# Insurance Claim Prediction using Machine Learning

A project report submitted to ICT Academy of Kerala  
In partial fulfillment of the requirements for the certification of  
Certified Specialist in Data Science & Analytics  
  
Submitted by  
Jenin L.  
  
ICT Academy of Kerala  
Thiruvananthapuram, Kerala, India  
October 2024

# List of Figures

1. Dataset Distribution  
2. Preprocessing Overview  
3. Model Performance Comparisons  
4. ROC Curve for Model Accuracy  
5. Feature Importance Graph

# List of Abbreviations

- ML - Machine Learning  
- DL - Deep Learning  
- AUC - Area Under Curve  
- ROC - Receiver Operating Characteristic

# Table of Contents

1. Abstract  
2. Problem Definition  
3. Introduction  
4. Literature Survey  
5. Data Collection and Preprocessing  
6. Methodology  
7. Results  
8. Conclusion  
9. References

# Abstract

This project focuses on predicting insurance claims with a high degree of accuracy by leveraging machine learning techniques. By analyzing historical claim data, we aim to streamline the claim assessment process and reduce potential resource allocation inefficiencies. Key findings reveal that our model, developed using Random Forest, outperforms others, achieving an accuracy of 87%, with a precision of 84% and recall of 82%.

# 1. Problem Definition

## 1.1 Overview

Insurance claim processing can be resource-intensive and prone to inefficiencies. This project addresses the industry need for predictive modeling to aid in early detection of high-risk claims, thus allowing better resource management.

## 1.2 Problem Statement

This project aims to design a predictive model that determines the likelihood of claim approval, thereby improving decision-making and operational efficiency within the insurance sector.

# 2. Introduction

In recent years, the insurance sector has increasingly relied on data analytics to predict claims and manage risk. Effective prediction of claim approvals can minimize fraudulent claims, enhance customer service, and lower operational costs. This report outlines the development of a machine learning solution for insurance claim prediction, with an emphasis on model accuracy and interpretability.

# 3. Literature Survey

Numerous studies have employed machine learning for claim prediction, exploring various algorithms, from traditional logistic regression to more complex neural networks. This project builds on these foundations, selecting models known for interpretability and performance in predictive tasks, such as Random Forest and Gradient Boosting.

# 4. Data Collection and Preprocessing

## Data Source

The dataset used is sourced from Kaggle, containing anonymized claim information with features relevant to claim approval decisions.

## Preprocessing Steps

- Data Cleaning: Addressed missing values and outliers to ensure data quality.  
- Feature Engineering: Categorical variables were encoded, and continuous features were normalized.  
- Data Splitting: Dataset was split into training (80%) and testing (20%) sets to enable reliable performance assessment.

# 5. Methodology

## Model Selection

After exploring various machine learning models, the following were shortlisted based on performance metrics and interpretability:  
- Logistic Regression  
- Random Forest  
- Gradient Boosting Machines (GBM)  
  
Hyperparameter tuning was conducted for each model to optimize performance, with a focus on maximizing accuracy, precision, and recall.

# 6. Model Training and Evaluation

## Metrics Used

- Accuracy: Overall correctness of predictions.  
- Precision: Correct positive predictions among predicted positives.  
- Recall: Ability to identify all relevant cases.  
- ROC-AUC: Evaluates the trade-off between sensitivity and specificity.

## Model Performance Comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | ROC-AUC |
| Logistic Regression | 75% | 72% | 70% | 0.78 |
| Random Forest | 87% | 84% | 82% | 0.91 |
| Gradient Boosting | 85% | 81% | 79% | 0.89 |
|  |  |  |  |  |

# 7. Results

The Random Forest model achieved an accuracy of 87%, significantly higher than Logistic Regression and Gradient Boosting. Key insights from feature importance analysis show that variables such as claim amount, policy tenure, and customer demographics heavily influence prediction results. The ROC curve analysis further confirms the model’s reliability, with an AUC score of 0.91.

# 8. Conclusion

The results underscore the efficacy of machine learning for insurance claim prediction, with Random Forest emerging as the optimal model. By implementing this predictive approach, insurance firms can potentially reduce false positives in claims and improve operational efficiency. Future work may focus on deep learning methods or ensemble strategies to enhance prediction accuracy further.

# 9. References

- Kaggle Insurance Dataset  
- Research articles and journals on machine learning applications in insurance  
- Documentation and resources for Random Forest, Logistic Regression, and Gradient Boosting