# Improving Field Survey Accuracy and Efficiency with an AI-Based Classification Tool

*A clear and concise title that reflects the problem statement*

| Details | |
|---|---|
| **Particulars** | **Details** |
| **Name of Ministry/Department/Division** | Ministry of Statistics and Programme Implementation/Computer Centre |
| **Address** | 10, East Block, Rama Krishna Puram, New Delhi, 110066 |
| **Name of the Nodal Officer and Designation** | |
| **Phone Number (Nodal Officer)** | |
| **Email ID** | |
| **Domain/Area: Problem Statement** | Official Statistics |
| **Category of problem statement**<br>*(Select all that are applicable)* | ☑ Data Collection<br>☑ Data Processing<br>☑ Data Quality<br>☑ Statistical Methodology<br>☐ Data Accessibility<br>☑ Data Integration and Interoperability<br>☐ Data Timeliness<br>☐ Data Standardization<br>☑ Data Utilisation and Analysis<br>☐ Data Visualization<br>☐ Data Transparency<br>☐ Technological Infrastructure<br>☑ Data Management<br>☑ Data Extraction and Pipeline Creation<br>☐ Others (please specify) |
| **What kind of support do you expect from the solution giver?** | ☑ Research and development of new methodology<br>☑ Development or modification of tools/process |
| **What kind of support/resources will you be able to share with MoSPI?** | ☑ Subject Matter Experts (SMEs)<br>☑ Existing Datasets<br>☑ Technical Resources<br>☐ Collaborative Networks<br>☑ IT Infrastructure<br>☑ Physical Infrastructure (space for partners to work from)<br>☐ Others (please specify) |

## Problem Statement

### A. Problem statement Identified (Max 200 words) *

*(Write a crisp and specific problem statement identified by the Ministry/Department/Divion. Include aspects such as what the problem is, whom it impacts, and scale of impact. Try to give data figures wherever possible. Make sure the problem statement is related to official statistics areas.)*

*National Industrial Classification (NIC) codes are an essential statistical standard used in industrial surveys, labour statistics, national income estimates, population census etc for capturing the data related to economic activity of different entities. All economic activities are grouped into section coded from A to U, every section into division with 2-digit numeric code, every division into group with 3-digit numeric code, every group into class with 4-digit numeric code, and every 4-digit class into 5-digit sub-class. The section level is the broader category and subclass level is the most granular level of every economic activity. The enumerators collecting data from households or an individual entering data on portal for various surveys need to select the corresponding industrial code based on the text description given. There are more than 500+ sub-class level codes with corresponding activity description available as free text. Atmost care and research were done in formulating the codes and their descriptions so that it should Includes the most non-ambiguous and interpretable text description for every code.*

*For selecting a NIC code of an economic activity, enumerator have to search the entire list using exact keyword-based search of the description available at class or subclass level. Not every economic activity can be described using the exact english words available in the text description of subclass list. Sometimes synonyms of the keywords mentioned in the description or related activity is provided which don't have any exact match with the text description representing the economic activity. So, in such cases either surveyor or user have to interpret the similarities between description available in the list and description of the actual economic activity and select the most appropriate NIC code. It may sometimes result in in-correctly capturing of the economic activity as well as consuming a lot of time and in return results in misleading data as well as reducing the productivity of the enumerator/user.*

### B. Methodology used for identifying the problem statement. *

*(Detail how the problem was identified, including any studies conducted, resources referred to, or methodologies applied.)*

*The problem was identified from inputs from past experience of enumerators*

*Mandatory

*For non-mandatory questions, enter "N/A" in the given box if not applicable* *and feedbacks available from enumerators or users and as well as from the supervisors who conducts trainings of the enumerators for NIC code classifications. Also, whenever the NIC codes or descriptions changes, the concerned officials must remember the changes. It again takes time to memorise the codes in the data collection exercise.*

## C. Challenges imposed and need for solving them. *

*(Explanation of the current situation, including relevant data and statistics that highlights the need for addressing this problem. List all key stakeholders affected by this problem, including internal teams, external partners, or end-users. Highlight the potential long-term impacts if the problem remains unsolved.)*

*Any solution available as of now for NIC code search is available with syntax-based keyword matching search option like searching using exact keyword match in the pdf document of the list. The current process lacks semantic based search option and the enumerator/user either needs to remember the exact words to find the correct Industry code or have to search entire list and select the code. It requires the surveyor/user to be experienced enough in language interpretation and trained/aware enough about the economic activities available in the NIC code list. It is a time-consuming process and may result in an inaccurate data collection. If this remains unsolved, it may lead to misleading data Insights required for policy formulation and execution and productivity loss of enumerators/users.*

***Key Stakeholders Impacted :***

*Enumerators, MoSPI and all ministries/departments, research institutions/think tanks relying on various surveys where economic activity specific Information is captured and used for economic and policy formulation and research.*

## D. Existing processes/systems in place to deal with the challenges (Max 150 words). *

*(How is the Ministry/Department/Division currently addressing the problem statement? In case no way has been found to manage it, kindly mention that as well.)*

*Currently, enumerators rely on their experience to identify the closest matching sub-division within the NIC classifications. This expertise is developed over time, primarily through trial and error, as agents gradually learn to navigate the NIC list. However, there is no formal system in place to accurately track or standardize this process, leaving NIC code selection process largely dependent on individual familiarity and intuition.*

## E. Expected outcome(s) for stakeholders' post resolution. *

*(Clearly outline the benefits and improvements the impacted stakeholders will experience once the problem is resolved. Also mention the essential features of the solution.)*

