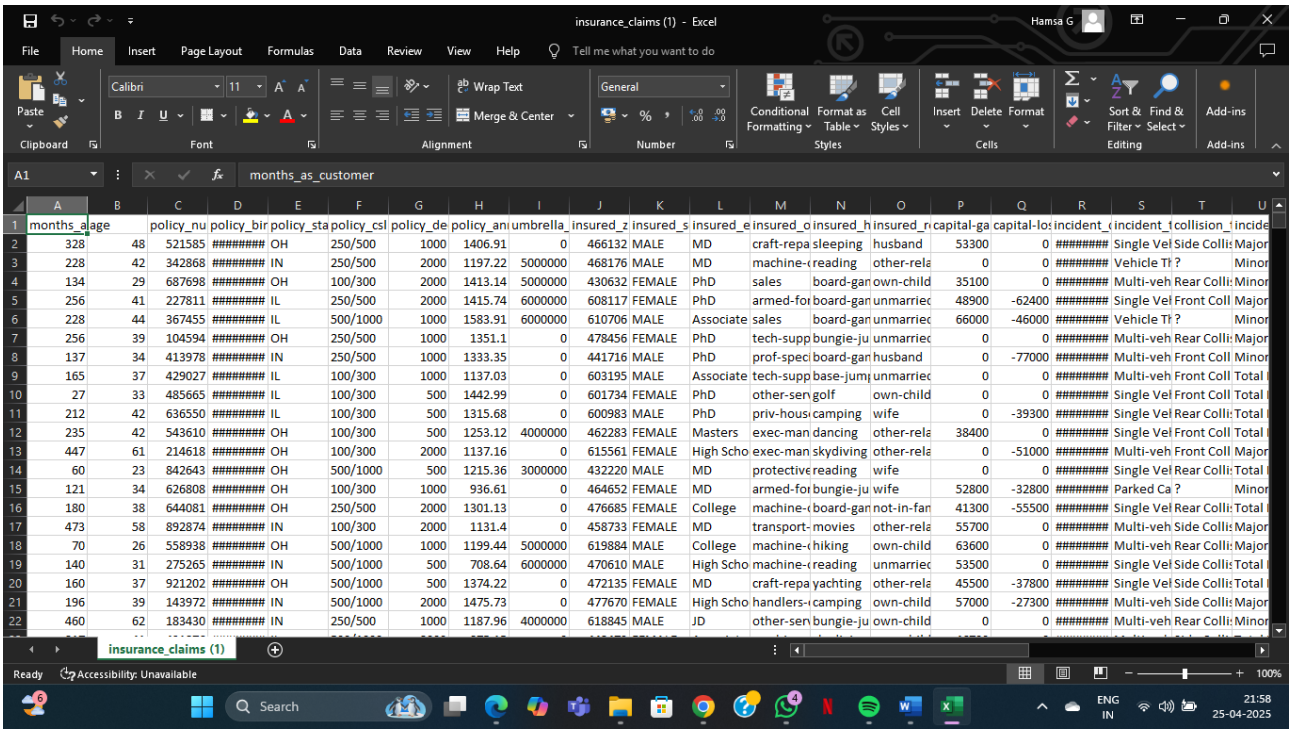


Figure 6.2.2.1 Vehicle Insurance Claims CSV File Extract

The dataset was distributed with data representing about 247 fraudulent claims, which made up 24.7% of the data, and 753 genuine claims, which made up 75.3% of the data, as seen in the bar graph in figure 6 below.

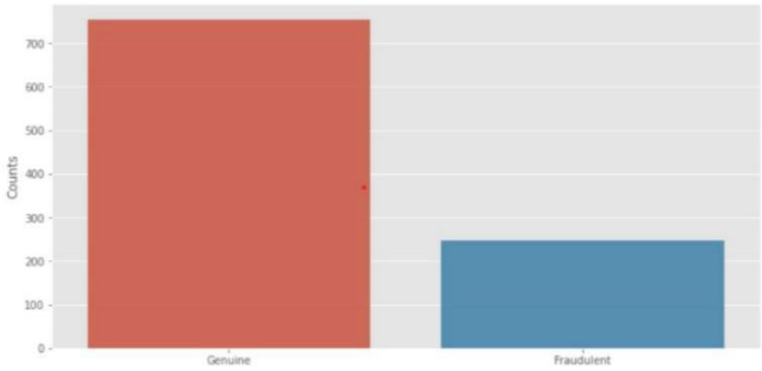


Figure 6.2.2.2 Vehicle Insurance Claims Distribution

A total of 1000 rows and 39 columns made up the dataset, which also included the following input variables as denoted in figure 7 below: *age, policy_number, policy_bind date, policy_state, policy_deductible, policy_annual premium, months_as customer insured, zip_code, insured sex, insured_level_of_education, insured _job, insured _pastimes, insured_relationship, covered_ capital gains, insured_ capital loss, occurrence date, occurrence type, collision type,*

incident_severity, authorities_contacted, occurrence _state, occurrence_city, occurrence_location, occurrence_hour_of_the_day, vehicle_make, vehicle_model, witnesses, occurrence_bodily injuries, occurrence_number_of_vehicles_involved, occurrence_ police_report _available, occurrence_total _claim amount, occurrence_property_damage, vehicle _year_of _manufacture and label to denote a claim as either genuine or fraudulent. Some of the input variables were later used to generate the features that were used to train the eight models used in this study, while others were eliminated for failing to reach the predetermined threshold.

```
Data columns (total 39 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   months_as_customer           1000 non-null   int64
 1   age                          1000 non-null   int64
 2   policy_number                1000 non-null   int64
 3   policy_bind_date             1000 non-null   object
 4   policy_state                 1000 non-null   object
 5   policy_csl                   1000 non-null   object
 6   policy_deductable            1000 non-null   int64
 7   policy_annual_premium        1000 non-null   float64
 8   umbrella_limit               1000 non-null   int64
 9   insured_zip                  1000 non-null   int64
 10  insured_sex                  1000 non-null   object
 11  insured_education_level      1000 non-null   object
 12  insured_occupation           1000 non-null   object
 13  insured_hobbies              1000 non-null   object
 14  insured_relationship         1000 non-null   object
 15  capital-gains                1000 non-null   int64
 16  capital-loss                 1000 non-null   int64
 17  incident_date                1000 non-null   object
 18  incident_type                1000 non-null   object
 19  collision_type               822 non-null    object
 20  incident_severity            1000 non-null   object
 21  authorities_contacted        1000 non-null   object
 22  incident_state               1000 non-null   object
 23  incident_city                1000 non-null   object
 24  incident_location            1000 non-null   object
 25  incident_hour_of_the_day     1000 non-null   int64
 26  number_of_vehicles_involved  1000 non-null   int64
 27  property_damage              640 non-null    object
 28  bodily_injuries              1000 non-null   int64
 29  witnesses                    1000 non-null   int64
 30  police_report_available      657 non-null    object
 31  total_claim_amount           1000 non-null   int64
 32  injury_claim                 1000 non-null   int64
 33  property_claim               1000 non-null   int64
 34  vehicle_claim                1000 non-null   int64
 35  auto_make                    1000 non-null   object
 36  auto_model                   1000 non-null   object
 37  auto_year                    1000 non-null   int64
 38  fraud_reported               1000 non-null   object
```

Figure 6.2.2.3 Dataset Columns showing Input Variables

The dataset characteristics are summarized in table 1 below.

Number of Claims	1000
Number of Attributes	39
Categorical Attributes	24
Genuine Claims	753
Fradulent Claims	247

Fradulent Claims Incident Rate	24.7%

Table 6.2.2 Dataset Features

The extracted dataset was also not entirely clean because several input variables had null values, as can be seen in figure 8 below, where collision type variable was missing 178 values, property_damage was missing 360 values, and police_ report_available variable was missing 343 values. However, no missing values were found in the data for the other input variables.

```
months_as_customer             0
age                            0
policy_number                  0
policy_bind_date               0
policy_state                   0
policy_csl                     0
policy_deductable              0
policy_annual_premium          0
umbrella_limit                 0
insured_zip                    0
insured_sex                    0
insured_education_level        0
insured_occupation             0
insured_hobbies                0
insured_relationship           0
capital-gains                  0
capital-loss                   0
incident_date                  0
incident_type                  0
collision_type               178
incident_severity              0
authorities_contacted          0
incident_state                 0
incident_city                  0
incident_location              0
incident_hour_of_the_day       0
number_of_vehicles_involved    0
property_damage              360
bodily_injuries                0
witnesses                      0
police_report_available      343
total_claim_amount             0
injury_claim                   0
property_claim                 0
vehicle_claim                  0
auto_make                      0
auto_model                     0
auto_year                      0
fraud_reported                 0
```

Figure 6.2.2.4 Dataset Columns showing Null Values

6.2.3 Data Preparation

Because the dataset was acquired in raw format, pre-processing was required to generate high-quality features that would be presented to the ML classifiers. To analyse and choose quality features, this

study employs a classical exploratory approach to data analysis. For machine learning to produce accurate and insightful results, data pre-processing is a crucial step. The reliability of the outcomes is inversely correlated with data quality. Real-world datasets are imperfect, inconsistent, and noisy in nature. Data pre-processing improves the data quality by addressing the gaps in the data, reducing noise, and addressing inconsistencies. Data preparation, according to Pandey (2019), entails cleaning, integrating, transforming, and reducing data to eliminate any duplicate or irrelevant data, leaving just the bits that provide valuable information to aid in establishing an efficient and effective classification. The stages in the procedure are as follows:

- Data cleaning which aims to remove outliers from the dataset and impute missing values.
- Application of data transformation techniques like normalization. For instance, normalization may increase the precision and effectiveness of distance-based mining algorithms.
- Data integration, which combines data from several sources into one data warehouse.
- Data reduction, which involves removing redundant features from the data to lower its size.
- Techniques for feature extraction and feature selection can be used.

6.2.3.1 Data Clean-Up

The data preparation process stated by checking for duplicate records and missing values. The missing values were then replaced with specified values using the fillna python method in the dataset. Figure 9 below demonstrates the checking of duplicate and null values and replacement of null values with specifies values.

```
#Checking for duplicate claims
df.drop_duplicates(inplace = True)
df.shape
```

```
#We replace missing values with np.nan
df.replace('?', np.nan, inplace = True)
```

Handling missing values

```
df['collision_type'] = df['collision_type'].fillna(df['collision_type'].mode()[0])
df['property_damage'] = df['property_damage'].fillna(df['property_damage'].mode()[0])
df['police_report_available'] = df['police_report_available'].fillna(df['police_report_available'].mode()[0])
```

Figure 6.2.3.1.1 Checking and Filling Null Values

The dependent variable, fraud reported, was used as the starting point for exploratory data analysis. Heatmaps were created for variables with at least a 0.3 Pearson's correlation coefficient, including the dependent variable, to better visualize the input variables within the dataset, aid in directing attention to areas of data visualizations that matter the most, and examine the relationships between

them. The Pearson's correlation coefficient, which measures the linear relationship between two sets of data, is the ratio of the standard deviations of two continuous variables. Because the result is always between -1 and 1, it is effectively a normalized measurement of covariance (Statistics Solutions,2022).

The heatmap analysis in figure 10 below demonstrates a strong correlation between month_as_customer and age, with a correlation of 0.92. This is most likely because people get vehicle insurance when they own a car, and because the time measure simply increases with age, therefore the "age" variable was dropped. The total _claim variable was also dropped because it was discovered that there was a strong correlation between the total _claim_amount, injury_claim, property_claim, and vehicle claim variables. Additionally, to avoid redundancy, several of the data variables with high correlation were dropped. Afterwards, it was noticed there seemed not to be any multicollinearity issues, other from the possibility that all the claims are correlated and the total claims have been considered. On the other hand, the other claims offer a level of granularity that is not otherwise covered by total claims. As a result, these variables were retained.

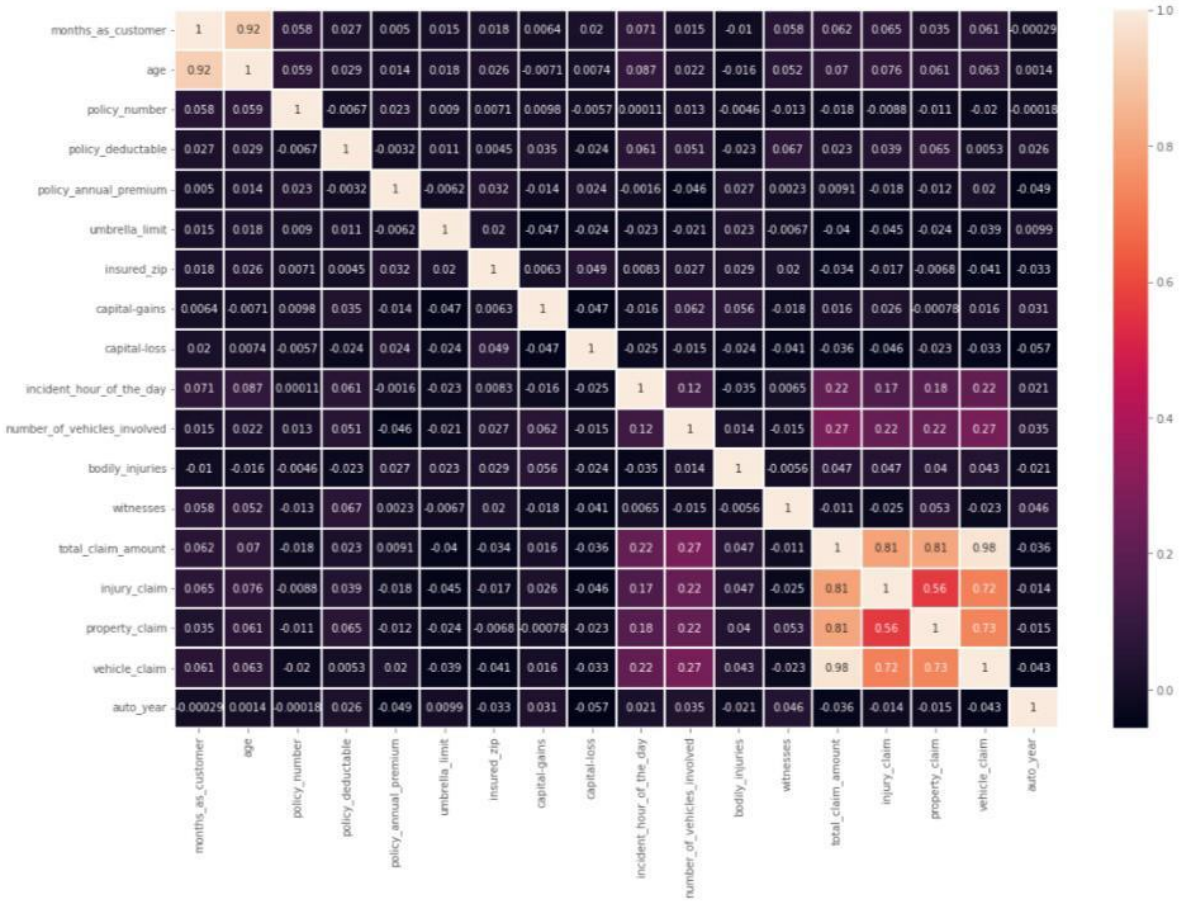


Figure 6.2.3.1.2 Correlation Heatmap among Data Variables

Additional analysis was performed on the remaining data variables by identifying the unique values to obtain the reputable features to be utilized for classification, in addition to employing multicollinearity to eliminate some features as indicated on the heatmap in figure 10 above. A feature was eliminated if it had many unique values because there wasn't anything to be learned from them. Figure 11 below illustrates the number of unique values present in the remaining data variables.

