Project Overview

Project Title: Comparative Analysis of RFE and LASSO in Predicting Insurance Fraud Using Machine Learning

Summary of the Background: The insurance business faces a rising issue with insurance fraud which generates annual financial losses that continue to grow. Intentional fraud takes the form of insured parties staging accidents along with exaggerating losses or fraud involves misrepresents facts to gain money from insurance payments. The detection of insurance fraud becomes complex because large volumes of data differences exist between genuine and fraudulent claims (Villegas-Ortega, Bellido-Boza, and Mauricio, 2021). Machine learning models serve as efficient fraud detection systems through their ability to learn previous data to find patterns. The selection process of features remains vital for improving detection models because it chooses optimal predictors along with eliminating unnecessary variables (Zhang *et al.*, 2023). The feature selection methods being widely utilized in research include Recursive Feature Elimination (RFE) together with Least Absolute Shrinkage and Selection Operator (LASSO). Feature selection techniques RFE examines and deletes unimportant variables one by one to improve model prediction while LASSO applies regularization that reduces coefficients of unimportant features to zero to enhance model clarity. This research evaluates RFE as well as LASSO in assessing important features needed for insurance fraud prevention and when used along with ML methods like XGBoost, Random Forest as well as Logistic Regression models to optimize the predictive accuracy.

Research Question

- ➢ Which of the feature selection techniques (RFE and LASSO) produces better results in identifying predictive features for insurance fraud?

Project Objectives

- To review existing research in feature selection along with machine learning models with focus on detecting insurance fraud.
- To collect and preprocess the insurance fraud dataset from Kaggle by removing missing values and outliers.
- To apply RFE and LASSO to identify key features from the dataset.
- To implement ML models (XGBoost, Random Forest, Logistic Regression) using the features selected through RFE and LASSO.
- To evaluate the performance of each model using metrics such as accuracy, mean square error (MSE), as well as F1 score.
- To compare the efficiency of RFE and LASSO in improving model performance for predicting insurance fraud.

Reference List

Zhang, B., Dong, X., Hu, Y., Jiang, X. and Li, G., 2023. Classification and prediction of spinal disease based on the SMOTE-RFE-XGBoost model. PeerJ Computer Science, 9, p.e1280.

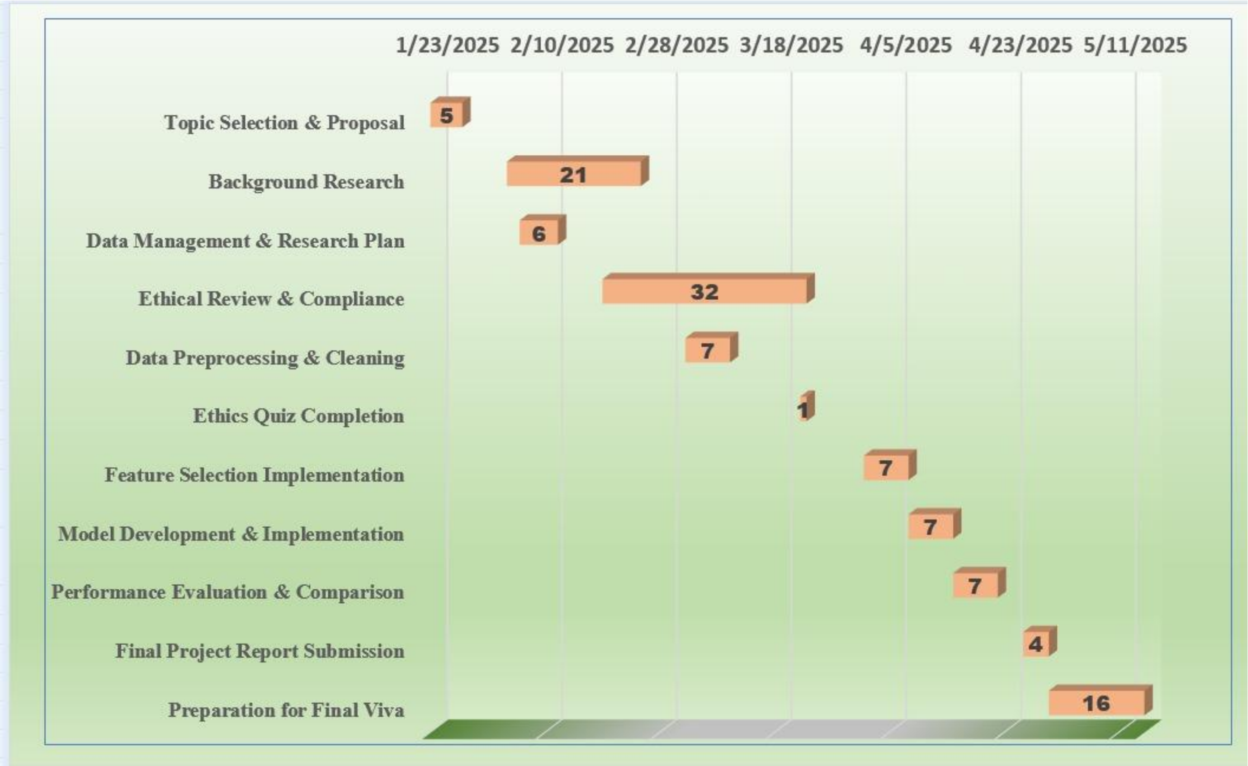
Villegas-Ortega, J., Bellido-Boza, L. and Mauricio, D., 2021. Fourteen years of manifestations and factors of health insurance fraud, 2006–2020: a scoping review. Health & justice, 9, pp.1-23.

