Development of an AI/ML based predictive model for estimation of Monthly Per Capita Expenditure (MPCE)

Details	
Particulars	**Details**
Name of Ministry/Department/Division	Ministry of Statistics and Programme Implementation/Data Informatics and Innovation Division
Address	10, East Block, Rama Krishna Puram, New Delhi, 110066
Name of the Nodal Officer and Designation	Ms. Madhura Roy, Director
Phone Number (Nodal Officer)	8860570648
Email ID	madhura.roy@gov.in
Domain/Area: Problem Statement	Official Statistics in general
Category of problem statement *(Select all that are applicable)*	☐ Data Collection ☐ Data Processing ☐ Data Quality ☑ Statistical Methodology ☐ Data Accessibility ☑ Data Integration and Interoperability ☑ Data Timeliness ☐ Data Standardization ☐ Data Utilisation and Analysis ☐ Data Visualization ☐ Data Transparency ☐ Technological Infrastructure ☐ Data Management ☐ Data Extraction and Pipeline Creation ☐ Others (please specify)
What kind of support do you expect from the solution giver?	☑ Research and development of new methodology ☐ Development or modification of tools/process
What kind of support/resources will you be able to share with MoSPI?	☐ Subject Matter Experts (SMEs) ☑ Existing Datasets ☐ Technical Resources ☐ Collaborative Networks ☐ IT Infrastructure ☐ Physical Infrastructure (space for partners to work from) ☐ Others (please specify)

*Mandatory

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

Problem Statement

A. Problem statement Identified (Max 200 words) *

(Write a crisp and specific problem statement identified by the Ministry/Department/Division. 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.)

The Household Consumption Expenditure Survey (HCES) is designed to collect information on consumption of goods and services by the households. The survey also collects some auxiliary information on household characteristics and demographic particulars of the members of the households. Information collected in HCES is useful for understanding the consumption pattern, standard of living and hence, well-being of the households. Besides, the data of the survey provides budget shares of different commodity groups that is used for preparation of the weighting diagram for compilation of official Consumer Price Indices (CPIs). The data collected in HCES is also utilized, inter-alia, for deriving various other macroeconomic indicators. The survey covered the whole of the Indian Union except a few inaccessible villages in the Andaman and Nicobar Islands. Information in the survey was collected from **8,723 villages** and **6,115 urban blocks** spread over the entire country covering **2,61,746 households (1,55,014 in rural areas and 1,06,732 in urban areas**). The survey also collected some auxiliary information on household characteristics and demographic particulars of the members of the households in **Questionnaire - HCQ: Household Characteristics**. The survey is conducted every five years leading to non-availability of latest MPCE estimates, leading to lag in the availability of consumption pattern at centre and the state level for various strata of society. Therefore, it is proposed to create a predictive model for estimating expenditure data at the household level so that the consumption pattern can be found.

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 through a detailed review of existing processes and feedback received from various stakeholders, which highlighted the significant gap between two consecutive Household Consumption Expenditure Survey. Insights from this exercise, combined with an analysis of research paper available on this topic, led to the identification of this problem.

*Mandatory

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

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.)

The survey is conducted every five years leading to non-availability of latest MPCE estimates, leading to lag in the availability of consumption pattern at the Centre and at the State level for various strata of society. Therefore, it is proposed to create a predictive model for estimating expenditure data at the household level so that the consumption pattern can be found. This hackathon represents a pioneering initiative to leverage cutting-edge AI/ML techniques for enhancing the utility of HCES data, enabling more timely and insightful analysis. This initiative may help in applying model-based estimates using our survey data to fill the data gap in the interim.

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.)

In the current system, HCES is generally conducted once in every 5 years.

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.)

- **Enable expenditure metric to be released more frequently**: The predictive model for estimation of MPCE using demographics and other household characteristics will enable predicting the MPCE from surveys which are conducted more frequently than HCES. It will help the researchers in predicting the trend of MPCE before actual release of the MPCE after 5 years.
- **Create the interoperability framework for cross-survey predictions**: This cross-survey prediction will establish a framework for metrics derived from other surveys like PLFS to be predicted using the common features present in two surveys leading to filling of the data gap.

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 at a later time without major consequences.

G. What is your expected timeline for resolution? *

(Mention the duration within which you expect a resolution)

A resolution is expected within 6 to 9 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.)

World bank is working on survey imputation models and have published several papers on this topic.

(Complete this section only if the Ministry/Department/Division 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.)

Participants are expected to build a model to estimate household consumption expenditure based on survey data. The predictive model for estimation of MPCE using demographics and other household characteristics will enable predicting the MPCE from surveys which are conducted more frequently than HCES. It will help the researchers in predicting the trend of MPCE before actual release of the MPCE after 5 years.

*Mandatory

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

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 ris

The proposed solution is feasible and it is evaluated on different metrics such as MAPE and R^2. It is also checked for data drift handling as well as on Explainibilty of model.

*Mandatory

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