I have capstone project in AIML in BITS Pilani. In which I have to create a project with python code for medical claim fraud. Can you give me step by step plan explaining each step. I need to get the good dataset please help me to do

**🔹 Step-by-Step Plan for Medical Claim Fraud Detection (AIML Capstone Project)**

This plan will guide you through **data collection, preprocessing, feature engineering, model training, and evaluation** to build a **fraud detection system** using AI/ML.

**🔷 Phase 1: Understanding the Problem & Data Collection**

**✅ Step 1: Define the Problem Statement**

* **Objective**: Detect fraudulent health insurance claims.
* **Key Challenges**:
  + Imbalanced data (fraud cases are rare).
  + Data quality (missing, inconsistent data).
  + Feature extraction from text-based data (medical reports).

**✅ Step 2: Find or Create a Dataset**

**Option 1: Public Datasets**

* **Kaggle**: Search for "Medical Claim Fraud Detection" datasets.
  + Example Dataset
* **CMS Medicare Fraud Data** (USA): Link
* **Synthetic Data**: Generate a dataset using Python.
  + I can help create a **synthetic dataset** matching real-world fraud patterns.

**✅ Step 3: Understand the Data Structure**

* Explore dataset fields like:
* ['Patient ID', 'Total Amount', 'Hospital', 'Diagnosis Report',
* 'Discharge Summary', 'Fraud Status', 'Policy Number', 'Claim Date', etc.]
* Check missing values, data types, and distribution.

**🔷 Phase 2: Data Preprocessing & EDA**

**✅ Step 4: Clean & Preprocess Data**

* Handle missing values.
* Convert categorical variables (policy type, hospital name) into numerical form.
* Standardize text fields (diagnosis reports, discharge summaries).

**✅ Step 5: Exploratory Data Analysis (EDA)**

* **Data Imbalance Check**: Fraud vs. Non-Fraud cases.
* **Feature Correlation**: Identify important variables.
* **Fraud Trend Analysis**: Identify common fraud patterns.
* **Visualization**:
  + Pie chart: Fraud vs. Non-Fraud ratio.
  + Histogram: Claim amount distribution.
  + Box plot: Claim frequency per policyholder.

**🔷 Phase 3: Feature Engineering & Model Training**

**✅ Step 6: Feature Engineering**

* **Statistical Features**
  + Claim frequency per patient.
  + Deviation from average claim amount.
  + Time gap between successive claims.
* **Text Processing (NLP)**
  + Convert medical reports into structured features using **TF-IDF or Word2Vec**.
* **Anomaly Detection Features**
  + Flagging unusually high claim amounts.
  + Identifying suspicious patterns (e.g., frequent hospital changes).

**✅ Step 7: Split Data into Train/Test Sets**

**Step 1: Define the Problem Statement**

**Objective**

The goal of this project is to **detect fraudulent health insurance claims** using **machine learning techniques**. Fraudulent claims can lead to financial losses for insurance companies and increase premium costs for genuine customers. By developing a **fraud detection model**, we aim to identify suspicious claims before they are processed.

**Key Challenges & Considerations**

1️⃣ **Imbalanced Data**

* Fraudulent claims are rare (typically <10% of total claims).
* Machine learning models may be biased toward non-fraudulent claims.
* Need techniques like **oversampling, undersampling, or synthetic data augmentation**.

2️⃣ **Data Quality Issues**

* **Missing Data:** Some fields (e.g., diagnosis reports, hospital bills) may be incomplete.
* **Inconsistent Data:** Different formats for claim dates, billing codes, etc.
* **Data Cleaning & Preprocessing** is essential before modeling.

3️⃣ **Feature Engineering from Text Data**

* Medical reports, discharge summaries, and prescriptions contain crucial fraud indicators.
* Need **Natural Language Processing (NLP)** to extract insights.
* Convert text into features using **TF-IDF, word embeddings, or transformer-based models**.