Healthcare Dataset Documentation

1. Overview

Context This is a synthetic healthcare dataset designed to mimic real-world healthcare data. It was created to serve as a resource for data science, machine learning, and data analysis enthusiasts, enabling users to practice data manipulation and analysis skills within the context of the healthcare industry without compromising patient privacy.

Inspiration Real-world healthcare data is often sensitive and restricted by privacy regulations (such as HIPAA). To address the gap for educational and research data, this dataset was generated using Python's **Faker** library. It mirrors the structure and attributes commonly found in hospital records to foster innovation and knowledge sharing.

2. Dataset Specifications

- **Type:** Synthetic / Simulated Data
- **Domain:** Healthcare / Hospital Administration
- **Target Problem:** Multi-Class Classification

3. Data Dictionary

The dataset contains columns representing patient demographics, admission details, and clinical information.

Column Name	Description	Data Type / Examples
Name	Name of the patient.	String
Age	Age of the patient at time of admission.	Integer (Years)
Gender	Gender of the patient.	"Male", "Female"
Blood Type	Patient's blood group.	"A+", "O-", etc.
Medical Condition	Primary diagnosis or condition.	"Diabetes", "Hypertension", "Asthma"
Date of Admission	Date the patient was admitted.	Date
Doctor	Name of the attending physician.	String
Hospital	Name of the healthcare facility.	String
Insurance Provider	Entity covering medical costs.	"Aetna", "Blue Cross", "Medicare", etc.
Billing Amount	Amount billed for services.	Float
Room Number	Room accommodated during stay.	Integer
Admission Type	Circumstances of admission.	"Emergency", "Elective", "Urgent"

Column Name	Description	Data Type / Examples
Discharge Date	Date the patient was discharged.	Date
Medication	Medication prescribed/administered.	"Aspirin", "Penicillin", "Lipitor", etc.
Test Results	Outcome of medical tests.	"Normal", "Abnormal", "Inconclusive"

Export to Sheets

4. Usage Scenarios

This dataset is suitable for various analytical tasks:

- **Predictive Modeling:** Developing models to predict clinical outcomes.
- **Data Cleaning:** Practicing techniques to handle categorical and numerical data.
- **Visualization:** Creating dashboards to analyze healthcare trends (e.g., billing vs. condition).
- **Education:** Teaching machine learning concepts in a healthcare context.

Recommended Challenge

Multi-Class Classification: Users can treat `Test Results` as the target variable.

- **Classes:** 3 (Normal, Abnormal, Inconclusive)
- **Goal:** Predict the test result category based on patient demographics and medical conditions.

5. Acknowledgments & Disclaimers

- **Privacy:** This dataset is entirely synthetic. It does not contain real patient information and does not violate privacy regulations.
- **Image Credit:** Image by BC Y from Pixabay.