As a result, the following features were removed: policy_ number, policy bind_date, policy_state, insured zip, incident_location, incident_date, incident_state, incident_city, insured_hobbies, auto_make, auto_model and auto_year.

months_as_customer	391
age	46
policy_number	1000
policy_bind_date	951
policy_state	3
policy_csl	3
policy_deductable	3
policy_annual_premium	991
umbrella_limit	11
insured_zip	995
insured_sex	2
insured_education_level	7
insured_occupation	14
insured_hobbies	20
insured_relationship	6
capital-gains	338
capital-loss	354
incident_date	60
incident_type	4
collision_type	3
incident_severity	4
authorities_contacted	5
incident_state	7
incident_city	7
incident_location	1000
incident_hour_of_the_day	24
number_of_vehicles_involved	4
property_damage	2
bodily_injuries	3
witnesses	4
police_report_available	2
total_claim_amount	763
injury_claim	638
property_claim	626
vehicle_claim	726
auto_make	14
auto_model	39
auto_year	21
fraud_reported	2

Figure 6.2.3.1.3 Unique Values Present in Data Variables

6.2.3.2 Data Transformation

Data was transformed into formats that machine learning classifiers could interpret. For instance, text values must be converted into integer values since machine learning classifiers cannot interpret text values. We converted the categorical data into integer format to enable categorical data encoding which enables categorical values to be fed into different models. This improved the predictions of our models. For the models to use the data with converted categorical values to produce and enhance the predictions, Verma (2021) defines categorical data encoding as the process of turning categorical data into integer format. He continued by defining categorical data as information that has been obtained and is organized into groups and has a limited number of possible values. The dataset's categorical data was extracted for conversion, and each column's unique values printed. Policy_csl, insured_sex. insured_education level, insured_occupation, insured relationship, incident _type, collision_type. incident_severity, authorities_contacted, property_damage, and police_report _available were the columns extracted. Figure 12 and 13 displays the unique values of the retrieved categorical data columns and the converted categorical data into integer format respectively.

```
policy_csl:
['250/500' '100/300' '500/1000']

insured_sex:
['MALE' 'FEMALE']

insured_education_level:
['MD' 'PhD' 'Associate' 'Masters' 'High School' 'College' 'JD']

insured_occupation:
['craft-repair' 'machine-op-inspct' 'sales' 'armed-forces' 'tech-support'
 'prof-specialty' 'other-service' 'priv-house-serv' 'exec-managerial'
 'protective-serv' 'transport-moving' 'handlers-cleaners' 'adm-clerical'
 'farming-fishing']

insured_relationship:
['husband' 'other-relative' 'own-child' 'unmarried' 'wife' 'not-in-family']

incident_type:
['Single Vehicle Collision' 'Vehicle Theft' 'Multi-vehicle Collision'
 'Parked Car']

collision_type:
['Side Collision' 'Rear Collision' 'Front Collision']

incident_severity:
['Major Damage' 'Minor Damage' 'Total Loss' 'Trivial Damage']

authorities_contacted:
['Police' 'None' 'Fire' 'Other' 'Ambulance']

property_damage:
['YES' 'NO']

police_report_available:
['YES' 'NO']
```

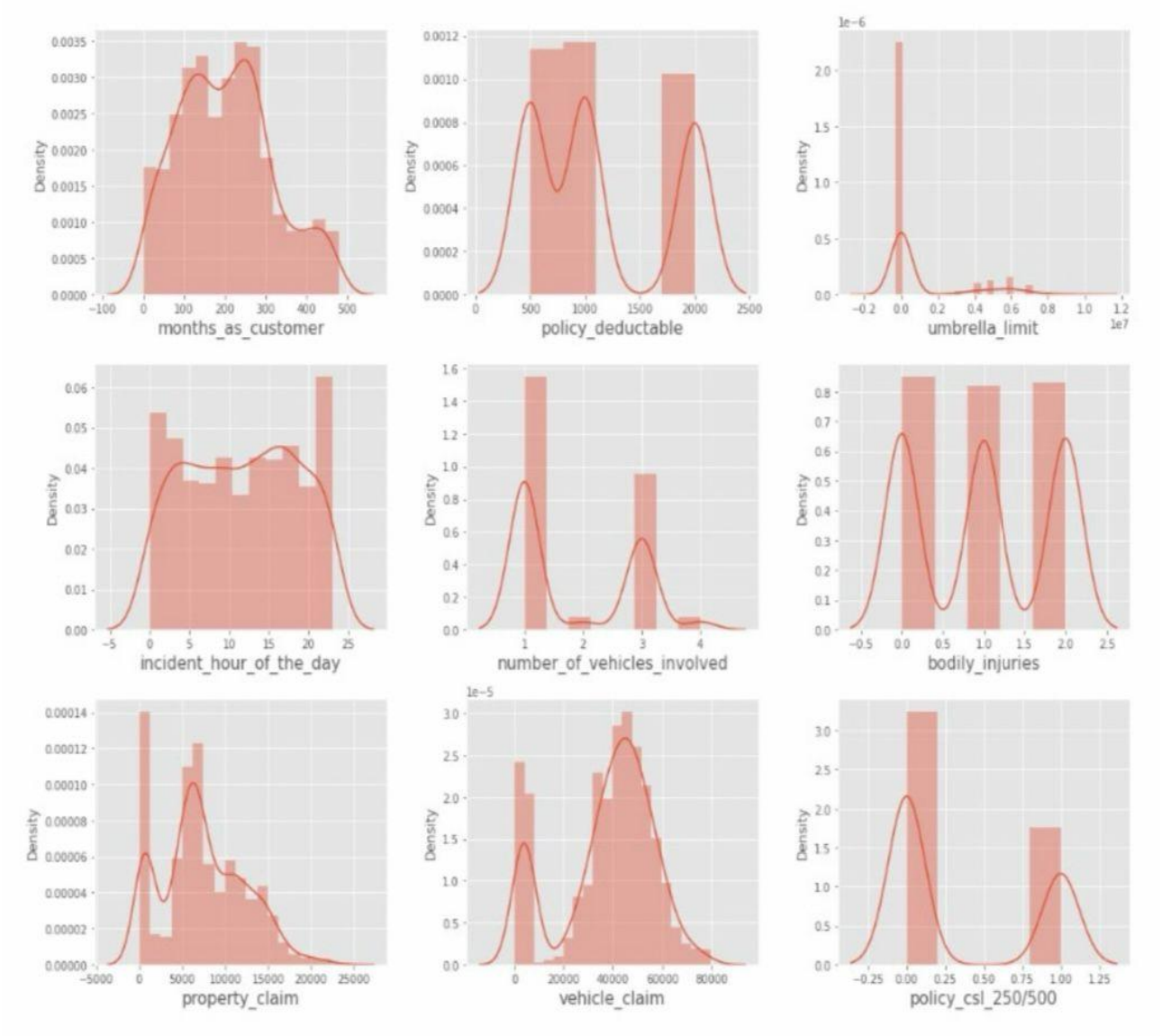
Figure 6.2.3.2.1 Categorical Data Columns Unique Values

	policy_csl_250/500	policy_csl_500/1000	insured_sex_MALE	insured_education_level_College	insured_education_level_High School
0	1	0	1	0	0
1	1	0	1	0	0
2	0	0	0	0	0

Figure 6.2.3.2.2 Converted Categorical Data Columns into Integer Values

6.2.3.3 Data Integration

To create the final dataset that would be utilized for both training and testing the various models, the columns that included numerical values were also extracted and combined with the converted numerical values from the categorical data. Figure 14 displays the distribution plot to show the variation in the final dataset's data distribution.



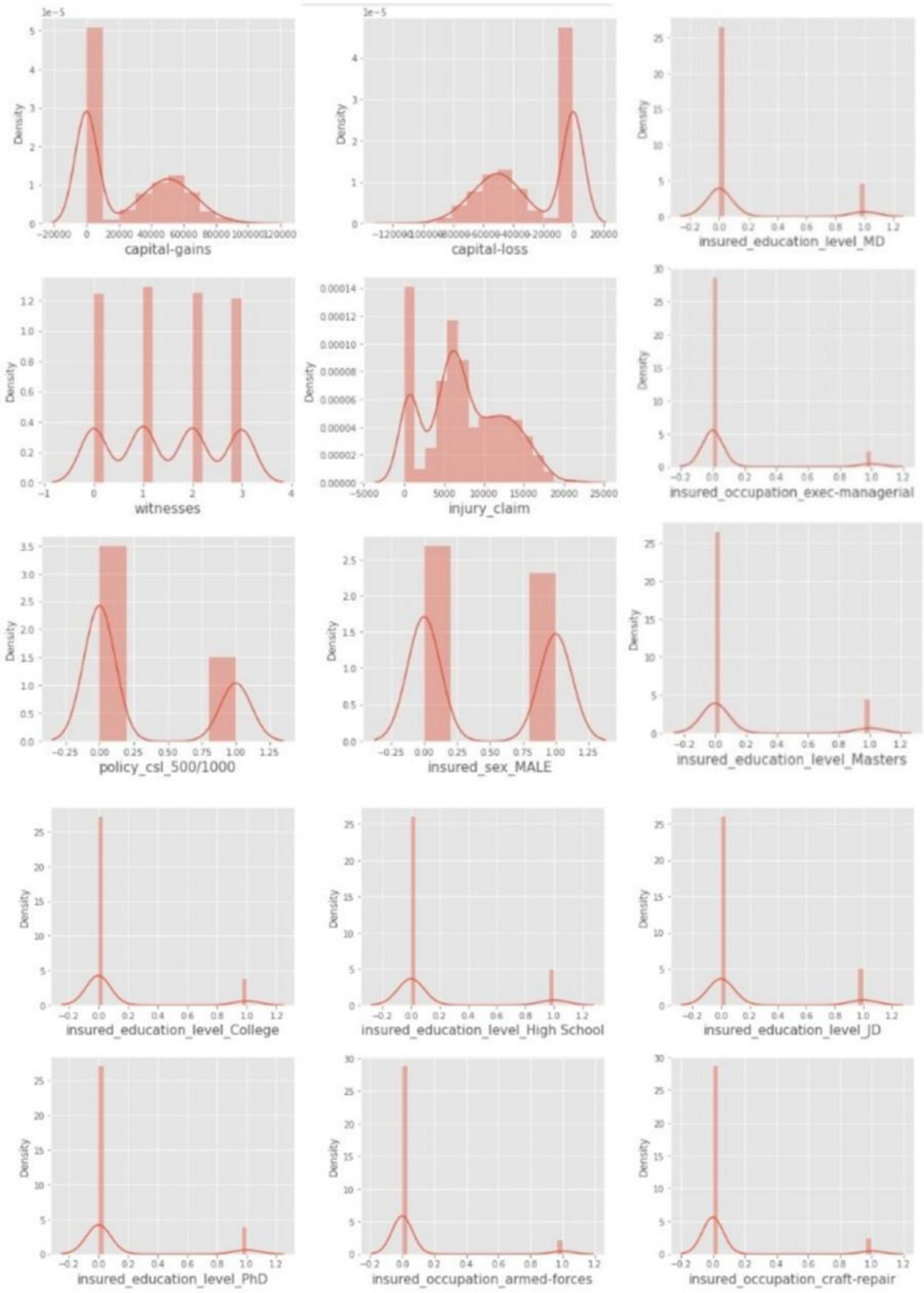


Figure 6.2.3.3 Final Dataset Data Distribution Plot

6.2.3.4 Feature Selection

We performed an analysis on the data to identify anomalies and outliers before splitting the dataset into training and testing sets. As a result, we scaled the numerical columns that had the outliers and deemed the data suitable for training. According to Tang et al. (2016), an outlier is an observation that deviates significantly from other values in a sample drawn at random from a population, almost as if the data were produced differently, or the potential for a data collecting error. The Inter Quantile Range (IQR) was utilized in this study to identify outliers. According to Tang et al. (2016), when values are sorted from lowest to highest, the IQR describes the median 50% of those values as shown in figure 15 below. The median (middle value) of the lower and upper half of the data is found first before calculating the IQR. These numbers are in the first quartile (Q1) and third quartile (Q3). The difference between Q3 and Ql is the IQR. Some numerical columns were found to contain outliers; thus, scaling was applied to them.

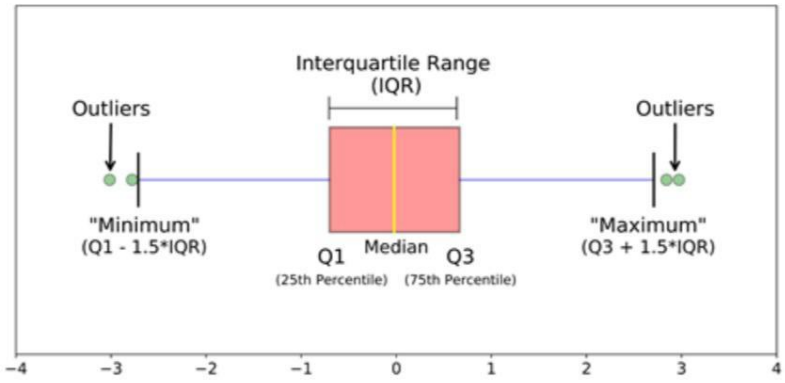


Figure 6.2.3.4.1 Inter Quantile Range Calculation Graph

The remaining dataset also included several attributes derived from the two variables that were identified for this study, such as information about insurance policies and information about insurance claims that matched the targeted level of fraud in the vehicle insurance claims. The following features as shown in figure 16 below were maintained for machine learning models classification as they satisfied the two variables identified for this study: *months as customer, policy csl, policy _deductable, policy annual premium, umbrella limit, insured sex, insured education level, insured occupation, insured relationship, capital-gains, property damage, bodily injuries, witnesses, incident_hour_of_the_day, number_of_vehicles_ involved, injury_claim, property_claim. vehicle _claim and fraud reported.*

```
Data columns (total 25 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   months_as_customer           1000 non-null   int64
 1   policy_csl                   1000 non-null   object
 2   policy_deductable            1000 non-null   int64
 3   policy_annual_premium        1000 non-null   float64
 4   umbrella_limit               1000 non-null   int64
 5   insured_sex                  1000 non-null   object
 6   insured_education_level      1000 non-null   object
 7   insured_occupation           1000 non-null   object
 8   insured_relationship         1000 non-null   object
 9   capital-gains                1000 non-null   int64
 10  capital-loss                 1000 non-null   int64
 11  incident_type                1000 non-null   object
 12  collision_type               1000 non-null   object
 13  incident_severity            1000 non-null   object
 14  authorities_contacted        1000 non-null   object
 15  incident_hour_of_the_day     1000 non-null   int64
 16  number_of_vehicles_involved  1000 non-null   int64
 17  property_damage              1000 non-null   object
 18  bodily_injuries              1000 non-null   int64
 19  witnesses                    1000 non-null   int64
 20  police_report_available      1000 non-null   object
 21  injury_claim                 1000 non-null   int64
 22  property_claim               1000 non-null   int64
 23  vehicle_claim                1000 non-null   int64
 24  fraud_reported               1000 non-null   object
```

Figure 6.2.3.4.2 Features Maintained for Classification

6.2.4 Modeling

In this study, eight machine learning models—XGBoost, AdaBoost, SVM, NB, RF, ANN, DT, and LR—were trained and tested to identify the best-performing algorithm for detecting fraudulent vehicle insurance claims using both unbalanced and balanced datasets. The dataset was split 80% for training and 20% for testing.

To address class imbalance, the SMOTE (Synthetic Minority Oversampling Technique) method was applied, generating synthetic samples for the minority class. After balancing, the data was again split into training and testing sets.

To ensure robust evaluation, 10-fold cross-validation was used. This technique divides the dataset into 10 parts, rotating each as the test set while training on the remaining nine, minimizing model bias and improving performance reliability.

6.2.4 Experiment Environment

The study employed Google Colaboratory Notebook, popularly known as Colab, for modeling purposes (Google Inc., 2017). The Google notebook provides simple data sharing, allowing programmers to write and run Python in their browsers with no setup fees and free Graphics Processing Unit (GPU) access.

