DSCI 510 Final Project Proposal

Project Title: Fraud Risk Analytics for Health Insurance Claims

Team Member: Shivam Kumar | **USC ID:** 9153656592

Email: shivamku@usc.edu | **GitHub** **https://github.com/shivam8764/DSCI-510-PROJECT**

1. Problem Statement and Business Proposition

Health insurance companies lose substantial amounts of money each year due to fraudulent claims, including inflated bills, unnecessary procedures, and falsified hospitalizations. These losses eventually lead to higher premiums for honest customers and reduced funds for genuine patients.

In this project, I will design a fraud risk analytics pipeline that learns patterns from claims data and assigns a fraud propensity score to each claim or provider. The business idea is straightforward: if suspicious items can be identified early, investigators can focus on the riskiest cases, reduce financial leakage, and help maintain fair premiums for everyone.

2. Data Sources and Collection Strategy

To meet the course requirement, I will combine one realistic fraud dataset with external health and demographic data collected through code.

- Primary dataset: a public, realistic synthetic healthcare claims and fraud dataset from Kaggle that contains claim-level records, provider information, billing amounts, and a fraud label. I will download it programmatically using Python (requests or the Kaggle API) and load it into pandas.
- Context data: I will utilize the CDC Chronic Disease Indicators API to retrieve state-level metrics, including obesity, diabetes, and hospitalization rates, and the US Census API to obtain population and basic socioeconomic variables. All responses will be pulled using requests, parsed from JSON or CSV, and stored in structured Data Frames.

Datasets: https://www.kaggle.com/datasets/bonifacechosen/nhis-healthcare-claims-and-fraud-dataset

https://data.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators/hksd-2xuw/about_data

https://www.census.gov/data/developers/guidance/api-user-guide/example-api-queries.html

3. Planned Analysis and Modelling

- Feature engineering includes metrics such as claims per provider, average claim amount, proportion of high-cost procedures, and gaps between claims.
- The cleaned dataset is examined using descriptive statistics and correlation analysis.
- Baseline supervised models are developed, including logistic regression and a tree-based model.
- Model performance is assessed using precision, recall, and ROC AUC, with emphasis on the fraud class.
- Simple anomaly detection methods are applied to identify providers exhibiting unusual billing behaviour.

4. Planned Visualizations

Using Matplotlib and possibly Seaborn, I will create distribution plots for fraudulent versus non-fraudulent claims, a correlation heatmap for key features, bar charts of providers ranked by fraud risk, and scatter plots that compare disease burden with average claim cost or fraud rate by state. Together, these visuals will support a clear story that connects the technical model to its goal of reducing fraud losses and protecting honest policyholders.