1 About Dataset

The Financial Fraud Dataset is a collection of financial filings from various companies submitted to the U.S. Securities and Exchange Commission (SEC). It comprises data from 85 companies involved in fraudulent activities and an equal number of companies that are not involved in fraud. This dataset is designed for academic research, particularly in the field of financial fraud detection, utilizing both traditional machine learning models and large language models (LLMs).

Features:

- **Numéro d'identification / Number ID**: A unique identifier assigned to each fraud report for tracking purposes.
- **Date Received / Date reçue**: The date on which the fraud report was submitted to the relevant authorities.
- **Complaint Received Type**: The method through which the complaint was reported, such as online, phone, or email.
- **Type de plainte reçue**: The French equivalent of "Complaint Received Type," indicating how the report was submitted.
- **Country**: The nation where the fraud incident occurred or where the victim resides.
- **Pays**: The French term for "Country," used in reports to indicate the location of the fraud incident.
- **Province/State**: The specific administrative region within a country where the fraud took place.
- **Province/État**: The French term for "Province/State," used in reports to specify the geographical location of the incident.
- **Fraud and Cybercrime Thematic Categories**: Classifications of various types of fraud and cybercrime reported by victims, aiding in data analysis and prevention strategies.
- **Catégories thématiques sur la fraude et la cybercriminalité**: The French version of "Fraud and Cybercrime Thematic Categories," used in bilingual reports.
- **Solicitation Method**: The initial approach taken by the fraudster to contact the victim, which can include phone calls, emails, or in-person visits.
- **Méthode de sollicitation**: The French term for "Solicitation Method," detailing how victims were approached by fraudsters.
- **Gender**: The gender of the victim, which may be recorded for demographic analysis of fraud cases.
- **Genre**: The French term for "Gender," used in reports to categorize victims based on gender identity.

- **Language of Correspondence**: The language used in communication between the victim and the reporting authority or fraudster.
- **Langue de correspondance**: The French equivalent of "Language of Correspondence," indicating the language used in reports or communications.
- **Victim Age Range / Tranche d'âge des victimes**: The age group to which the victim belongs, providing insight into demographic trends in fraud cases.
- **Complaint Type**: A classification indicating the nature of the complaint, such as identity theft or phishing.
- **Type de plainte**: The French term for "Complaint Type," used to categorize reported incidents in a bilingual context.
- **Number of Victims / Nombre de victimes**: The total count of individuals affected by a specific instance of fraud reported.
- **Dollar Loss / pertes financières**: The total financial loss incurred by victims as a result of fraudulent activities.

Key Details:

- **Purpose**: To assist researchers in developing models that can effectively detect fraudulent activities, potentially saving significant financial resources for governments and banks.
- **Data Structure**: The dataset includes detailed financial statements and reports, specifically focusing on the Management Discussion and Analysis (MD&A) sections and other relevant filings.
- **Limitations**: The dataset is limited to cases filed with the SEC, which means it does not encompass all fraudulent cases globally. Additionally, the text data may require cleaning before further processing.