Assessment of Marginal Workers in Tamil Nadu - A Socioeconomic Analysis

PROBLEM STATEMENT:

Marginal workers in Tamil Nadu are confronted with a myriad of socioeconomic challenges, including irregular employment, insufficient income, substandard working conditions, lack of education and healthcare access, and limited social protections. These adversities hinder their potential for upward mobility and overall development, perpetuating a cycle of poverty and vulnerability.

PROJECT OBJECTIVE:

The objective of this project is to conduct a thorough socioeconomic analysis of marginal workers in Tamil Nadu, unraveling the root causes and manifestations of their challenges. By doing so, we aim to propose data-driven policy recommendations and interventions that can improve their living conditions, enhance employment prospects, and facilitate their socio-economic empowerment within the state. Through this analysis, we strive to advocate for equitable policies, targeted interventions, and a conducive environment that enables marginal workers to break the cycle of poverty and achieve a better quality of life.

DESIGN THINKING PROCESS:

1.Empathize:

- Begin by immersing yourself in the environment and context of marginal workers in Tamil Nadu. Understand their daily lives, challenges, and aspirations.
- Conduct interviews, surveys, and observations to gather insights directly from the marginal workers and other stakeholders.
- Listen to their stories, concerns, and experiences to develop a deep understanding of their needs and the factors affecting their socio-economic status.

2.Define:

- After gathering insights, define the specific problems and challenges faced by marginal workers in Tamil Nadu.
- Identify the key issues related to their employment, income, access to resources, and overall well-being.
- Create a clear and concise problem statement that outlines the challenges to be addressed in the analysis.

3.Ideate:

- Brainstorm potential solutions and strategies for addressing the identified challenges.
- Encourage creative thinking within your research team and consider a wide range of possibilities.
- Explore innovative ways to improve the socio-economic conditions of marginal workers, keeping their needs and aspirations in mind.

4.Prototype:

- Develop prototypes or models of the proposed solutions and analysis methods. This may include research frameworks, data collection tools, and analysis techniques.
- Create a plan for data collection, analysis, and visualization that can help in understanding the socio-economic conditions of marginal workers.
- Consider how to effectively test and refine your analysis tools and methods during the testing phase.

5.Test:

- Implement your research and analysis plan by collecting data and conducting the socio-economic analysis of marginal workers in Tamil Nadu.
- Continuously test and refine your methods to ensure they are effective in addressing the defined challenges.
- Seek feedback from stakeholders and experts to improve the accuracy and relevance of your findings.

PHASES OF DEVELOPMENT:

Data Collection:

In this project, data collection involves gathering relevant socioeconomic data pertaining to marginal workers in Tamil Nadu. This may include information such as income levels, employment types, education, living conditions, access to healthcare, and other socio-economic indicators.

Data sources could include government surveys, interviews, field studies, and existing databases.

Data Preprocessing:

- Once data is collected, it needs to be preprocessed to ensure its quality and suitability for analysis. This phase might involve tasks such as:
- Handling missing data: Address any missing values in the collected data.
- Data cleaning: Remove outliers or erroneous data points.
- Data transformation: Convert data into a suitable format, deal with categorical variables, and create new features if necessary.

Feature Scaling:

Feature scaling in this context could involve ensuring that the various socio-economic indicators are on a common scale to avoid any particular indicator dominating the analysis. For instance, if income is measured in thousands and education level is measured on a scale of 1-10, scaling would be needed.

Model Training:

- The model training phase in this project might involve using statistical or machine learning models to analyze the socio-economic data. Potential models could include regression analysis, clustering, or classification models, depending on the specific goals of the analysis.
- Training these models would entail using a subset of the preprocessed data to build relationships or make predictions related to marginal workers' socio-economic status in Tamil Nadu.

Model Evaluation:

- In the context of a socioeconomic analysis, model evaluation might involve assessing the accuracy and relevance of the models used. Specific evaluation metrics could be defined based on the goals of the analysis.
- For example, you might use metrics such as R-squared for regression analysis or silhouette score for clustering to evaluate how well the model represents the socio-economic characteristics of marginal workers.