6.2.5 Evaluation

After all the classifiers had been trained with both balanced and unbalanced datasets, the model's performances were evaluated using the test data to see if they could categorize claims as genuine or fraudulent. An analysis of the performance categorization indicators was done in this study to gauge the model's efficiency and effectiveness, as well as establish their risk threshold. Confusion matrix,

classification accuracy, classification report based on recall, precision, and F-1 score were the metrics employed. This found which classifier had the best levels of prediction performance and classification accuracy.

6.2.5.1 Confusion Matrix

A confusion matrix, according to Parab (2020), is a performance classification metric used to assess a machine learning algorithm's performance based on target classes. To determine the classification metrics above, the following values were first computed using a confusion matrix:

- True Positives (TP) - The amount of fraudulent vehicle insurance claims that were discovered.
- False Negatives (FN) - The amount of fraudulent vehicle insurance claims that went undiscovered.
- False Positives (FP) - The number of genuine vehicle insurance claims that were incorrectly categorized as fraudulent.
- True Negative (TN) - The proportion of genuine vehicle insurance claims that were not flagged as fraudulent.

6.2.5.2 Accuracy

According to Parab (2020), accuracy is the proportion of accurately predicted observations (True Positives) to all the input observations (sum of True Positives, False Positives, False Negatives, True Negatives).

$$Accuracy = \frac{Number\ of\ Correct\ Predictions\ (TP + TN)}{Total\ Number\ of\ Predictions\ Made\ (TP + TN + FP + FN)}$$

Figure 6.2.5.2 Accuracy

6.2.5.3 Precision

A measure of precision is the proportion of accurately predicted positive samples (also known as True Positives) to the total number of predicted positive samples (sum of True Positives and False Positives) (Parab, 2020).

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Postives(FP)}$$

Figure 6.2.5.3 Precision

6.2.5.4 Recall

This is the proportion of correctly predicted positive samples (True Positives) to all samples in the actual class (sum of True Positives and False Negatives).

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives(FN)}$$

Figure 6.2.5.4 Recall

6.2.5.5 F-1 Score

F1 Score is the weighted average between precision and recall. The formula used to compute it is:

$$F1\ Score = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

Figure 6.2.5.5 F-1 Score

6.2.6 Deployment

An effective, novel system, based on the machine learning classifier with the highest levels of prediction performance and classification accuracy was developed to identify fraudulent vehicle insurance claims. The long-term profitability and consumer satisfaction of insurance businesses will benefit greatly from this.

CHAPTER 7 SYSTEM DESIGN AND IMPLEMENTATION

Power BI

Power BI is a Data Visualization and Business Intelligence tool that converts data from different data sources to interactive dashboards and BI reports. It aims to provide interactive visualizations and business intelligence capabilities to create their own reports and dashboards for the end users. Our data may be in an Excel spreadsheet, CSV file, or a collection of a cloud based. **Pbix** file which is designed for to use with Power BI desktop.

Microsoft Power BI is used to find insights within an organization's data. Power BI can help connect disparate data sets, transform and clean the data into a data model and create charts or graphs to provide visuals of the data. All of this can be shared with other Power BI users within the organization.

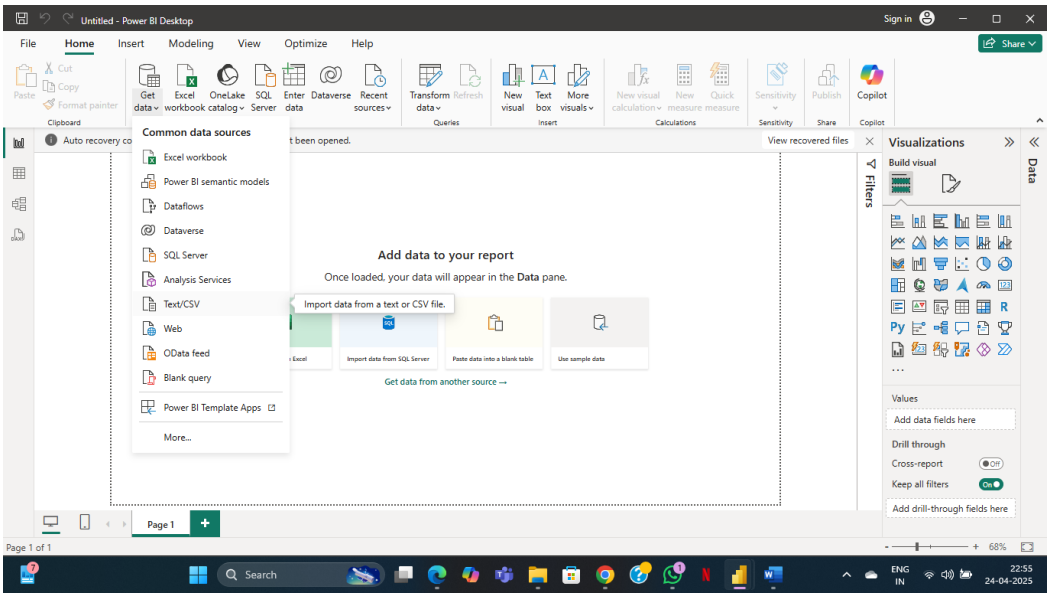
Power BI was initially released in 2014, operating system: Microsoft windows. Power BI suite provides multiple software, connector, and services - Power BI desktop, Power BI service based on Saas, and mobile Power BI apps available for different platforms. These set of services are used by business users to consume data and build BI reports. Power BI desktop app is used to create reports, while Power BI Services (Software as a Service - SaaS) is used to publish the reports, and Power BI mobile app is used to view the reports and dashboards. Power BI Desktop is available in both 32-bit and 64-bit versions.

To download the latest version, you can use the following link -
https://powerbi.microsoft.com/en-us/downloads/

Data Import:

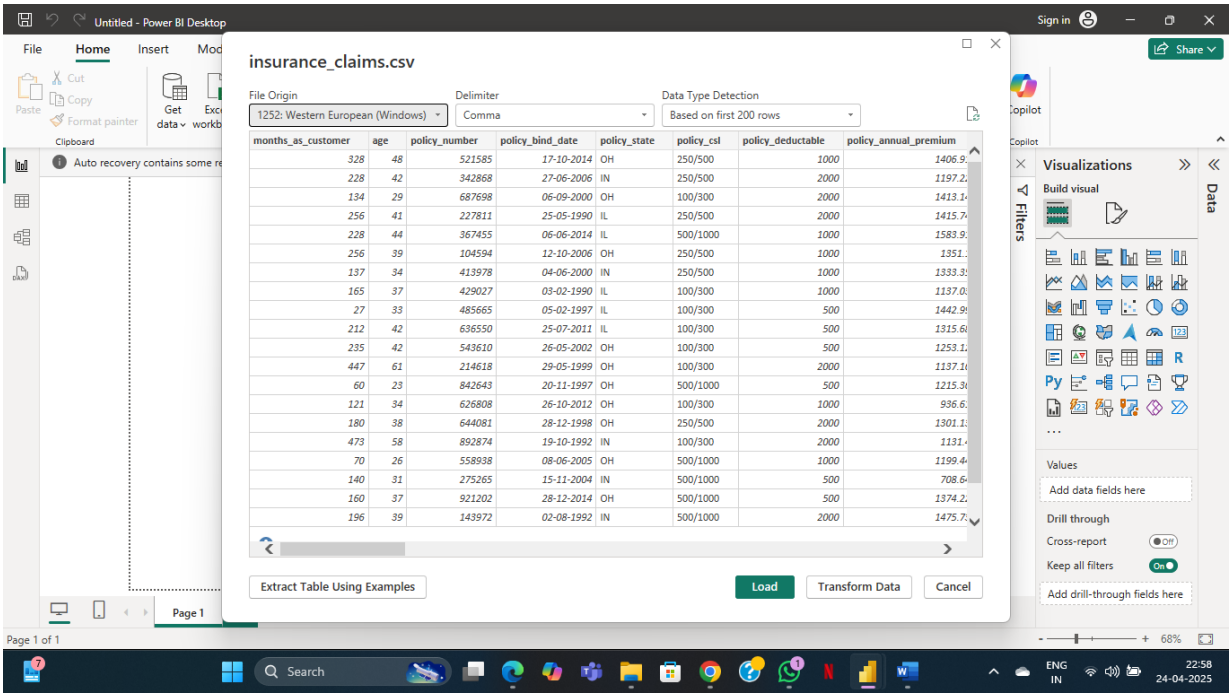
We import data from CSV file to Power BI desktop.

From the top menu choose "Get Data" -> Text/CSV

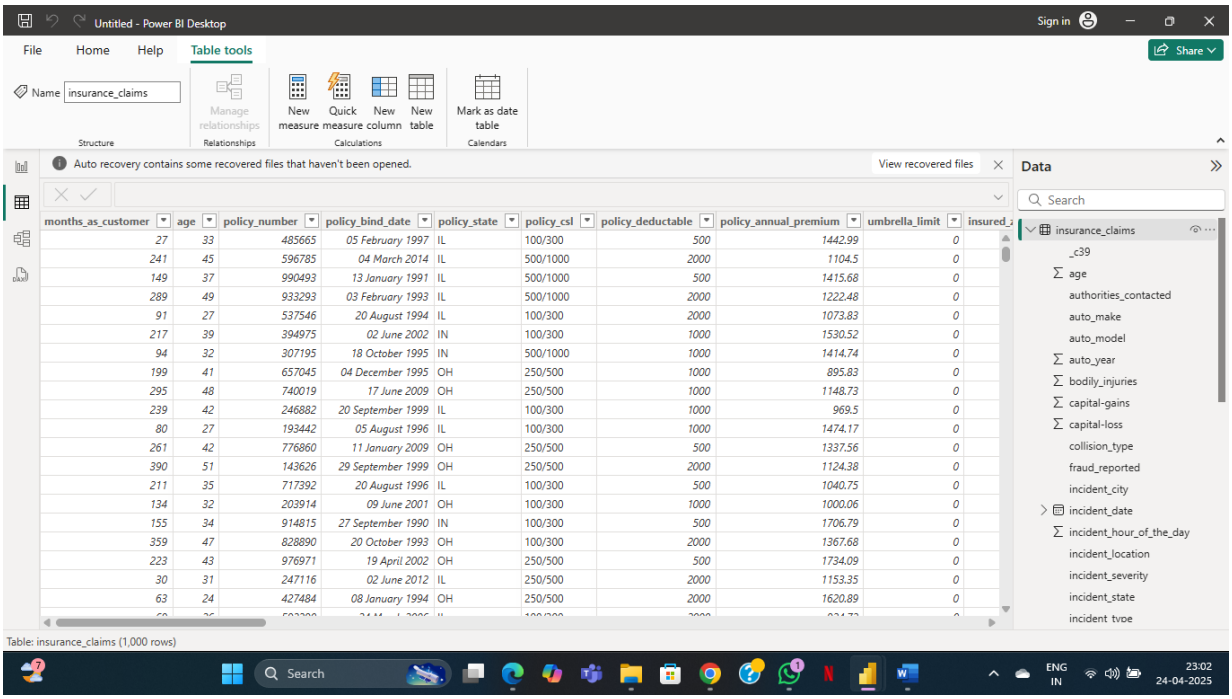


Open the file: insurance_claims.csv

Choose "Insurance Claims" worksheet (you will see the preview of the worksheet) and click "Load".



To see your data imported choose the icon on the left side of the screen.

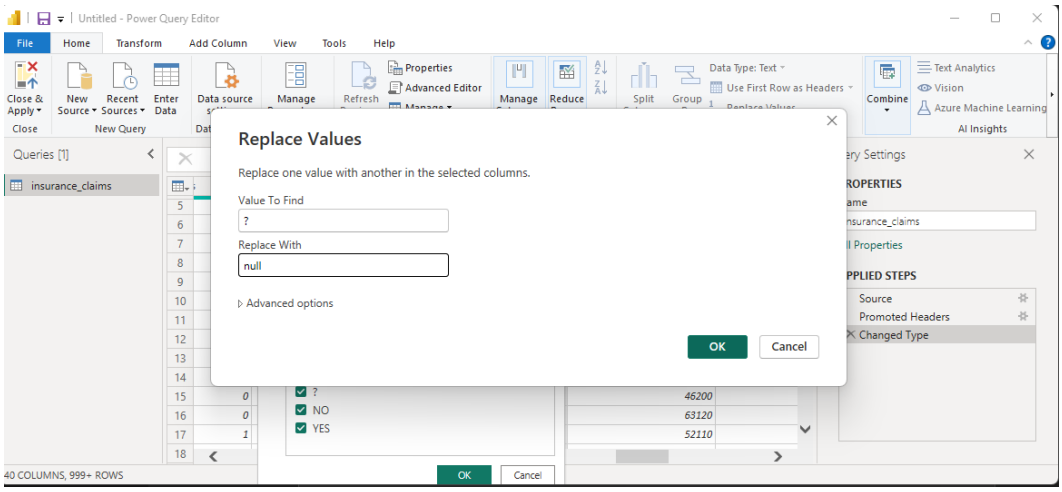


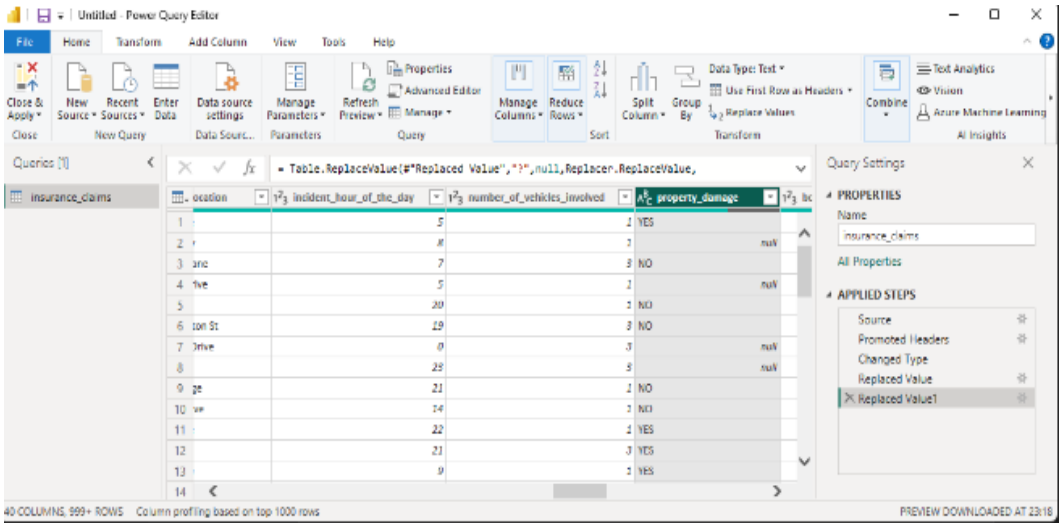
Data Preparation:

Step 1: Click **Transform Data** to open the **Power Query Editor**.

Step 2: Replace Missing Values

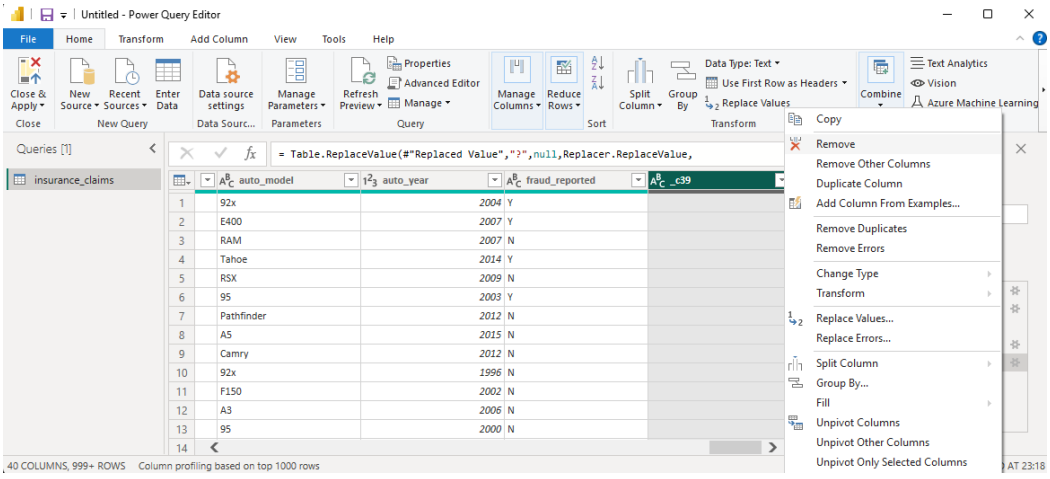
- In Power Query Editor:
  - Use **Home > Replace Values** or select the column → Right-click → Replace Values.
  - Replace ? with null value.