Project Plan:

Task list

Task	Description	Start Date	End Date
Topic Selection & Proposal	Select the research topic, finalize dataset, submit the project selection form.	23/01/2025	27/01/2025

Supervision Meeting 1	Discuss research objectives, methodology.	28/01/2025	03/02/2025
Background Research	Review existing studies on fraud detection.	04/02/2025	24/02/2025
Data Management & Research Plan	Develop the Data Management Plan and outline research workflow.	06/02/2025	11/02/2025
Supervision Meeting 2	Present the data management plan and receive feedback.	12/02/2025	18/02/2025
Ethical Review & Compliance	Study ethical guidelines and prepare for the Ethics Quiz.	19/02/2025	22/03/2025
Supervision Meeting 3	Submit and discuss the draft literature review.	25/02/2025	03/03/2025
Data Preprocessing & Cleaning	Handle missing data, encode categorical features.	04/03/2025	10/03/2025
Supervision Meeting 4	Present methodology, focusing on feature selection techniques.	11/03/2025	17/03/2025
Ethics Quiz Completion	Complete and submit the required ethics assessment.	22/03/2025	22/03/2025
Supervision Meeting 5	Review progress and initial modeling steps.	25/03/2025	31/03/2025
Feature Selection Implementation	Apply RFE and LASSO for selecting key features.	01/04/2025	07/04/2025
Model Development & Implementation	Train ML models (XGBoost, RF, LR).	08/04/2025	14/04/2025
Performance Evaluation & Comparison	Assess model performance, compare RFE and LASSO.	15/04/2025	21/04/2025
Supervision Meeting 6	Submit a draft of the FPR and discuss findings.	22/04/2025	25/04/2025
Final Project Report Submission	Submit the FPR.	26/04/2025	29/04/2025
Preparation for Final Viva	Prepare for Viva and attend the scheduled session	30/04/2025	13/05/2025



Data Management Plan

Overview of the Dataset: Insurance claim records within the Insurance Fraud Detection dataset include explanatory data points that enable researchers to develop predictive models which identify fraudulent insurance claims. This dataset comprises demographic and insurance and fraud-related information.

Data Collection: The dataset can be acquired from Kaggle website (https://www.kaggle.com/datasets/arpan129/insurance-fraud-detection) after it was first compiled at Indian Institute of Management Calcutta.

Metadata: The dataset consists of 38 explanatory variables together with one target variable "Fraud Reported" that shows insurance claim authenticity as 1 for fraudulent cases and 0 for legitimate cases. The CSV file contains a limited amount of data while maintaining efficient processing speed.

Document Control: A GitHub repository will function as the essential code storage platform at (https://github.com/gopireddy999/Fraud-Insurance-Cliam) where proper documentation and version control systems are maintained. Commitments for code modifications will happen weekly and every change.

ReadMe File: A ReadMe file inside the repository will serve to present vital project information to users. The file establishes details about the dataset while demonstrating installation and operation steps and methods for using data for detecting fraud.

Security and Storage: The project team members together with authorized reviewers will have restricted GitHub access for uploading all data along with code storage in online cloud backup which will run weekly backups to avoid data loss.

Ethical Requirements: The dataset contains no personal information which ensures the safety of individual privacy during GDPR and all relevant data protection standards. The research faces no ethical issues because the available Kaggle dataset can be used by academics.