1) ***Improved Efficiency:*** *An AI/ML solution with advanced keyword and context-based search will allow enumerators/users to quickly retrieve precise classifications, reducing the time spent on manual searches and enabling faster data collection.*
2) ***Enhanced Accuracy:*** *With automated search assistance, agents are less likely to misclassify entries, ensuring higher data accuracy and reliability in classification, which is critical for economic indices and policy analysis.*
3) ***Better Resource Utilization:*** *Field agents can redirect time and effort from tedious document searches/doubt clarification time to other survey tasks, increasing overall productivity and enhancing the quality of data collection.*

**F. How urgent do you consider it to solve this problem? ***

*(What degree of impact does the problem have on operations?)*

☐ **High Priority:** The problem significantly impacts daily operations; needs immediate attention.

☑ **Medium Priority:** The problem affects productivity or efficiency but does not halt operations. It should be addressed within a reasonable timeframe.

☐ **Low Priority:** The problem has minimal impact on overall operations and can be resolved later without major consequences.

**G. What is your expected timeline for resolution? ***

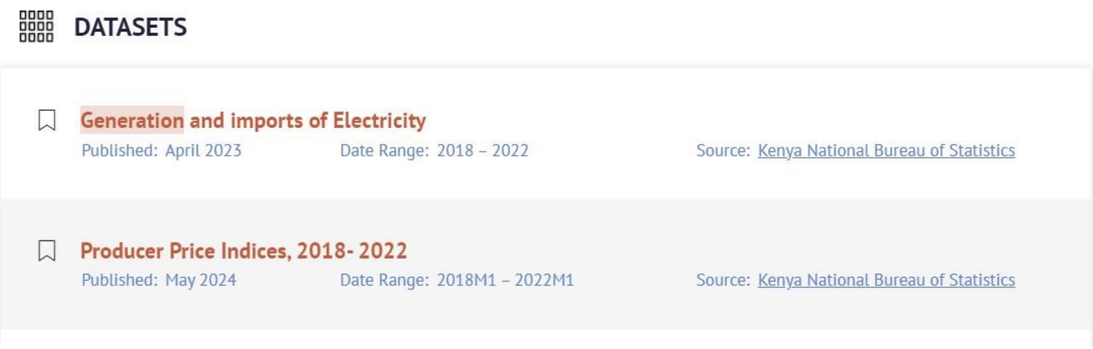
*(Mention the duration within which you expect a resolution)*

*A resolution is expected between 4 to 6 months.*

**H. Share any global best practices you'd like to highlight?**

*(Mention any global best practices you know of that could address your issue or be implemented to solve it. Include links where possible.)*

*Globally the several software companies implemented generic search engines as well a few NSO's/Statistics Bureau's implemented semantic* *based search for their resources in their portals or applications. E.g. Kenya National Bureau of Statistics implemented the semantic based search in their portal for searching open data available in the country. For a query "Electricity production", the following results are returned.*



DATASETS

Generation and imports of Electricity
Published: April 2023 Date Range: 2018 – 2022 Source: Kenya National Bureau of Statistics

Producer Price Indices, 2018- 2022
Published: May 2024 Date Range: 2018M1 – 2022M1 Source: Kenya National Bureau of Statistics

*The production and generation keywords are semantically related, therefore returned as the top result.*
*Similarly, Office of National Statistics, UK developed semantic search facility for SIC codes available in UK. ClassifyAI Is designed to classify the SIC codes based on text input.*

***(Complete this section only if the Ministry/Department submitting the proposal has potential solutions in mind)***

## I. Proposed solutions (Max 200 words)

*(Provide an overview of the proposed solutions, including the key milestones and tentative timelines for each phase of implementation. Try to post your idea in points /diagrams /infographics /pictures.)*

*The solution will include a web-app with an intelligent search feature like context-based search. A user needs to enter/speak the text/keyword within the search bar with the closest keyword to the code description available as free text. The implementation of underlying Information retrieval system requires following phases:*

1) ***Data Preparation Phase:*** *NIC 2008 data tables available as pdf files need to be ingested into a database. Any other relevant text description data where user input as a free text was captured in the past may help in training the ai/ml algorithms.*
2) ***Feature Engineering Phase****: Text based features can be selected as seed words for each industry and their synonyms/related text can be generated from alternate sources of data like web or ai enabled data generation open source tools or free text Inputs received from users during keyword-based searches for codes.*
3) ***Modelling Phase:*** *The ml-based information retrieval models like vector space models, probabilistic models, latent semantic analysis, neural ranking models etc can be Implement to retrieve the top-k best matches for the*

*corresponding economic activity.*

*4) **Model Performance Evaluation Phase:** The model will be evaluated on parameters like accuracy, precision and recall. Based on trade-off between precision and recall, a suitable hyperparameters of the model will be selected.*

*5) **Model Deployment Phase:** Since the model will be used as an enhanced feature of already existing data collection application, it should be deployed either in the web application or mobile application.*

## J. Analysis of the feasibility of the solution

*(Evaluate the viability of the proposed solution, considering its technical, financial, and operational aspects, along with identifying potential challenges and risks.)*

*The proposed solution is technically feasible leveraging open source NLP and ML algorithms. The solution will be a standalone application with ml features from open source models. The solution can be feasible using LLM as well as open-source solutions available for NLP like spacy etc.*

*Mandatory

*For non-mandatory questions, enter "N/A" in the given box if not applicable*