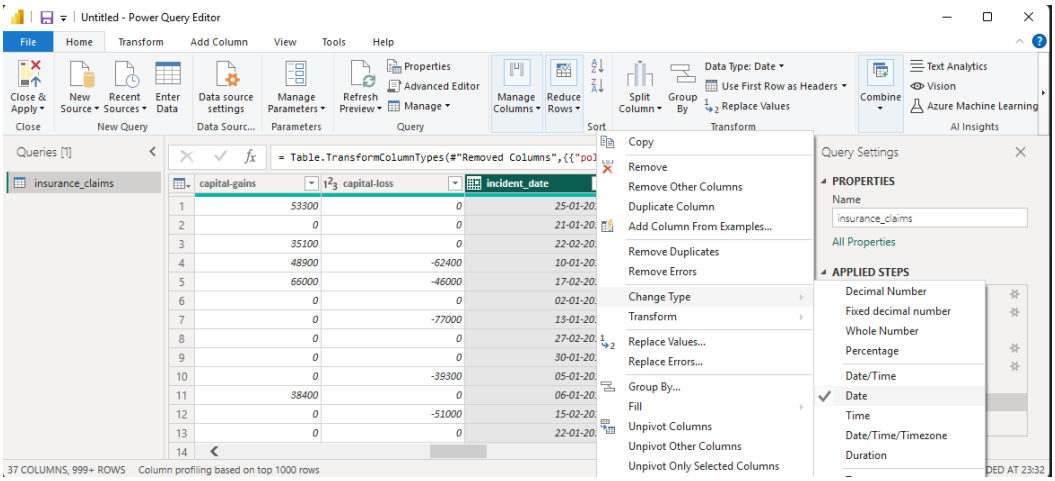
Step 3: Remove or Rename Columns

- Remove unwanted columns like _c39, policy_number, or incident_location:
  - Select the column → Right-click → **Remove**.
- Rename columns for clarity:
  - Double-click the column header and type a new name.



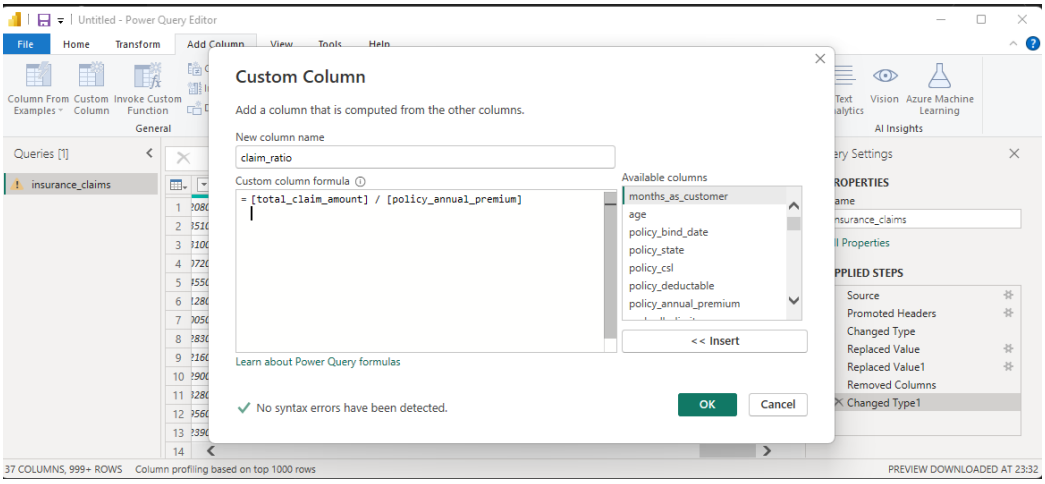
Step 4: Change Data Types

- For fields like policy_bind_date, incident_date, convert to **Date** type.
- Convert categorical fields (e.g., fraud_reported, insured_sex) to **Text**.
- Ensure numerical fields (e.g., policy_annual_premium, total_claim_amount) are set to **Decimal Number** or **Whole Number**.



Step 5: Create New Calculated Columns (if needed)

- **Example: Create a claim_ratio column:**
  - **Go to Add Column > Custom Column.**



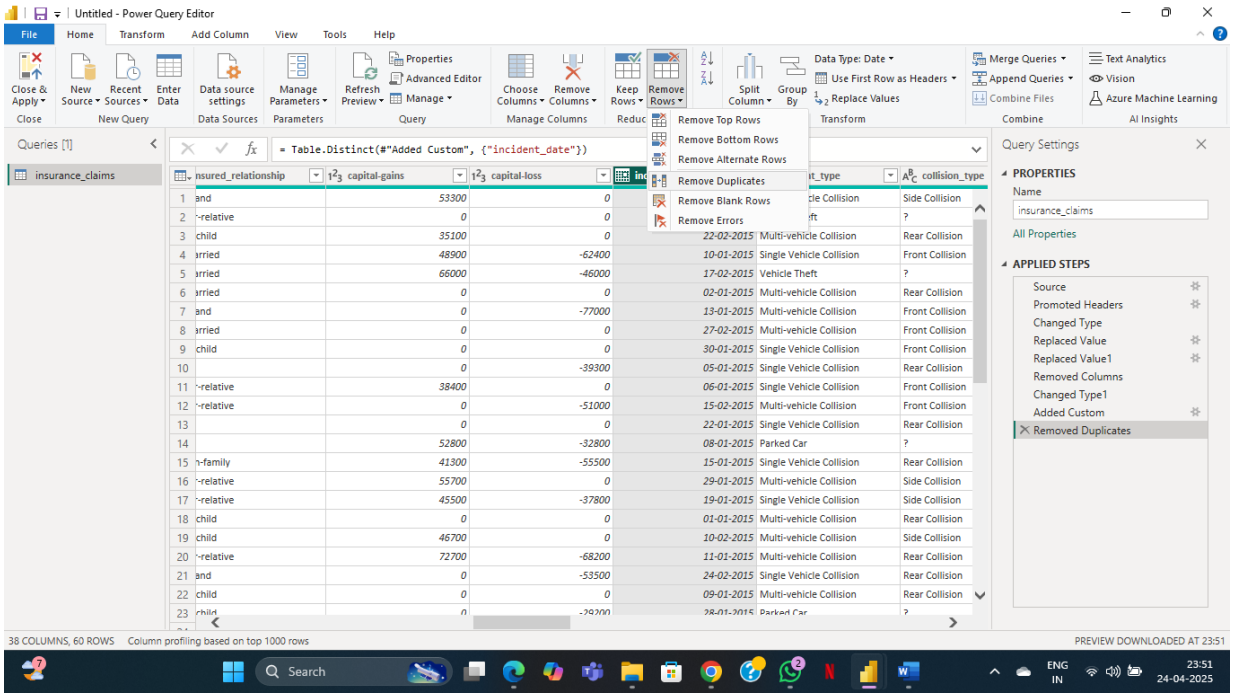
Step 6: Encode Categories (Optional for Modeling View)

Power BI supports categorical fields as-is for visuals. For ML in Power BI (e.g., with Azure ML), you'd need dummy encoding:

- **Use Transform > Pivot Column or Add Column > Conditional Column if necessary.**

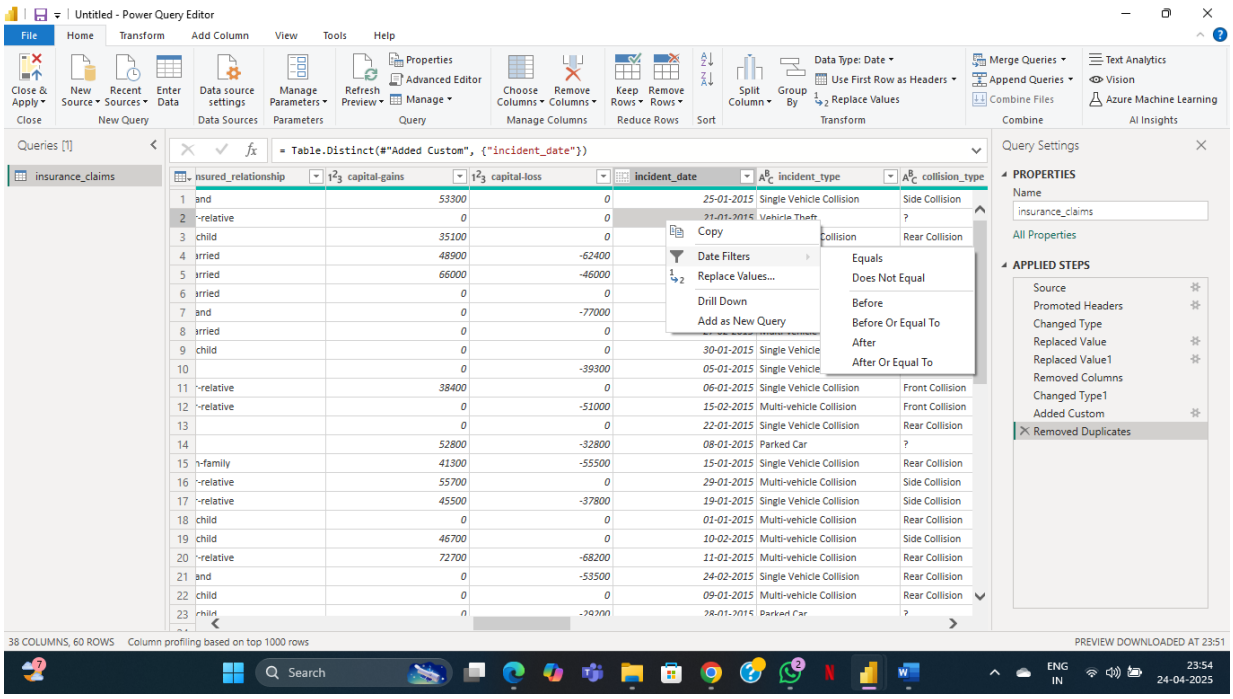
Step 7: Remove Duplicates (if any)

- **Go to Home > Remove Rows > Remove Duplicates.**
- **Select key columns like policy_number, incident_date, etc., to define uniqueness.**



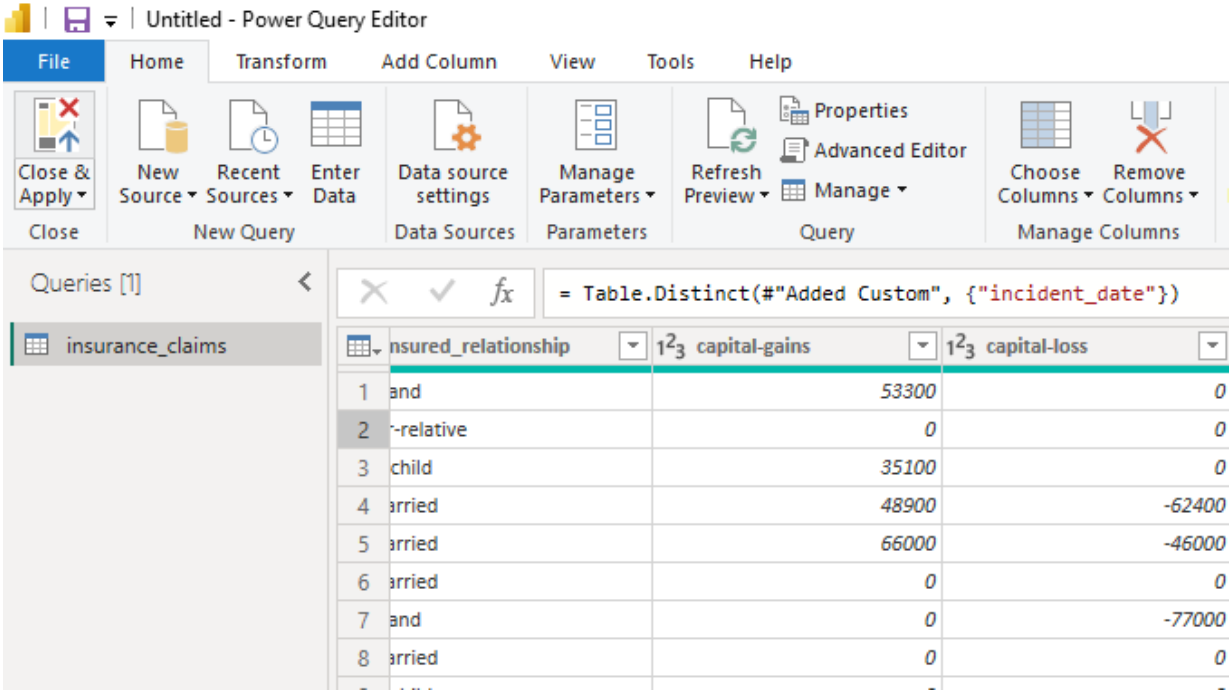
Step 8: Filter or Handle Outliers (Visual or Numeric Filters)

- Use **Filters** on columns to exclude outliers.
- Or use Conditional Column to flag outliers based on thresholds.



Step 9: Close & Apply

- **After all transformations, click Home > Close & Apply.**
- **The transformed data loads into Power BI for visualization.**

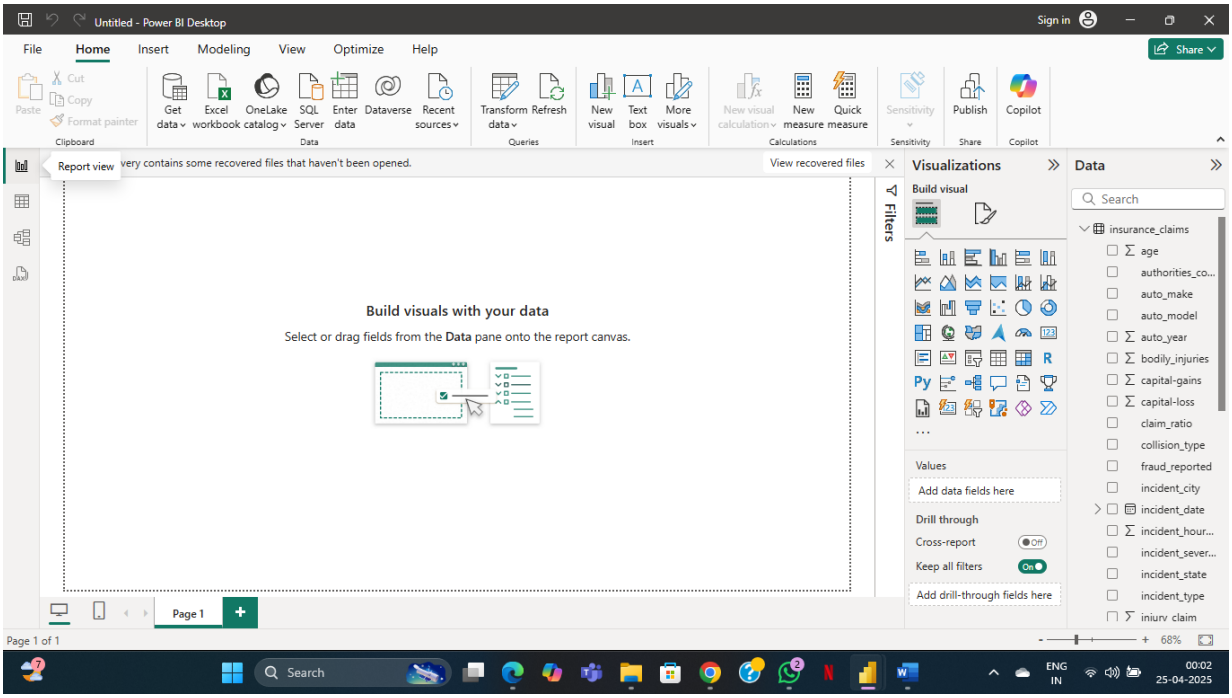


Visualizing data in charts and tables:

Now our task is to get some knowledge from the dataset.

