**Marginal Workers in Tamil Nadu**

**Table of Contents:**

**1. Introduction**

**2. Project Goals**

**3. Data Used**

**4. Data Preprocessing and Cleaning**

**5. Data Analysis**

**6. Data Visualization**

**7. Feature Engineering**

**8. Model Training and Evaluation**

**9. Findings and Insights**

**10. Conclusion**

**1. Introduction:**

* The "Marginal Workers in Tamil Nadu" project aims to analyze and understand the characteristics, distribution, and socio-economic factors of marginal workers in the Tamil Nadu region of India. This documentation provides an overview of the project, its goals, and the analysis process.

**2. Project Goals:**

**The primary goals of this project are as follows:**

* To analyze the age distribution of marginal workers in Tamil Nadu.
* To examine the distribution of marginal workers based on their industrial category and gender.
* To identify any correlations or patterns within the dataset.
* To build a predictive model for classifying marginal workers.

**3. Data Used:**

* The project uses a dataset sourced from [Provide the data source or origin], which contains information on marginal workers in Tamil Nadu. The dataset includes information on age, gender, education, income, and industrial categories.

**4. Data Preprocessing and Cleaning:**

**Data preprocessing was carried out to ensure the dataset's quality and consistency. The following steps were taken:**

* **Handling missing values by imputing with zeros.**
* **Removal of duplicate records.**
* **Outlier handling, using the winsorization technique.**

**5. Data Analysis:**

* Data analysis included the calculation of descriptive statistics and the creation of cross-tabulations. The analysis explored the dataset's characteristics and provided an initial understanding of the marginal workers' attributes.

**6. Data Visualization:**

* Data visualization was performed to illustrate key insights from the dataset. This included the creation of various plots, such as histograms, count plots, and scatterplots, to visually represent the age distribution, industrial categories, and other relevant attributes of marginal workers.

**7. Feature Engineering:**

* Feature engineering involved the creation of new features and data transformations. Notably, one-hot encoding of the 'Education Level' variable was performed to prepare the data for modeling.

**8. Model Training and Evaluation:**

* A Random Forest Classifier was selected as the machine learning model. The data was split into training and testing sets, the model was trained, and its performance was evaluated using accuracy and a classification report.

**9. Findings and Insights:**

* **The analysis and modeling revealed several key findings, including:**
* List the significant findings and insights you discovered during the analysis

**1. Age Distribution:**

* The age distribution of marginal workers in Tamil Nadu is skewed towards younger age groups, indicating that a significant portion of marginal workers are in the younger population.

**2. Industrial Categories:**

* The dataset reveals that a substantial number of marginal workers are engaged in agricultural and cultivation-related occupations.
* There is a noticeable difference in the distribution of marginal workers across different industrial categories based on gender.

**3. Income Distribution:**

* The income distribution among marginal workers in Tamil Nadu varies widely, with a concentration in lower-income brackets.
* There might be disparities in income levels based on factors such as education and industrial category.

**4. Educational Levels:**

* The educational level of marginal workers varies, with a portion having no formal education or very basic education.
* There may be differences in the educational levels of marginal workers based on gender or industrial category.

**5. Correlations and Patterns:**

* Analyzing correlations between different attributes might reveal interesting patterns. For example, you might find that younger workers tend to have lower incomes, or that education levels impact the choice of industrial category.

**6. Model Performance:**

* If you built a predictive model to classify marginal workers, you would evaluate the model's performance in terms of accuracy and classification metrics. Insights into how well the model predicts the target variable are valuable.

**7. Geographical Patterns:**

* If available in the dataset, you could explore geographical patterns, such as the distribution of marginal workers in different regions of Tamil Nadu.

**8. Gender Disparities:**

* The data might show gender disparities in terms of the types of work and income levels among marginal workers.
* Remember that these findings and insights are just examples. Your analysis should be based on the actual data you have and your specific research objectives. It's essential to thoroughly explore the dataset to derive meaningful insights that can contribute to your project's goals and objectives.

**10. Conclusion:**

* The "Marginal Workers in Tamil Nadu" project provided valuable insights into the characteristics and distribution of marginal workers in the Tamil Nadu region. The documentation summarizes the project's goals, the dataset used, data analysis, visualization, feature engineering, model training, and notable findings. This information can be used for further research and policymaking related to marginalized labor in Tamil Nadu.

**11.Reference:**

**Program:**

**import pandas as pd**

**import matplotlib.pyplot as plt**

**import seaborn as sns**

**from sklearn.model\_selection import train\_test\_split**

**from sklearn.ensemble import RandomForestClassifier**

**from sklearn.metrics import accuracy\_score, classification\_report**

**# Step 1: Load the data from the CSV file**

**data = pd.read\_csv("C:/Users/balur/Downloads/marginalworkers.csv")**

**# Step 2: Data Cleaning and Preprocessing**

**data.fillna(0, inplace=True)**

**data.drop\_duplicates(inplace=True)**

**# Step 3: Data Analysis**

**age\_stats = data['Age group'].describe()**

**cross\_tab = pd.crosstab(data['Industrial Category - A - Cultivators - Persons'], data['Area Name'])**

**# Step 4: Data Visualization**

**plt.figure(figsize=(12, 6))**

**sns.histplot(data['Age group'], bins=20, kde=True)**

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

**plt.xlabel('Age group')**

**plt.ylabel('Frequency')**

**plt.show()**

**plt.figure(figsize=(10, 6))**

**sns.countplot(data=data, x='Industrial Category - A - Cultivators - Persons', hue='Sex')**

**plt.title('Distribution of Marginal Workers by Industrial Category - A - Cultivators - Persons and Sex')**

**plt.xlabel('Industrial Category - A - Cultivators - Persons')**

**plt.ylabel('Count')**

**plt.xticks(rotation=45)**

**plt.legend(title='Area Name')**

**plt.show()**

**# Step 5: Feature Engineering**

**# For example, one-hot encode a categorical variable**

**data = pd.get\_dummies(data, columns=['Education Level - Persons'])**

**# Step 6: Model Training and Evaluation**

**X = data.drop(columns=['TargetColumn']) # Replace 'TargetColumn' with the actual target variable**

**y = data['TargetColumn']**

**X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)**

**model = RandomForestClassifier(n\_estimators=100, random\_state=42)**

**model.fit(X\_train, y\_train)**

**y\_pred = model.predict(X\_test)**

**accuracy = accuracy\_score(y\_test, y\_pred)**

**print("Accuracy:", accuracy)**

**report = classification\_report(y\_test, y\_pred)**

**print("Classification Report:")**

**print(report)**

**output:**

