**Project 4 Proposal: Fraud Analysis in Canada**

**Objective**

This project aims to build a machine learning model that predicts whether a reported case involves fraud or no fraud, using victim counts and dollar loss amounts as key indicators. Leveraging Canadian fraud data, the project will also analyze the financial impact of fraud and uncover patterns across demographic and geographic variables to help stakeholders understand and respond to fraud risks more effectively.

Fraud is a growing concern in Canada, with reported losses to the Canadian Anti-Fraud Centre (CAFC) exceeding $2 billion since 2021 alone:

* $383 million in 2021
* $530 million in 2022
* $569 million in 2023
* $638 million in 2024

The most commonly reported fraud types include investment scams, identity theft, merchandise fraud, romance scams, and fraudulent services. Yet, only a small percentage of victims formally report these incidents to authorities—posing a significant challenge to fraud prevention and enforcement.

Fraud reports continue to climb annually, with Ontario, Quebec, and British Columbia consistently reporting the highest case numbers. These trends underline the urgent need for data-driven models to support early detection, reduce financial losses, and inform strategic policy responses across Canada.

**Dataset**

We will use the Canadian Anti-Fraud Centre Reporting Data (2021–2025), available at: [**https://open.canada.ca/data/en/dataset/6a09c998-cddb-4a22-beff-4dca67ab892f**](https://open.canada.ca/data/en/dataset/6a09c998-cddb-4a22-beff-4dca67ab892f)

This dataset contains detailed records of reported fraud cases in Canada, including fraud type, region, victim demographics, and financial losses.

**Methodology**

* **Machine Learning -** We will use two classification algorithms—Random Forest and K-Nearest Neighbors (KNN)—to build and optimize a predictive model that classifies reports as fraud or non-fraud.
* **Model Evaluation Metrics**
  + Accuracy – how often the model makes correct predictions
  + Precision – how many predicted fraud cases are truly fraud
  + Recall – how many actual fraud cases the model identifies
  + F1 Score – harmonic mean of precision and recall
* **Data Cleaning & Structuring -** Handle missing values, normalize financial loss categories, and encode categorical variables.
* **Visualization – Matplotlib** to visualize trends across provinces, age groups, and fraud types to identify significant fraud patterns.

**Stakeholders & Impact**

* **Law Enforcement & Government –** Identify high-risk fraud zones and optimize resource allocation
* **Financial Institutions –** Improve fraud detection systems and enhance consumer protection
* **Public & Media –** Raise awareness of fraud patterns and support public education initiatives

**Team Members**

Eder Ortiz, Geraldine Valencia, Oludemilade Adenuga