Switch to the Report view.

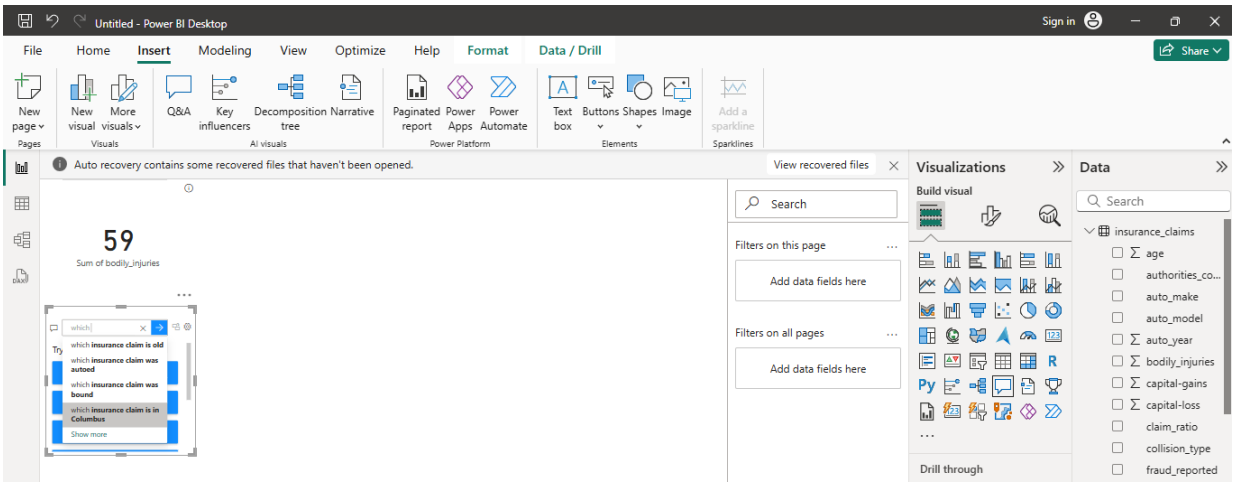
Here we can make new tables and charts.

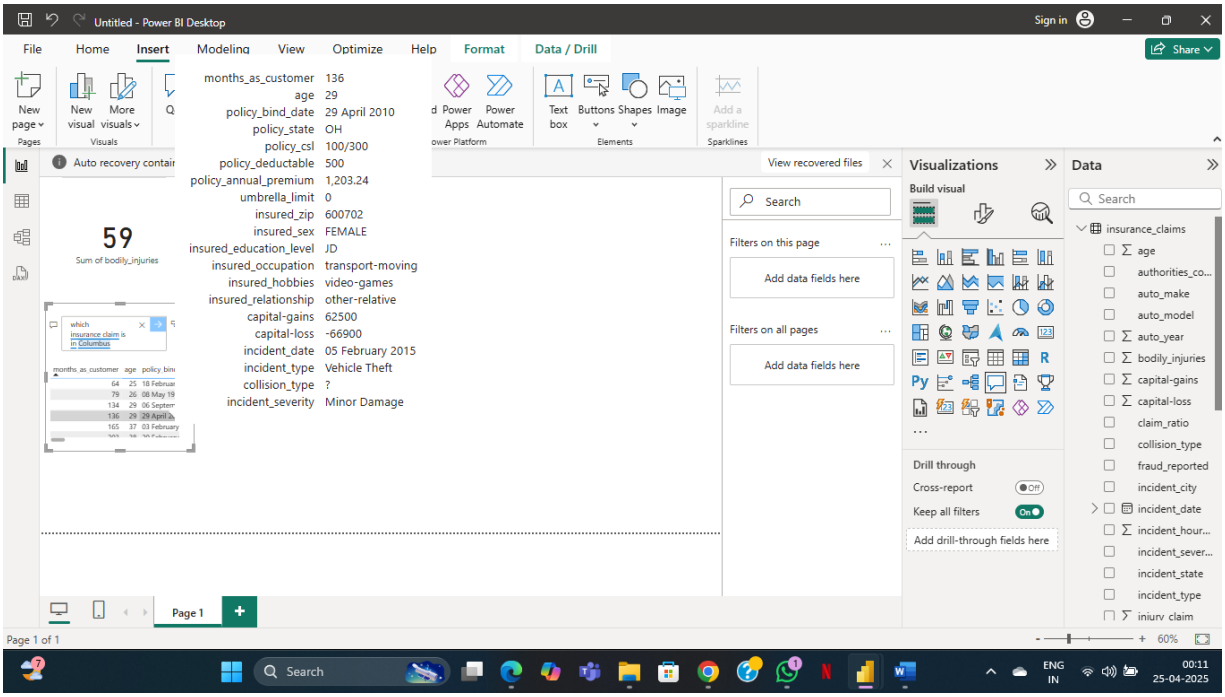


Natural language queries

In Power BI we can ask (not very sophisticated) queries in natural language. It is useful feature that saves time.

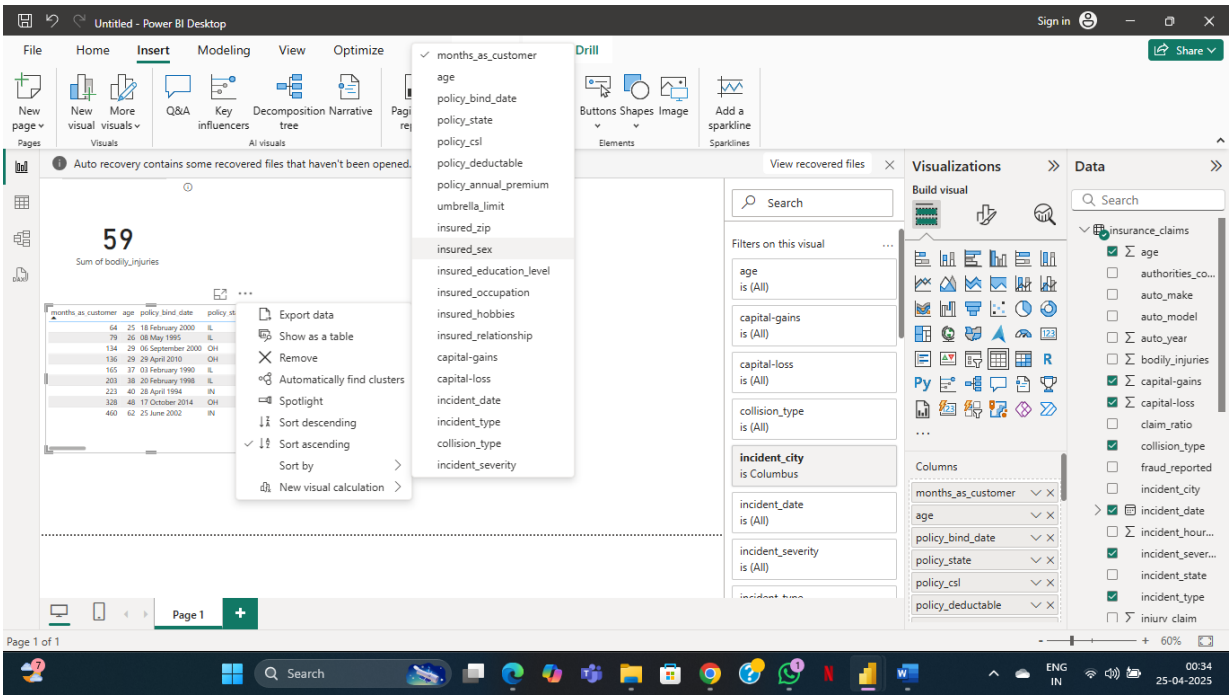
To try this feature, click on "Ask a Question"



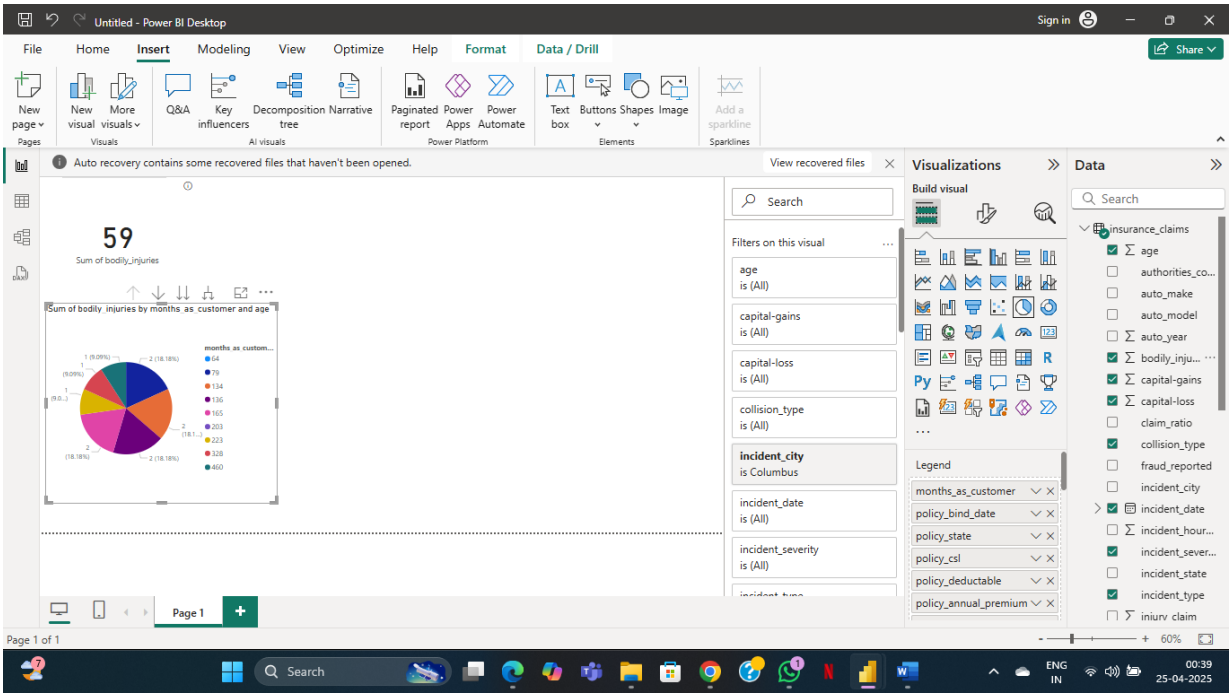


Changing the layout and format

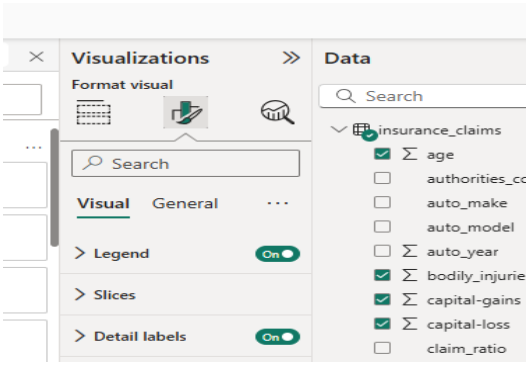
You can sort the data on the chart by sales value or manufacturer.



Try to add a pie chart to your report. Use the right side of the screen (drag&drop the fields)



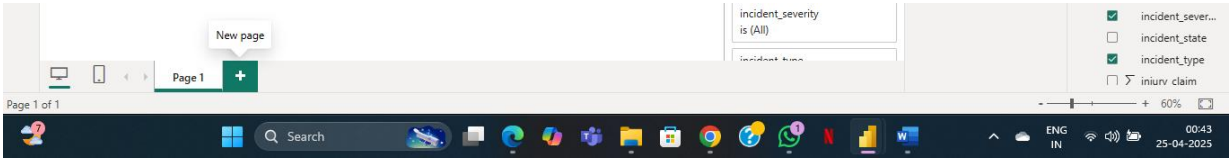
You can format the chart and any element of your report (font size, color, etc.) by clicking on painting tool



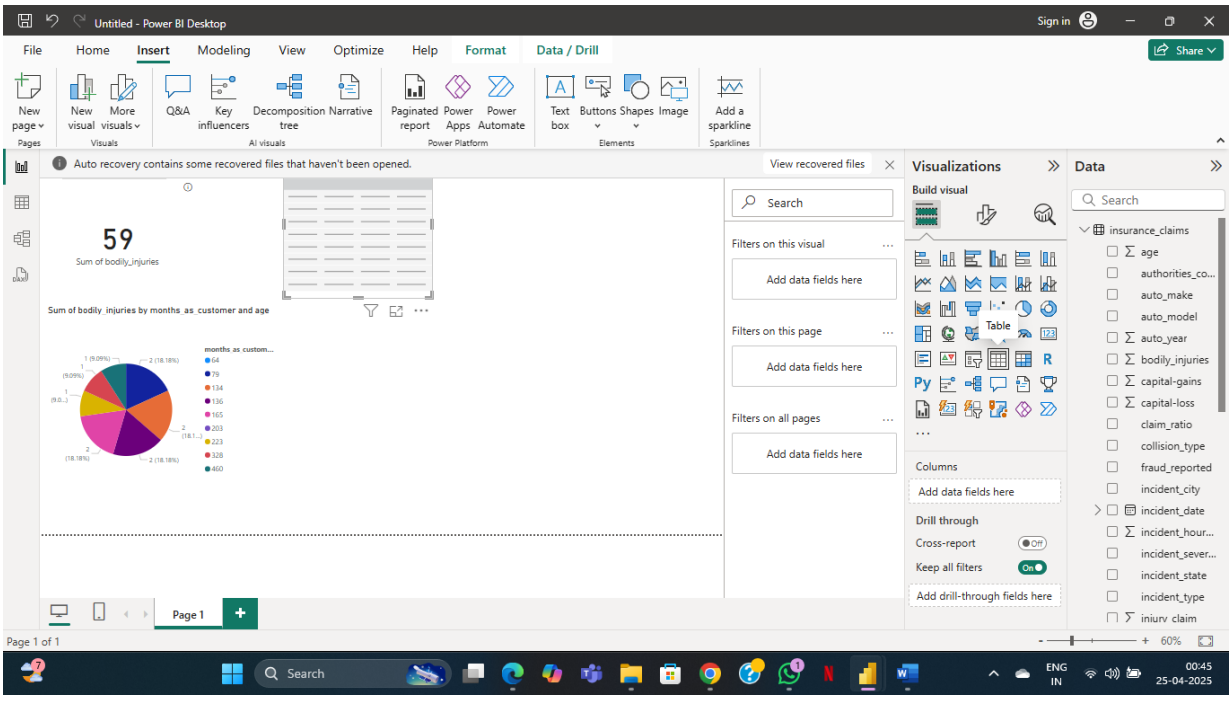
Try some options, feel free to choose colors and fonts, add a legend...

Adding tables and charts

We can compose our report on many pages. To add another page click on the green + at the bottom of the screen. You can name your new page.

