**HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS**

**Motivation**

Healthcare Provider Fraud is one of the biggest problems faced in Medical insurance. According to the government, the total Medicare spending increased exponentially due to frauds in Medicare claims. Healthcare fraud is an organized crime which involves peers of providers, physicians, beneficiaries acting together to make fraud claims.

Rigorous analysis of Medicare data has yielded many physicians who indulge in fraud. They adopt ways in which an ambiguous diagnosis code is used to adopt costliest procedures and drugs. Insurance companies are the most vulnerable institutions impacted due to these bad practices. Due to this reason, insurance companies increased their insurance premiums and as result healthcare is becoming costly matter day by day.

Healthcare fraud and abuse take many forms. Some of the most common types of frauds by providers are:

a) Billing for services that were not provided.

b) Duplicate submission of a claim for the same service.

c) Misrepresenting the service provided.

d) Charging for a more complex or expensive service than was actually provided.

e) Billing for a covered service when the service actually provided was not covered.

- For what purpose was the dataset created?

The goal of this project is to " predict the potentially fraudulent providers " based on the claims filed by them. Along with this, we will also discover important variables helpful in detecting the behaviour of potentially fraud providers. Further, we will study fraudulent patterns in the provider's claims to understand the future behaviour of providers.

- Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? Who funded the creation of the dataset?

The dataset is available to everyone on Kaggle under the name ‘HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS’. Its original contributor is Rohit Anand Gupta.

[HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS (kaggle.com)](https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis/code)

**Composition**

Dataset size:

* Training data: 32958 records which was split into training, validation and test set
* Features: 54 columns (e.g., total\_claims, unique\_benefeciaries, inpatient\_days)
* Target variable: ‘PotentialFraud’ (binary: ’Yes’ or ‘No’)

For the purpose of this project, we are considering Inpatient claims, Outpatient claims and Beneficiary details of each provider.

A) Inpatient Data

This data provides insights about the claims filed for those patients who are admitted in the hospitals. It also provides additional details like their admission and discharge dates and admited diagnosis code.

B) Outpatient Data

This data provides details about the claims filed for those patients who visit hospitals and not admitted in it.

C) Beneficiary Details Data

This data contains beneficiary KYC details like health conditions, region they belong to etc.

The dataset includes following key features:

* TotalClaims: Number of claims submitted by the provider.
* InpatientDays: Total number of inpatient days claimed.
* UniqueBeneficiaries: Number of unique patients for the provider
* OutpatientDays: Total number of outpatient days claimed.
* TotalReimbursetmentAmount: Total money reimbursed by insurance.

**Distribution**

The license applicable to this dataset is [CC0: Public Domain](https://creativecommons.org/publicdomain/zero/1.0/)

**Maintenance**

Dataset is maintained by Kaggle