HiLabs Hackathon: Provider Data Quality Analytics & Dashboard

Organized By: HiLabs in collaboration with IIT Roorkee

Date: Sept 6 - Sept 7, 2025

Problem Statement: Smart Provider Credentialing Analytics Platform

Introduction

HiLabs invites IITR students to participate in an exciting hackathon focused on building intelligent healthcare provider data analytics systems. This event aims to foster innovation and identify new approaches for ensuring provider directory accuracy and compliance in the healthcare industry.

At HiLabs, we use AI to solve real world problems in the healthcare space. Provider credentialing accuracy is critical for patient safety, regulatory compliance, and operational efficiency. Errors in provider directories cost the healthcare industry over $3 billion annually and can lead to patient harm, claim denials, and compliance violations. Solutions that allow healthcare administrators to quickly identify and resolve provider data quality issues using interactive dashboards and intelligent analytics will significantly improve healthcare operations.

The Challenge

Your task is to create an intelligent analytics platform that takes provider credentialing datasets and builds both the analytical engine and user interface to identify, analyze, and present data quality issues. The platform should include data processing algorithms AND an interactive presentation layer (dashboard, chatbot, or web application) that healthcare administrators can use to explore and resolve credentialing problems.

Core Problems to Solve (Choose 2-3 focus areas)

1. Provider Entity Resolution & Deduplication Build algorithms to identify duplicate provider records across databases despite name variations, spelling differences, or formatting inconsistencies.

2. License Validation & Compliance Tracking Cross-reference provider licenses with state medical board databases to detect expired credentials, invalid license numbers, and specialty mismatches.

3. Data Quality Assessment & Standardization Analyze provider data for formatting inconsistencies, missing information, and anomalies. Implement standardization algorithms for phone numbers, addresses, and other key fields.

4. Interactive Analytics Dashboard/Chatbot (Must Include in submission): Create a user-friendly interface that presents your analysis results through visualizations, interactive queries, or conversational AI that healthcare administrators can use to explore provider data quality.

Sample Input Questions Your Platform Should Handle

Data Quality Queries: - "How many providers have expired licenses in our network?" - "Show me all providers with phone number formatting issues" - "Which providers are missing NPI numbers?" - "Find potential duplicate provider records"

Analytics & Insights: - "What is our overall provider data quality score?" - "Which specialties have the most credentialing issues?" - "Show me a summary of all data quality problems by state" - "Generate a compliance report for expired licenses"

Interactive Exploration: - "Filter providers by license expiration date" - "Show providers practicing in multiple states with single licenses" - "Compare provider information across different databases" - "Export a list of providers requiring credential updates"

What we provide

1. **Provider Directory Dataset** (`provider_roster_with_errors.csv`) - 500+ provider records with data quality issues including duplicate entries, expired licenses, and formatting inconsistencies
2. **NY State Medical License Database** (`ny_medical_license_database.csv`) - Medical licensing authority records for cross-validation
3. **CA State Medical License Database** (`ca_medical_license_database.csv`) - California medical board licensing data
4. **Mock NPI Registry** (`mock_npi_registry.csv`) - National Provider Identifier database for validation

Rules

- Only open-source libraries and frameworks are allowed. You **cannot upload datasets to any servers, nor can you make API calls to proprietary services.** All processing must be done locally
- You must submit a repository of your code, along with a script to run your application and a readme that has clear instructions on setting up and running your platform. If we cannot run your code, you will be disqualified
- Your solution must include BOTH analytical algorithms AND a user interface (web dashboard, desktop app, or chatbot)

Suggestions and notes

- You are free to use any technique to build your platform. We recommend focusing on 2-3 core problems rather than trying to solve everything superficially
- Since you are scored on both analytical accuracy and user experience, balance your time between backend algorithms and frontend presentation

- Good architectural skills will be valued. Consider building modular components that can be easily extended
- **Performance matters**: Your platform should handle the provided datasets efficiently and provide responsive user interactions

Submission Format

Your submission should include:

1. **Analytics Platform** that processes provider datasets and identifies data quality issues through both algorithms and interactive interface
2. **Summary metrics for data issues found through the analytics**
3. **User Interface** - Web dashboard, desktop application, or chatbot that allows healthcare administrators to explore your analysis results

   Basic Queries that should be supported through bot or visualization (Only minimum and not exhaustive):

   1. *"How many providers have expired licenses in our network?"*
   2. *"Find potential duplicate provider records"*
   3. *"What is our overall data quality score?"*
   4. *"Show me all providers with phone number formatting issues"*

4. A readme doc detailing your approach, architecture decisions, and how to run your platform
5. Code as a repository with above readme. Please ensure the repo is publicly accessible to avoid access issues, and put the link in the first line of the readme
6. **Demo script or video** showing your platform in action with real results from the provided datasets
7. **You cannot upload the datasets to any third-party servers or make API calls to proprietary services**

Platform Architecture Requirements

Your solution should demonstrate:

- **Data Processing Layer**: Efficient algorithms for entity resolution, validation, and quality analysis
- **Analytics Engine**: Statistical analysis, scoring, and insight generation
- **Presentation Layer**: Interactive dashboard, chatbot, or web application

- **Integration**: Seamless connection between analysis results and user interface
- **Documentation**: Clear setup instructions and usage examples
  - Include a README, run script, repo link, and details about the required libraries, OS, and language version.
- Expected query responsiveness: <2 sec for 500+ rows

Bonus: Dockerizing the solution would be a plus as it make easier to run the solution consistently.

Evaluation Criteria

- **Analytical Accuracy** (30%) - Correctness of duplicate detection, license validation, and data quality analysis
- **User Experience** (25%) - Intuitive interface design, ease of use, and clarity of information presentation
- **Technical Implementation** (20%) - Code quality, architecture decisions, and system performance
- **Innovation & Creativity** (15%) - Novel approaches to credentialing challenges or unique interface designs
- **Completeness & Documentation** (10%) - Working end-to-end solution with clear setup instructions

Bonus Features (+5 points each) - Real-time data processing capabilities - Export functionality for compliance reports - Machine learning-based risk prediction - Mobile-responsive design - Advanced data visualizations