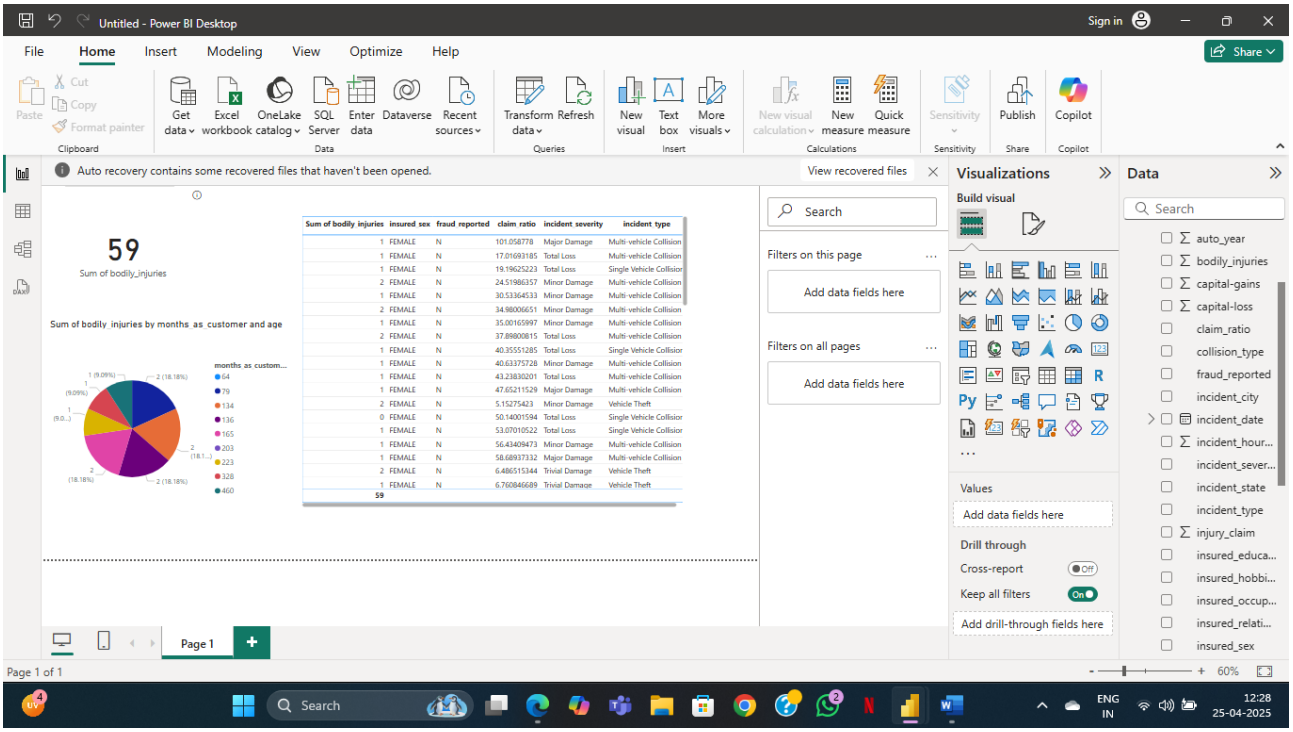


To add a table to your report choose from the visualizations menu on the right.



Add a table presenting AVERAGE Bodily Injured Insurer claiming insurance based on the aspects of their sex, fraud reported, claim ratio, incident severity and type of incidents occurred. Format the table (change colors and font size, sort the data)

Add a column chart and do formatting as well



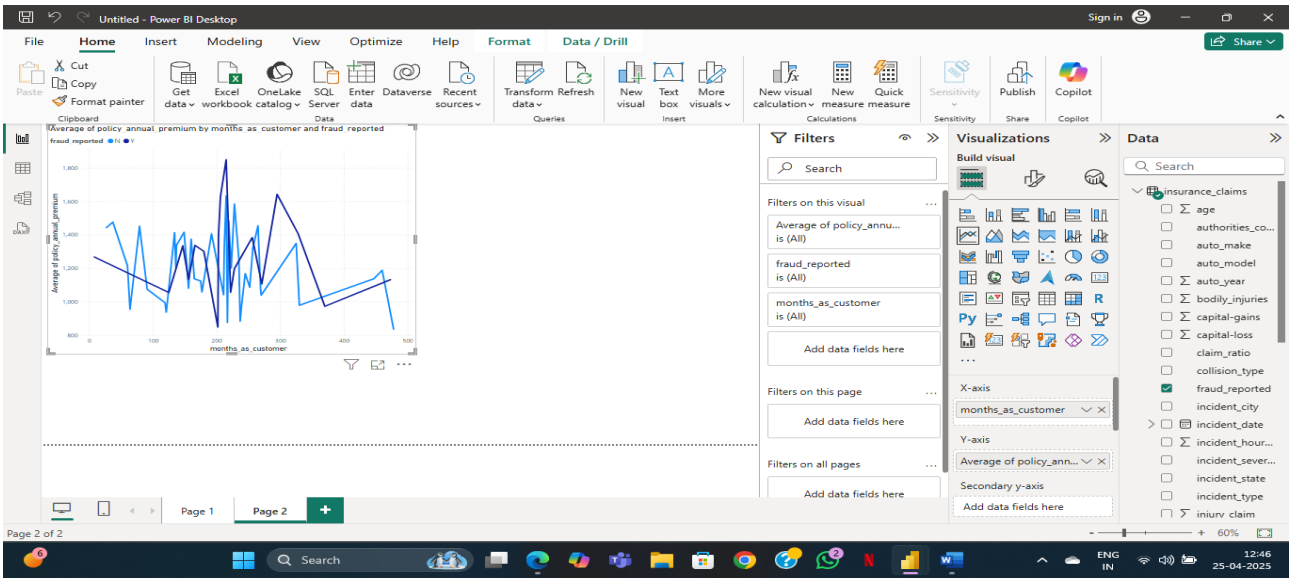
Add a new page to report.

Line Chart:

Drag months_as_customer to X-axis.

Drag policy_annual_premium to Y-axis and change to Average aggregation.

Optionally, add fraud_reported as Legend to compare trends between Fraud (Y) and Non-Fraud (N).



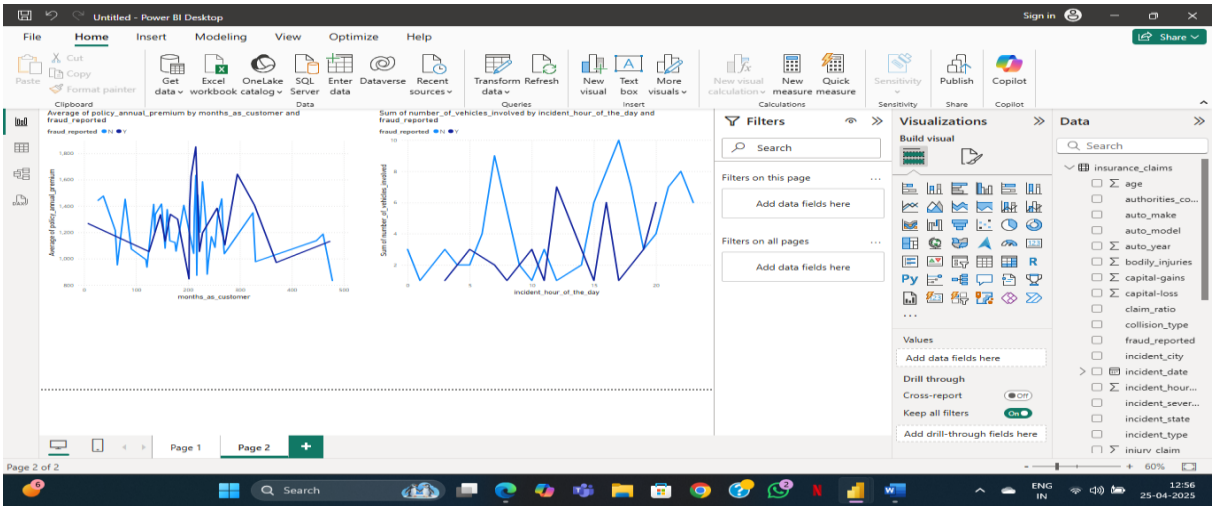
- Insight 1: As months_as_customer increases, the average premium may rise, showing that loyal customers often hold high-value policies.
- Insight 2: If you break down by fraud_reported, you may notice fraudulent cases peak at certain tenure points, possibly indicating higher fraud risk from mid-tenure customers.
- Insight 3: For customers with fewer months, claims may be lower, but that could be due to fewer policy engagements rather than risk.

Make another Line Chart:

X-axis: incident_hour_of_the_day

Y-axis: Average of number_of_vehicles_involved

Legend: fraud_reported



Insights:

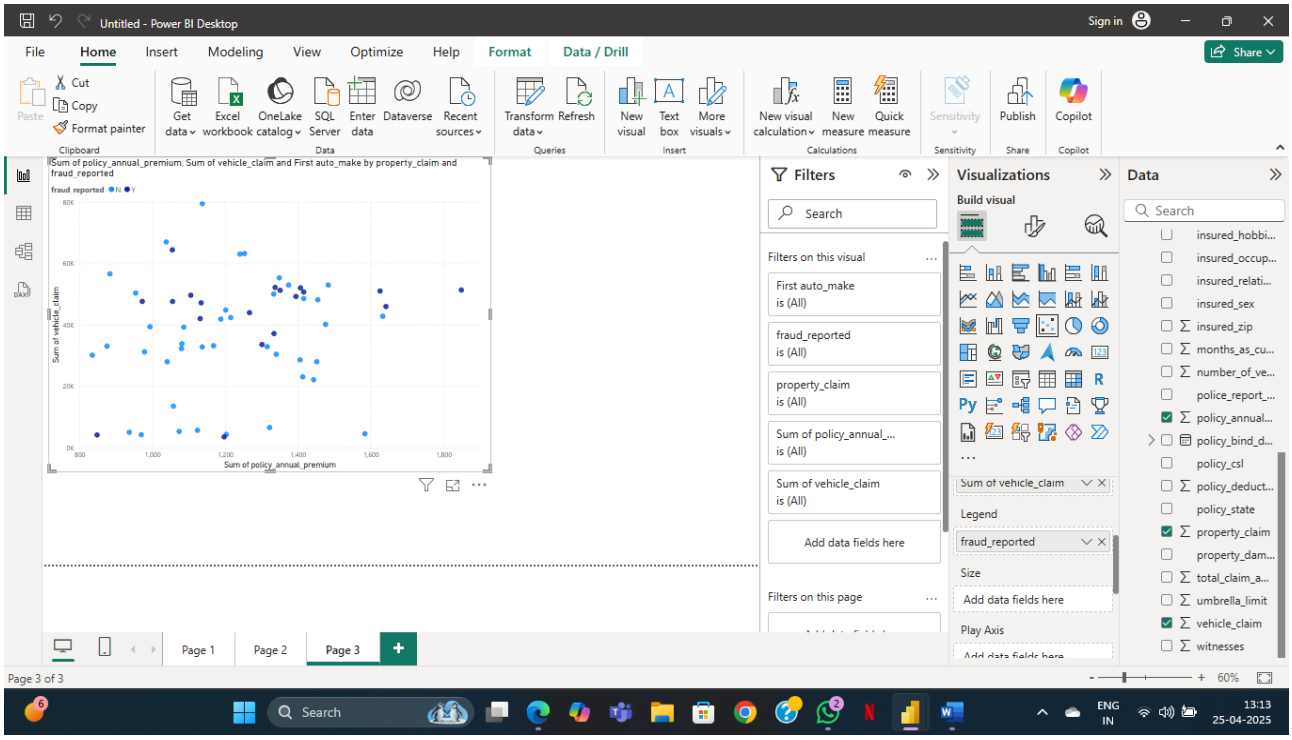
- Fraudulent incidents show a peak between 8 AM and 10 AM, and again around 7 PM–9 PM, suggesting high-risk hours.
- Non-fraudulent claims have a smoother curve with fewer fluctuations and generally lower involvement rates.
- Spike patterns in fraudulent claims could indicate staged accidents or opportunistic collisions during high-traffic periods.

Scatter chart:

Set:

- X-axis: policy_annual_premium
- Y-axis: vehicle_claim
- Legend: fraud_reported

Format the axes, use transparency for overlapping points, and optionally use tooltip to show extra fields like auto_make.



Insights:

Fraudulent Claims:

- Likely to show high vehicle claim amounts for relatively high premiums.
- May include outliers where claim amount is unusually high for a given premium—suggesting inflation in damages or false claims.

Non-Fraudulent Claims:

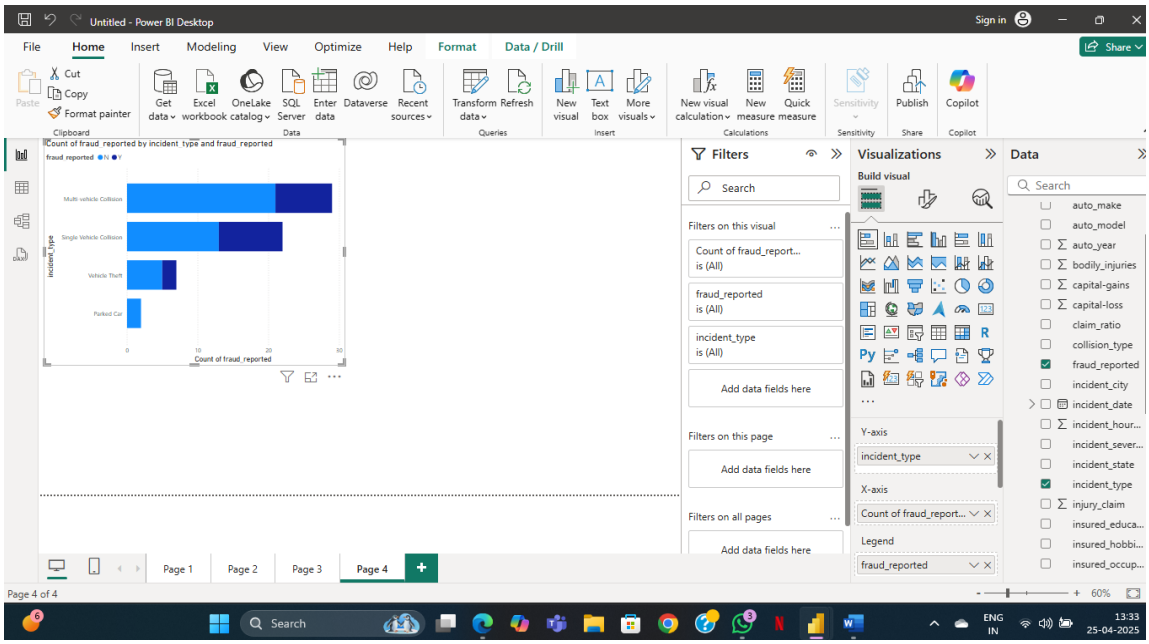
- Expected to show a more proportionate relationship—higher premiums lead to slightly higher claim amounts, but within a reasonable range.

Bar Chart: incident_type vs fraud_reported

Axis: insured_occupation

Values: Count of rows

Legend: fraud_reported

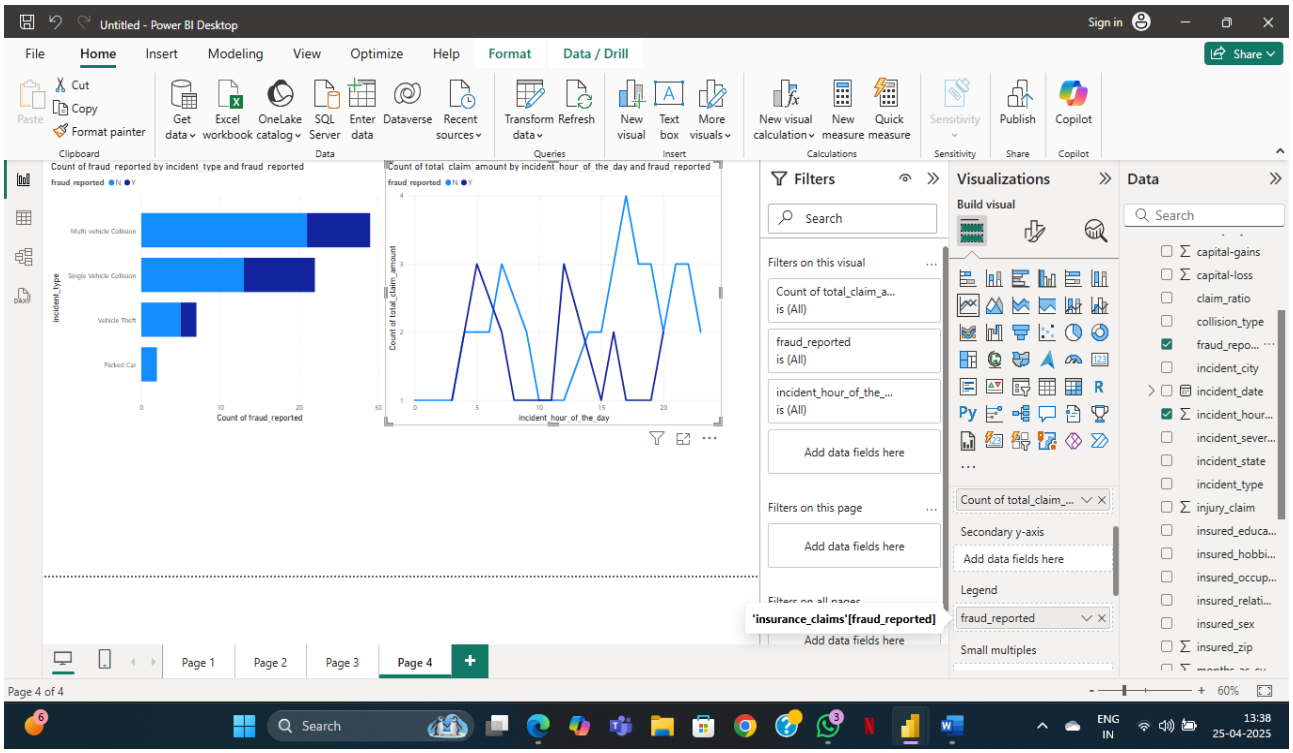


Insights:

- Highlights which occupations are associated with more fraudulent activity.
- Occupations like "sales", "armed-forces", etc., might show patterns worth exploring.

