

Insurance Claim Analysis – Summary Report

1. Objective

This project analyzes insurance claim data to understand key risk factors and build predictive machine learning models for claim occurrence. The process includes EDA, preprocessing, model training, and evaluation.

2. Dataset Overview

The dataset consists of customer demographics, policy details, and historical claim records. The target variable indicates whether a claim was made, with noticeable class imbalance.

3. Exploratory Data Analysis

- 1 Explored distributions of numerical and categorical features.
- 2 Identified important factors influencing claim probability.
- 3 Detected skewness, outliers, and variable relationships.

4. Data Preprocessing

- 1 Handled missing and inconsistent data.
- 2 Encoded categorical variables and scaled numerical features.
- 3 Performed train-test split for model validation.

5. Model Building

- 1 Logistic Regression used as a baseline model.
- 2 Tree-based models captured non-linear patterns.
- 3 Random Forest provided the best overall performance.

6. Model Evaluation

Models were evaluated using accuracy, precision, recall, F1-score, and confusion matrices. Comparative analysis helped identify the most reliable model for claim prediction.

7. Conclusion

The project demonstrates that data-driven models can effectively support insurance risk assessment and decision-making when combined with domain knowledge.