Metadata-Driven Static & Dynamic AI Preprocessing for Government Survey Data

Project Description

Introduction

Government surveys in India, such as those from the **Ministry of Statistics & Programme Implementation (MOSPI)**, generate massive datasets through schemes like **NMDS** and **CMDS**.

These datasets are vital for policy-making but often arrive **unprocessed** — with missing values, coded responses, inconsistent formats, and requiring deep statistical expertise to interpret.

Current preprocessing is **manual, rigid, and fragile**, relying on dataset-specific scripts that break when formats change, causing delays in insights delivery.

Problem

- **Slow & manual** — Cleaning, mapping codes, and applying weights take days.
- **Low adaptability** — Scripts fail when dataset structures change.
- **Expert-heavy** — Needs statisticians familiar with dataset-specific rules.
- **Limited accessibility** — Processed outputs are mostly in English.

Solution

We propose an **AI-powered, metadata-driven preprocessing system** with **two parallel engines**:

1. **Static Engine** –For known datasets (NMDS, CMDS, CPI)
    - Applies predefined cleaning/mapping rules from official metadata.
    - Ensures speed, accuracy, and consistency.

2. **Dynamic Engine** – For unknown or changed datasets
    - Uses **NVIDIA Nim (LLaMA 3 70B)** to infer schema, detect anomalies, and recommend corrections.
    - Learns from past patterns — no need to rewrite scripts.

Metadata Integration:

Reads JSON/Excel metadata to **auto-map codes**, identify **design weights**, and apply **harmonization rules**

Workflow

1. **Data & Metadata Ingestion** – Upload raw dataset + metadata, or infer metadata.
2. **Schema Recognition** – Automatically select Static or Dynamic Engine.
3. **Data Cleaning** – Handle missing values, type conversions, outliers.
4. **Weight Application** – Apply design weights for statistical accuracy.
5. **Harmonization** – Standardize variables across datasets/years.
6. **Validation Loop** – Errors trigger automated reprocessing.
7. **Insights Generation** – Weighted summaries, margins of error, multilingual reporting.
8. **Output Delivery** – Clean dataset + dashboard-ready formats.

Key Technologies & Their Significance

1. NVIDIA Nim API (LLaMA 3 70B)

Purpose:

- Powers the **Dynamic Engine** to handle *unknown or unstructured datasets*.
- Automatically infers schema, detects anomalies, and suggests corrections without hardcoding.

Significance:

- Adapts to **any dataset format change** instantly.
- Reduces manual effort for new survey data preprocessing.
- Brings AI-driven intelligence to government dataset handling.

2.JSON SCHEMA

Purpose:

- Validates datasets against **metadata standards** (e.g., NMDS, CMDS).

Significance:

- Ensures incoming data meets required structure before processing.
- Prevents corrupted or incomplete datasets from entering the pipeline.
- Increases trust in automated preprocessing results.

3. **OpenAI Whisper**

Purpose:

- Converts **voice input** (in multiple languages) to text for dataset queries and instructions.

Significance:

- Makes the system **accessible for illiterate and rural users**.
- Enables natural, hands-free interaction with the platform.
- Supports multilingual inclusivity.

Key Benefits

- **Faster** – Reduces processing time from days to minutes.
- **Adaptive** – Works with both known and unknown dataset formats.
- **Accessible** – Multilingual outputs for wider reach.
- **Scalable** – Handles 24+ NMDS datasets without manual rewriting.

Impact

This solution **bridges the gap** between raw statistical data and **decision-ready insights**, empowering policymakers, researchers, and the public with **reliable, standardized, and multilingual** survey results — **delivered in minutes, not days**.