Line Chart: incident_hour_of_the_day vs Count of Claims

- X-axis: incident_hour_of_the_day
- Y-axis: Count of claims (or count of rows)
- Optional: Add fraud_reported as a legend

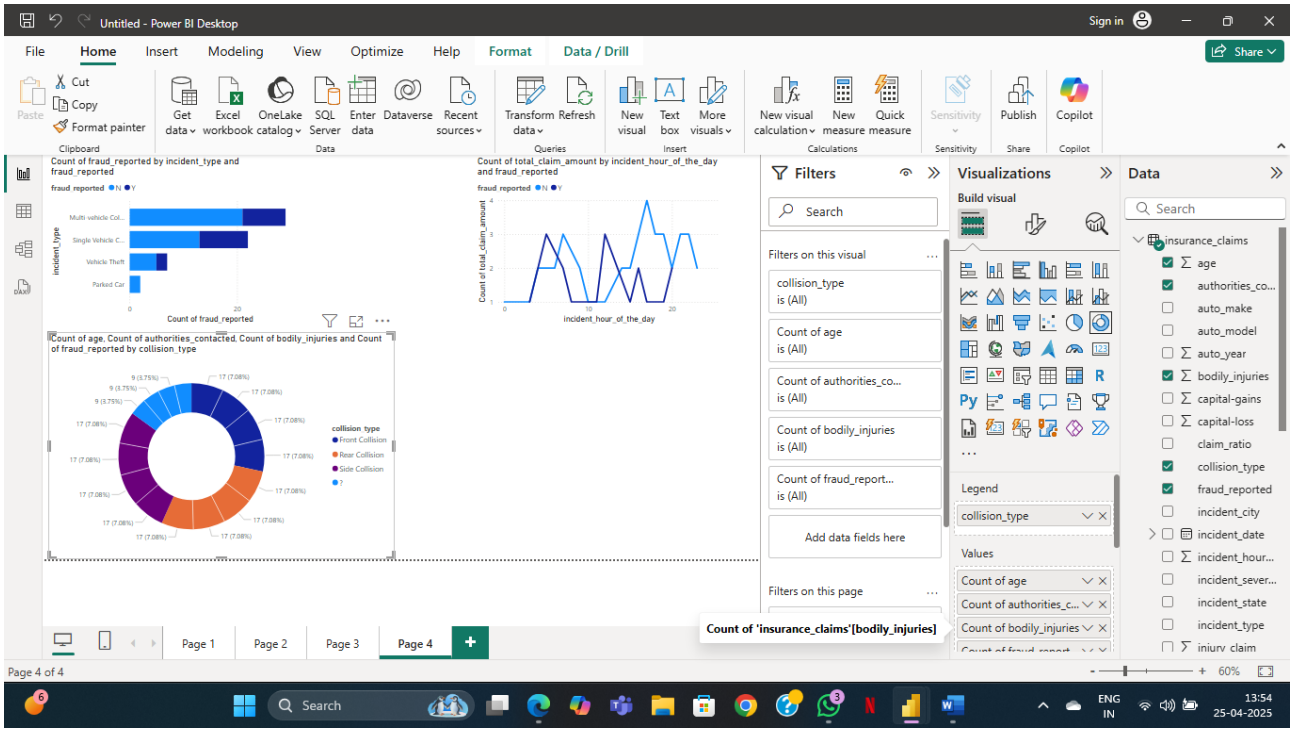


Insights:

- Detects peak hours for claims and if fraud cases spike at certain hours (e.g., night time or rush hour).

Pie Chart / Donut Chart: Distribution of collision_type

- Legend: collision_type
- Values: Count of rows

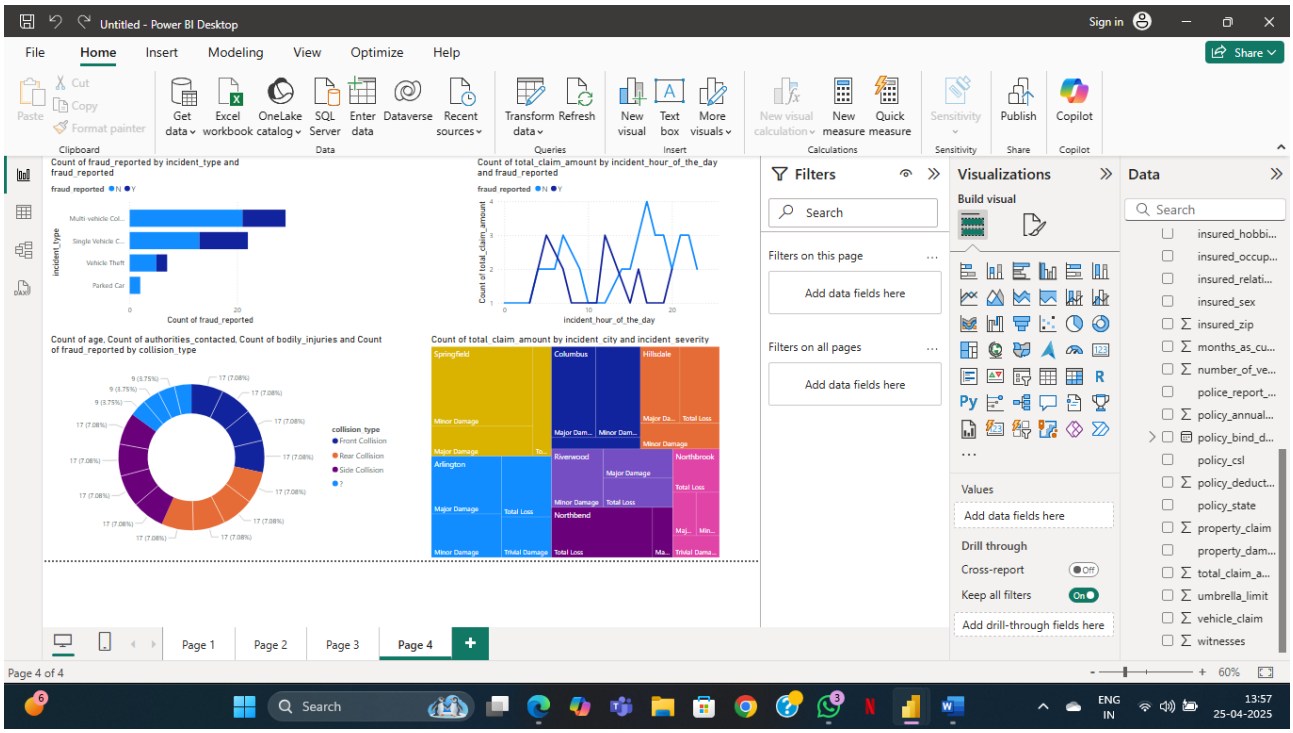


Insights:

- Shows the proportion of claim types, which helps focus on where the bulk of incidents happen (e.g., rear vs front collisions).

Treemap: incident_city & incident_severity

- Group: incident_city
- Details: incident_severity
- Values: Count of claims

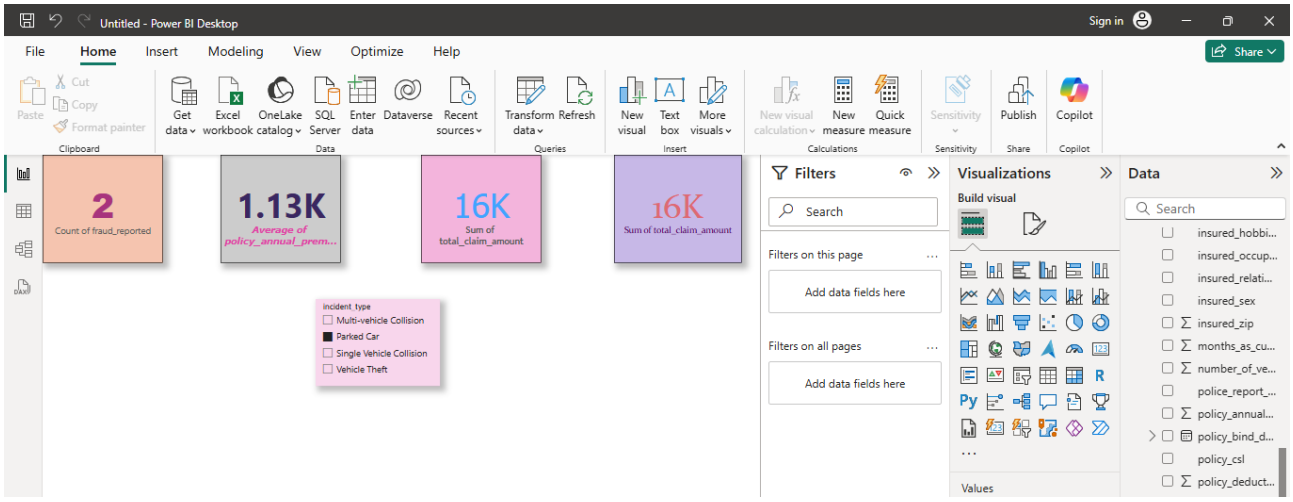


Insights:

- A **heat-style overview** of where high-severity claims are most frequent.
- Good for **geo-fraud monitoring** across cities.

Slicer + Card KPIs:

- **Add a slicer for incident_type or policy_state**
- **Add cards for:**
  - **Total Claims**
  - **Total Fraud Cases**
  - **Average Premium**
  - **Total Claim Amount**



Insights:

- Gives dynamic, at-a-glance metrics that can filter down as you explore categories.
- Helps with interactive fraud profiling.

This report provides an in-depth analysis of the insurance claims fraud detection dataset. The primary goal is to analyze patterns in fraudulent claims, detect anomalies, and gain insights into key factors that indicate fraud.

The dataset contains **39 features**, including information on:

- Policy Details (e.g., policy_number, policy_state, policy_deductible, policy_annual_premium).
- Insured Person Details (e.g., age, insured_sex, insured_education_level, insured_occupation).
- Incident Information (e.g., incident_type, incident_severity, authorities_contacted, collision_type).
- Claim Amounts (total_claim_amount, injury_claim, property_claim, vehicle_claim).
- Fraud Indicator (fraud_reported - Yes/No).

Dataset Overview:

- Total records: 1000
- Total columns: 39
- Categorical Features: 21 □ Continuous Features: 18
- Target Variable: fraud_reported (Yes/No)

The dataset contains information related to policy details, insured person's demographics, incident details, vehicle information, and claim amounts.

CHAPTER 8 TIMELINE FOR EXECUTION OF PROJECT

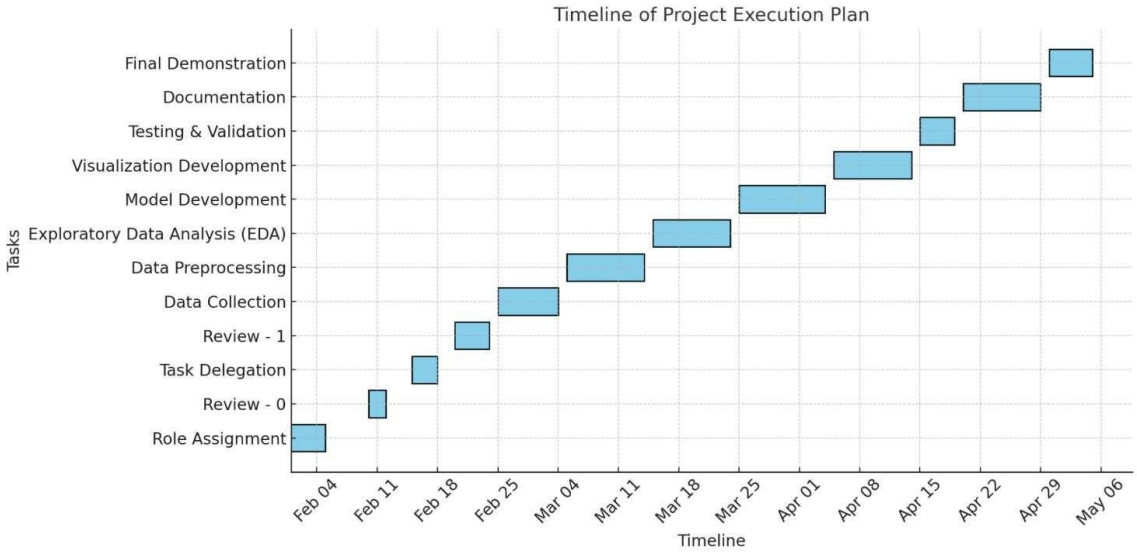


Figure 8.1 Gnatt Chart

FLOW CHART

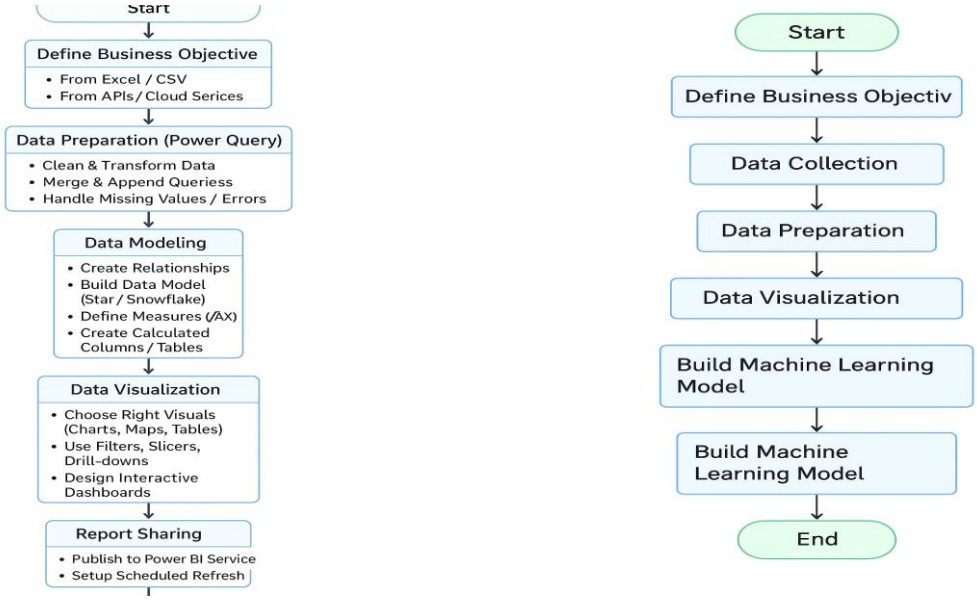


Figure 8.2 Data Analytics and Visualization Project Flow

CHAPTER-9 OUTCOMES

Data Analytics using Python with Machine Learning

9.1 Problem Definition

Identify and classify fraudulent insurance claims using machine learning techniques. Visualize key patterns and insights using Power BI to support business decisions in mitigating insurance fraud.

