ADS_PHASE1

ASSESSEMENT OF MARGINAL WORKERS IN TAMILNADU-A SOCIO ECONOMIC ANALYSIS

CONTENTS:

Problem definition:

The primary goal of this project is to comprehensively assess the socioeconomic conditions and characteristics of marginal workers in Tamil Nadu, with a focus on gaining insights into their employment patterns, living conditions, and well-being. This assessment aims to address the following key questions:

- **Definition of Marginal Workers:** Clearly define the criteria and characteristics that categorize individuals as marginal workers in the context of Tamil Nadu.
- **Population Size:** Determine the size and distribution of the marginal worker population across different regions within Tamil Nadu.
- **Employment Patterns:** Analyze the types of employment and sectors in which marginal workers are engaged. Identify any seasonal variations or shifts in employment.
- **Income and Earnings:** Investigate the income levels, earnings, and wage disparities among marginal workers compared to other occupational categories.
- **Living Conditions:** Assess the housing conditions, access to basic amenities, and healthcare facilities available to marginal workers and their families.
- **Social and Economic Inclusion:** Examine the level of social and economic inclusion of marginal workers, including access to education, social security, and government welfare schemes.
- **Challenges and Vulnerabilities:** Identify the unique challenges, vulnerabilities, and risks faced by marginal workers, such as job insecurity, health issues, or discrimination.
- **Policy Evaluation:** Evaluate the effectiveness of existing government policies and programs targeted at improving the livelihoods and well-being of marginal workers.

Design thinking:

- Data collection
- Data preprocessing
- Feature engineering
- Model selection
- Model training
- Evaluation

- Data collection:

The process of gathering and analyzing accurate data from various sources to find answers to research problems, trends, and probabilities, etc., to evaluate possible outcomes is Known as Data collection.

Methods of data collection:

• Surveys, quizzes, and questionnaires.

• Interviews.

• Focus groups.

• Direct observation.

- Data preprocessing:

  Data processing, manipulation of data by a computer. It includes the conversion of raw data to machine-readable form, flow of data through the CPU and memory to output devices, and formatting or transformation of output. Any use of computers to perform defined operations on data can be included under data processing.

  The four main stages of data processing cycle are:

  - Data collection.
  - Data input.
  - Data processing.
  - Data output.

- Feature engineering:

  Feature engineering involves a set of techniques that enable us to create new features by combining or transforming the existing ones. These techniques help to highlight the most important patterns and relationships in the data, which in turn helps the machine learning model to learn from the data more effectively.

  Types of feature Engineering:

  •7 of the Most Used Feature Engineering Techniques. Hands-on Feature Engineering with Scikit-Learn, TensorFlow, Pandas and Spicy.

  •Encoding. Feature encoding is a process used to transform categorical data into numerical values that can be understood by ML algorithms.

  •Feature Hashing.

  •Binning / Bucketizing.

  •Transformation.

- Model Selection:
  - ✓ Time Series Forecasting

  There are four components that a time series forecasting model is comprised of:

        - Trend
        - Seasonality
        - Cyclical variations
        - Random or irregular variations
    - ✓ Autoregressive (AR)
    - ✓ Autoregressive Integrated Moving Average (ARIMA)
    - ✓ Seasonal Autoregressive Integrated Moving Average (SARIMA):
- Model training:

Model training is the phase in the data science development lifecycle where practitioners try to fit the best combination of weights and bias to a machine learning algorithm to minimize a loss function over the prediction range. The purpose of model training is to build the best mathematical representation of the relationship between data features and a target label (in supervised learning) or among the features themselves (unsupervised learning). Loss functions are a critical aspect of model training since they define how to optimize the machine learning algorithms. Depending on the objective, type of data and algorithm, data science practitioner use different type of loss functions. One of the popular examples of loss functions is Mean Square Error (MSE)

- Evaluation:

Maybe the most popular and simple error metric is MAE: MAE: The Mean Absolute Error is defined as: While the MAE is easily interpretable (each residual contributes proportionally to the total amount of error), one could argue that using the sum of the residuals is not the best choice, as we could want to highlight especially whether the model incur in some large errors.