Module 1 Final Project: Dataset Proposal

**Dataset: Car Insurance Claim Prediction**

****

**Mohit Maithani**

Northeastern University: College of Professional Studies

ALY 6040: Data Mining

Professor Justin Grosz

February 23, 2025

**Dataset Description:** The Car Insurance Claim Prediction dataset is designed to facilitate the prediction of insurance claims within a six-month period based on various policyholder attributes. This dataset comprises extensive details about policyholders, including policy tenure, age of the car, age of the car owner, population density of the city, and technical specifics like make, model, power, and engine type of the car. The primary objective is to utilize this dataset to predict whether a policyholder will file a claim within the next six months, making it an ideal candidate for binary classification tasks.

**Rationale for Selection:** The choice of this dataset stems from its direct applicability to real-world scenarios in the insurance industry. The ability to predict claims accurately can significantly influence the decision-making process within insurance companies, enhancing their capacity to manage risks and allocate resources efficiently. Additionally, this dataset provides a clear binary target variable (is\_claim), which indicates whether a claim was filed, aligning perfectly with the predictive modeling focus of our data mining course.

**Initial Interest:** My initial interest in this dataset is driven by its complexity and relevance to the financial sector, particularly in risk management. The dataset’s attributes offer a deep dive into factors that might influence claim frequencies, providing a rich ground for applying various data mining techniques, including data preprocessing, exploratory data analysis, and predictive modeling.

**Learning Objectives:** Through the analysis of this dataset, I aim to:

* Develop a robust understanding of the relationships between different variables and their impact on insurance claims.
* Enhance my skills in handling imbalanced datasets, as claim incidents typically represent a smaller proportion of the total dataset.
* Apply and compare multiple machine learning algorithms to identify the most effective model for this prediction task.

**Group Members:**

* Ayush Anand
* Mohit Maithani

**Conclusion:** This dataset not only meets the academic and practical requirements of our data mining course but also challenges us to apply our theoretical knowledge in a context that mirrors real industry problems. By the end of this project, we expect to deliver actionable insights that could potentially benefit the insurance sector, particularly in enhancing their predictive analytics capabilities.

**Dataset Source:**

<https://www.kaggle.com/datasets/ifteshanajnin/carinsuranceclaimprediction-classification/data>