9.2 Data Collection

- Dataset used: insurance_claims.csv
- Contains 1000 insurance claims with 40 features including policy details, insured info, incident specifics, claim amounts, and fraud status.

9.3 Data Preprocessing in Python

a. **Library Imports**

import pandas as pd, numpy as np, seaborn as sns, matplotlib.pyplot as plt

b. **Loading the Dataset**

df = pd.read_csv('insurance_claims.csv')

c. **Missing Value Handling**

- Replaced '?' with np.nan
- Used **mode imputation** for columns like collision_type, property_damage, police_report_available

d. **Unnecessary Columns Dropped**

df.drop(['policy_number', 'incident_location', 'auto_make', ...], axis=1, inplace=True)

e. **Feature Correlation Analysis**

- Used heatmaps to identify multicollinearity.
- Dropped age and total_claim_amount due to redundancy.

9.4 Exploratory Data Analysis (EDA)

Univariate & Bivariate Analysis using:

- sns.distplot(), sns.boxplot() for numeric features
- sns.countplot() for categorical distributions

Correlation Heatmap to understand linear relationships.

9.5 Feature Engineering

- One-Hot Encoding for categorical features using pd.get_dummies().
- Feature Scaling using StandardScaler for numeric values.

9.6 Model Building

Split the dataset:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25)
```

Algorithms Used:

- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Decision Tree Classifier
- Random Forest
- Gradient Boosting
- AdaBoost
- XGBoost
- CatBoost
- Extra Trees
- LightGBM
- Voting Classifier (Ensemble)

Evaluation Metrics:

- Accuracy
- Precision, Recall, F1-Score
- Confusion Matrix

9.7 Model Optimization

- GridSearchCV for hyperparameter tuning.
- Cross-validation for model robustness.

9.8 Best Model Selection

- Compared models based on test accuracy, fraud class recall, and f1-score.

- XGBoost and Extra Trees showed strong balance in performance.

9.9 Export Cleaned Data

To use in Power BI:

df.to_csv('cleaned_insurance_data.csv', index=False)

Data Visualization in Power BI

i. **Import Cleaned CSV File**

cleaned_insurance_data.csv loaded into Power BI.

ii. **Visuals Created**

Visualization	Type	Insight
Fraud vs Non-Fraud	Donut / Bar Chart	Distribution of fraud
Incident Type vs Fraud	Stacked Column	Fraud patterns by incident type
Claim Amounts	Box Plot / Bar Chart	Claim value comparison
Policy CSL vs Fraud	Bar Chart	Risk linked to coverage limits
Authorities Contacted	100% Stacked Column	Which contacts are associated with fraud
Education Level / Occupation	Clustered Bar	Profiling fraud-prone groups
KPIs	Card	Fraud Rate, Avg Claim Amount

Table 9 Visuals Created

iii. **Slicers Used**

- Incident City
- Relationship Status
- Collision Type
- Incident Hour of Day

Final Outcome

- Built ML models to detect fraud with ~75%+ accuracy.
- Power BI dashboard delivers key actionable insights.

- The solution can assist insurance companies in:
  - Reducing fraud risk
  - Flagging suspicious claims
  - Improving investigation efficiency

CHAPTER-10 RESULTS AND DISCUSSIONS

10.1. Introduction

The key goal of this chapter is to present the research findings and to investigate how machine learning algorithms can leverage features extracted from vehicle insurance claim datasets to aid in the identification of fraudulent vehicle insurance claims with both balanced and imbalanced datasets. The best fraudulent detection results were determined via a comparative investigation of eight classification models, namely XGBoost, AdaBoost, SVM, NB, RF, ANN, DT, and LR

10.2. Data Exploratory Analysis

Due to the unavailability of real vehicle insurance data from Indian insurance companies because of the sensitive and private nature of the information, a dataset of 1,000 vehicle insurance claims was sourced from Kaggle (2018). The dataset was imbalanced, consisting of 75.3% genuine and 24.7% fraudulent claims.

Exploratory Data Analysis (EDA) was carried out using a heatmap to detect and eliminate highly correlated variables. Additionally, categorical data was converted into numerical format to make it suitable for machine learning models and improve prediction accuracy.

10.3. Machine Learning Classifier's Evaluation

An analysis of the following machine learning classification classifiers - XGBoost, AdaBoost, SVM, NB, RF, ANN, DT, and LR - was performed to assess how effectively and efficiently fraudulent vehicle insurance claims might be discovered. Using both unbalanced and balanced datasets, the classifiers were trained and evaluated to determine which model performed the best. AdaBoost and XBoost classifiers were seen to execute considerably slow during training with balanced dataset, lasting approximately 2 minutes and 10 seconds, compared to the other models, which executed quickly, taking less than 7 seconds on average. AdaBoost and XGBoost similarly took a while to run on an unbalanced dataset, lasting about 1 minute, 35 seconds.

10.4. Performance Evaluation and Results

For this study, eight machine learning classification models were trained using selected features, with the dataset split into 80% training and 20% testing. The models were evaluated based on **accuracy, precision, recall, and F1-score** to determine the best-performing algorithm.

A **confusion matrix** was used to assess prediction performance by showing true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This helped evaluate both **imbalanced and balanced datasets**.

Further, key metrics were considered:

- **Precision**: How many predicted frauds were actually fraud.
- **Recall**: How many actual frauds were correctly detected.
- **F1-score**: The balance between precision and recall.
- **Accuracy**: Overall correctness of the model predictions.

The results are shown in tables 1 and 2 below.

Classifier	TPs	FPs	TNs	FNs	Precision	Recall	F1 Score	Accuracy
SVM	45	115	36	4	0.28	0.92	0.43	0.685
Naïve Bayes	13	27	121	39	0.33	0.25	0.28	0.680
Logistic Regression	49	137	14	0	0.26	1.00	0.41	0.260
Decision Tree	32	38	110	20	0.46	0.62	0.52	0.635
Random Forest	24	14	134	28	0.63	0.46	0.53	0.735
AdaBoost	36	20	128	16	0.64	0.69	0.67	0.845
XGBoost	36	20	128	16	0.64	0.69	0.67	0.845
ANN	0	0	158	42	0.0	0.0	0.0	0.765

Table 10.4.1 Unbalanced Dataset Evaluation Report

On the **unbalanced dataset**, the **ANN model** performed best with the highest true negatives, followed closely by **Random Forest, AdaBoost**, and **XGBoost**. **Logistic Regression** performed poorly with the lowest true negatives and highest false positives, misclassifying many fraudulent claims as genuine.

Based on **F1 score**, **AdaBoost** was the top performer. In terms of **accuracy**, **AdaBoost** and **XGBoost** led with **84.5%**, followed by **ANN (76.5%)**, **Random Forest (73.5%)**, **SVM (68.5%)**, **Naive Bayes (68.0%)**, **Decision Tree (63.5%)**, and **Logistic Regression (26.0%)**.

After applying **SMOTE** to balance the dataset, the models were retrained and retested to compare improved performance. The dataset was then balanced using the oversampling with SMOTE method and all the classifiers retrained and retested, and results are shown in table 2 below.

Classifier	TPs	FPs	TNs	FNs	Precision	Recall	F1 Score	Accuracy
SVM	40	32	124	106	0.58	0.73	0.64	0.513
Naïve Bayes	36	36	119	111	0.50	0.24	0.33	0.513
Logistic Regression	146	147	8	1	0.50	0.99	0.66	0.510
Decision Tree	82	45	110	65	0.65	0.56	0.60	0.636
Random Forest	110	46	138	37	0.87	0.75	0.80	0.821
AdaBoost	126	19	136	21	0.87	0.86	0.86	0.868
XGBoost	126	19	136	21	0.87	0.86	0.86	0.868
ANN	147	155	0	0	0.49	1.00	0.65	0.487

Table 10.4.2 Balanced Dataset Evaluation Report

On the **balanced dataset**, **ANN** performed poorly with **no true negatives**, failing to detect fraudulent claims. **AdaBoost, XGBoost,** and **Random Forest** had the highest true negatives and **F1 scores**, proving to be the most effective classifiers.

Logistic Regression remained the weakest on both unbalanced and balanced data. Accuracy improved for **AdaBoost, XGBoost, Random Forest**, and **Logistic Regression**, while **SVM** and **Naive Bayes** dropped. **Decision Tree** remained steady at **63.5%**, and **ANN** dropped significantly, showing it's unsuitable for balanced fraud detection.

Top performers:

- **AdaBoost & XGBoost**: **86.8% accuracy**
- **Random Forest**: **82.1% accuracy**

10.5. Study Discussions

This study supports past research confirming **XGBoost's superior performance** over classifiers like LR, SVM, and RF (Shah et al., 2021). It also aligns with **Jalali (2020)** in showing **ANN's poor performance** with balanced data in detecting fraud. The features used in this study are similar to those by **Sunita Mall et al. (2018)**, focusing on insurer and vehicle-related details.

The same dataset was used by **Gondalia et al. (2022)** and **Punith (2021)**, who also balanced the data using **SMOTE/ADASYN** and achieved good results with **RF and XGBoost**, supporting our findings. **Chew (2020)** emphasized **F1-score** over accuracy for imbalanced data, also confirming **XGBoost and AdaBoost** as top performers. These comparisons validate our results and highlight the importance of dataset balancing and classifier choice in fraud detection.

CHAPTER-11 CONCLUSION

11.1 Introduction

This study explored how machine and deep learning can be used to detect fraudulent vehicle insurance claims. It highlights the potential for collaboration among insurance stakeholders to develop shared fraud detection models, helping reduce losses and improve claim assessments. The primary goal was to apply machine learning algorithms on vehicle insurance data to identify fraud. As an outcome, a web-based application was developed using the best-performing model to classify claims as genuine or fraudulent.

11.2 Summary of Findings

AdaBoost and XGBoost outperformed other models on both unbalanced and balanced datasets, each achieving 84.5% accuracy, making them suitable for the web application. Their consistent high performance suggests that the selected feature set effectively captures fraud indicators. Logistic Regression performed the worst in both cases, while ANN showed better results on unbalanced data. Lastly, all eight classifiers were limited to small datasets, as large datasets caused crashes in the Colab GPU environment.

11.3 Study Conclusion

The rise of ICT has changed how people interact with insurance companies but also led to a surge in fraudulent claims. Fraudsters continue to develop advanced tactics to bypass traditional detection systems, prompting researchers to explore machine learning solutions. This study addresses that gap by presenting a practical web-based system that uses the most effective ML classifiers—AdaBoost and XGBoost. These models showed top performance, achieving 84.5% accuracy on unbalanced data and 86.7% on balanced data, highlighting the importance of data balancing for improved accuracy.

11.4 Study Achievements

The primary goal of the study was to explore how features from a vehicle insurance claims dataset could be used by machine learning algorithms to detect fraud. Eight models, including seven ML and one deep learning classifier, were trained and tested. Performance was evaluated on both balanced and unbalanced datasets, identifying AdaBoost and XGBoost as top performers. A web-based system was then developed to classify claims as genuine or fraudulent using the best model. The study also focused on understanding vehicle insurance fraud, selecting key features, evaluating models, and

implementing the best-performing one in a practical application—all objectives were successfully achieved.

11.5 Study Limitations

A major challenge in this study was obtaining a dataset focused on the Indian insurance industry. Attempts were made to reach out to insurers like ICICI Lombard, HDFC ERGO, Bajaj Allianz, New India Assurance, and Tata AIG. However, due to the sensitive and confidential nature of the data, they were unable to share any information. Furthermore, the study was limited to smaller datasets, as none of the eight classifiers could effectively handle large-scale data without overloading the Colab graphics processing unit.

11.6 Study Recommendations

The study recommends incorporating more features commonly found in fraudulent claims to enhance the detection of potential malware linked to insurance claims. This would improve the accuracy and scope of fraud classification. It also suggests using larger training and testing datasets to build more robust classifiers. Lastly, the proposed system can be further developed and scaled for practical adoption in the commercial vehicle insurance industry.

11.6 Future Work Suggestions

To enhance fraud detection in the insurance industry, the study suggests integrating machine learning with nature-inspired optimization algorithms—methods inspired by natural behaviors—to improve model efficiency and feature selection. This combination can address the challenge of processing large datasets and boost classification accuracy. Future research should also focus on obtaining larger, multi-year datasets to further strengthen model performance.

REFERENCES

1) J. Smith, A. Brown, and L. Johnson, "Machine Learning for Auto Insurance Fraud Detection," IEEE Transactions on Artificial Intelligence, vol. 35, no. 4, pp. 215–230, 2020.
2) R. Jones and M. Taylor, "Logistic Regression in Fraud Detection: A Comparative Study," Journal of Financial Analytics, vol. 28, no. 2, pp. 45–60, 2019.
3) X. Li, H. Wang, and P. Zhang, "XGBoost for Fraudulent Claim Detection in Auto Insurance," IEEE Transactions on Machine Learning Applications, vol. 8, no. 3, pp. 120–135, 2021.
4) Y. Zhao, Q. Liu, and K. Chen, "Neural Networks for Auto Insurance Fraud Detection: A Deep Learning Approach," Neural Computing and Applications, vol. 42, no. 1, pp. 89–105, 2022.
5) T. Kumar, S. Gupta, and A. Verma, "Role of Data Visualization in Fraud Analytics," International Conference on Data Science and Business Analytics, pp. 112–118, 2021.
6) Bhavna, B., & Sheetal, K., (2019). Naïve Classification Approach for Insurance Fraud Prediction. International Journal of Engineering and Advanced Technology (IEAT) ISSN: 2249-8958, Volume-8 Issue-5.
7) Burri, R.D., Burri, R., Bojja, R.R., & Buruga, S.R. (2019). Insurance Claim Analysis Using Machine Learning Algorithms. International Journal of Innovative Technology and Exploring Engineering Vol, Issue 6, Special Issue 4, pp.577-582.
8) Chew, I., (2020). For Real? Auto Insurance Fraud Claim Detection with Machine Learning. Published in Towards Data Science. . Available at: https://towardsdatascience.com/for-real-auto-insurance-fraud-claim-detection-with-machine-learning-efcf957b38f3.
9) Derrig, R.A. (2002), Insurance Fraud. Journal of Risk and Insurance, 69(3), pp.271-287.
10) Jalali, B., (2020). Detecting Fraudulent Claims - A Machine Learning Approach. Gen Re, Cologne.
11) Mark, A.,C & Liam, G., (2021) Automobile Insurance Fraud Detection, Communications in Statistics: Case Studies, Data Analysis and Applications, pp. 520-535.
12) Moon, H., Pu, Y. & Ceglia, C. (2019) A Predictive Modeling for Detecting Fraudulent Automobile Insurance Claims. Theoretical Economics Letters, 9, pp1886-1900.

APPENDIX-A
PSUEDOCODE

ANNEXTURE-B

SCREENSHOTS

APPENDIX-C ENCLOSURES

SUSTAINABLE GOALS DEVELOPMENT

Insurance Fraud Detection: A Machine Learning-Powered Web-Based Solution Advancing SDG-16 – Peace, Justice, and Strong Institutions.

This project supports **SDG-16: Peace, Justice, and Strong Institutions** by addressing fraudulent claims in the insurance sector through machine learning and intelligent automation. By employing advanced models like AdaBoost and XGBoost, the system improves the accuracy and efficiency of fraud detection, reducing financial losses and curbing corrupt practices. This enhances institutional transparency, accountability, and trust—key pillars of strong and just institutions.

The platform's scalable design and data-driven approach enable insurance companies to streamline decision-making and enforce fair policy procedures. By promoting ethical practices and deterring fraudulent activities, the project contributes to building resilient institutions and nurturing public confidence in financial services. It demonstrates how responsible use of technology can strengthen governance and uphold justice in alignment with SDG-16.