**INNOVATIVE: MARGINAL WORKERS IN TAMILNADU**

DESCRIPTION:

Tamil Nadu, an Indian state, houses a sizable number of marginalized workers, including those in agriculture, construction, and the informal sector. They often grapple with economic hardships, limited access to healthcare and education, and vulnerability to exploitation. The problem statement seeks to thoroughly assess their living conditions, job patterns, and well-being in order to identify key issues and shape policy recommendations for improving their quality of life.

ALGORITHM:

import pandas as pd

import matplotlib.pyplot as plt

# Step 1: Load the data

data = pd.read\_csv('marginal\_workers\_data.csv')

# Step 2: Data Cleaning and Preprocessing (Adapt as needed)

# Handle missing values, duplicates, and outliers

data.dropna(subset=['age'], inplace=True)

data.drop\_duplicates(inplace=True)

# Additional data preprocessing steps

# Step 3: Data Exploration

# Example: Plot a histogram of the age of marginal workers

plt.hist(data['age'], bins=20)

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.title('Age Distribution of Marginal Workers')

plt.show()

# Step 4: Feature Engineering (Add as needed)

# Example: Create a new feature 'is\_urban' based on a location column

data['is\_urban'] = data['location'].apply(lambda x: 1 if x == 'Urban' else 0)

# Step 5: Data Analysis (Adapt for your specific analysis)

# Define your analysis goals and perform necessary computations

# Example: Calculate the average age of marginal workers

average\_age = data['age'].mean()

print(f'Average age of marginal workers: {average\_age:.2f}')

# Step 6: Model Building (if applicable)

# If your project involves predictive modelling, build and train models

# Step 7: Visualisation and Reporting

# Create more visualisations and reports to communicate findings

# Example: Save the histogram plot to a file

plt.hist(data['age'], bins=20)

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.title('Age Distribution of Marginal Workers')

plt.savefig('age\_distribution.png')

THE COLUMNS THAT COULD BE USED:

1. Demographic information (age, gender, education, caste, etc.)
2. Employment information (occupation, income, etc.)
3. Social security information (access to schemes, credit, healthcare, etc.)
4. Migration information (place of birth, current residence, reasons for migration, etc.)
5. Other relevant information (skills, training, experience, etc.)

LIBRARIES TO BE USED:

Python libraries:

NumPy: A library for scientific computing.

Pandas: A library for data manipulation and analysis.

Scikit-learn: A library for machine learning.

R libraries:

dplyr: A library for data manipulation with R.

ggplot2: A library for data visualization with R.

caret: A library for machine learning in R.

**The library can be downloaded from the following websites:**

* NumPy: https://numpy.org: https://numpy.org
* Pandas: https://pandas.pydata.org: https://pandas.pydata.org
* Scikit-learn: https://scikit-learn.org: https://scikit-learn.org
* dplyr: https://dplyr.tidyverse.org: https://dplyr.tidyverse.org
* ggplot2: https://ggplot2.tidyverse.org: https://ggplot2.tidyverse.org:
* caret: https://caret.r-forge.r-project.org: <https://caret.r-forge.r-project.org>

TESTING AND TRAINING:

Training set:

The training set should be a representative sample of the population of marginal workers that you are interested in studying. The training set should contain a variety of different types of workers, including workers from different industries, occupations, and demographic groups.

Test set:

The test set should be a holdout sample of the population that is not used to train the model. The test set should be used to evaluate the performance of the trained model on unseen data. The test set should be similar in composition to the training set, but it should not contain any of the same data points.

**Some common metrics that can be used include:**

1. Accuracy: This metric measures the percentage of data points that the model predicts correctly.
2. Precision: This metric measures the percentage of data points that the model predicts as positive that are actually positive.
3. Recall: This metric measures the percentage of positive data points that the model predicts correctly.
4. F1 score: This metric is a harmonic mean of precision and recall. It is a good metric to use when both precision and recall are important