DATASET DESCRIPTION:

- ➢ The dataset appears to be Marginal Workers classified by Age , Industrial category and sex census 2011- India and states.
- ➢ Columns in the dataset include 'Industrial Category', 'Worked for months' and 'Age group'
- ➢ The data appears to be time-series data .

DATA PREPROCESSING:

1.Loading the Dataset:

The code begins by importing the necessary libraries, such as pandas for datamanipulation and scikit-learn for preprocessing and modeling.

It loads the dataset from a CSV file using the pd.read_csv() function.

2. Handling Missing Values:

The code checks for missing values in the dataset using the df.isnull().sum() method to identify columns with missing data.

It employs the SimpleImputer from scikit-learn with the 'mean' strategy to fill missing values with the mean values of the respective columns. This step ensures that the dataset is complete and ready for analysis.

3. Selecting Relevant Columns:

The code selects only the relevant columns from the dataset, which include 'worked_hours', 'salary'.

4. Data Transformation:

It ensures that the 'hours' column is in datetime format using the pd.to_datetime() method.

The 'hours' column is set as the index of the DataFrame to make it the time series index.

MODEL TRAINING PROCESS:

Model Selection:

Based on the project's objectives (e.g., classification, regression) and the nature of the data, select a set of candidate machine learning models. Common choices include:

1. Logistic Regression (for binary classification)

2. Random Forest

3. Support Vector Machines (SVM)

4. Gradient Boosting Machines (e.g., XGBoost, LightGBM)

Model Training:

Train each selected model using the preprocessed training dataset and the features identified through feature selection. Optimize hyper parameters using techniques like grid search, random search, or Bayesian optimization to improve model performance.

Model Evaluation:

Evaluate the models using appropriate evaluation metrics based on the project's objectives:

1. For Classification: Accuracy, Precision, Recall, F1-Score, ROC-AUC, etc.

2. For Regression: Mean Squared Error (MSE), R-squared, Mean Absolute Error (MAE), etc.

INSIGHTS:

Feature engineering:

It involves selecting, transforming, and enhancing the features (attributes or variables) in the dataset to ensure they provide valuable and relevant information for analysis and modeling.

Feature engineering aims to maximize the predictive power of the models by extracting meaningful insights from the available data, ultimately aiding in achieving accurate and meaningful outcomes in the project's analysis and recommendations.

Visualization:

The code includes visualizations to show the working hours of marginal workers, which is crucial for understanding how the data is transformed for modeling.

Additionally, it visualizes the rolling mean feature, which helps users interpret the impact of this feature on predictions.

RECOMMENDATIONS:

Data Collection Plan:

Plan the data collection process carefully. Identify the sources of data, such as government surveys, field studies, or interviews. Determine the variables to be collected, including age, sex, worked hours, and any other relevant socio-economic indicators.

Data Quality Assurance:

Implement measures to ensure data quality. This includes validating the collected data for accuracy, consistency, and completeness. Address any issues related to data quality during the collection phase.

Ethical Considerations:

Ensure that data collection and analysis adhere to ethical standards and protect the privacy and rights of the participants. Obtain necessary permissions and consent when collecting data, especially if personal information is involved.

Data Preprocessing:

As mentioned earlier, conduct data preprocessing, including handling missing values, data cleaning, data transformation, and feature scaling, to prepare the dataset for analysis.

Feature Engineering:

Consider creating new features or aggregating data to better represent the socio-economic characteristics of marginal workers. Feature engineering can enhance the quality of your analysis.

Data Visualization:

Utilize data visualization techniques to explore and present your data effectively. Visualizations such as charts, graphs, and maps can help convey complex socio-economic information in an accessible manner.

Statistical Analysis and Modeling:

Select appropriate statistical methods or machine learning models for your analysis. Regression, clustering, or classification models may be relevant, depending on your research objectives.

Interpretation and Context:

Ensure that the analysis is interpreted within the socio-economic context of Tamil Nadu. Consider the regional and cultural factors that might influence the findings.