**PROJECT TITLE: Serverless IOT data processing.**

**PHASE 3: Development part 1**

**1. Data Collection:**

* Determine the dataset you need for your project.
* Download or gather the dataset. This could be in the form of CSV files, text data, images, or any other relevant format.

**2. Data Loading:**

* Depending on the format of your dataset, you'll need to use appropriate libraries or methods to load the data into your programming environment. Common libraries for data loading include **pandas** for CSV files, **numpy** for numerical data, an  
  d various image libraries for image datasets.

**3. Data Exploration:**

* Take a look at the dataset to get a better understanding of its structure and content.
* Examine the first few rows of the dataset to see what it looks like.
* Check for any missing values or inconsistencies in the data.

**4. Data Preprocessing:**

* Handle missing data, either by imputing values or removing rows with missing data, depending on the nature of your dataset.
* Encode categorical variables if necessary (e.g., one-hot encoding).
* Normalize or scale numerical features if needed.
* Perform any other data transformations that are specific to your project's requirements.

**5. Data Splitting:**

* Divide your dataset into training, validation, and test sets (or any other splits you require for your project).
* Ensure that these sets are representative of the data distribution to avoid bias in your model.

**6. Save the Preprocessed Data:**

* Once you've completed the preprocessing, save the preprocessed dataset if necessary. This can be done using libraries like **pandas** (for DataFrames) or by saving arrays in various formats (e.g., NumPy's **.npy** files).

**7. Documentation:**

* Make sure to document the steps you've taken for data preprocessing. This will be helpful for tracking your progress and sharing your work with others.

**8. Version Control:**

* If you're using a version control system like Git, commit your changes at this stage to keep a record of your wor

**Step 1: Loading and Preprocessing the Dataset**

1. **Choose and Acquire the Dataset:**
   * Determine the dataset you need for your project.
   * Download or gather the dataset in a format compatible with your programming environment.
2. **Load the Dataset:**
   * Depending on the format of your dataset, use the appropriate libraries or methods to load it. For example, if you're working with CSV files, you can use the **pandas** library in Python.

Example (loading a CSV file with pandas):

python

import pandas as pd data = pd.read\_csv('your\_dataset.csv')

1. **Data Exploration and Preprocessing:**
   * Explore the dataset to understand its structure and content.
   * Handle missing data, encode categorical variables, normalize numerical features, and perform any other necessary preprocessing steps.

rows with missing values

1. **Data Splitting:**
   * Divide the dataset into training, validation, and test sets.

**data = data.dropna() # Removes rows with missing valuesStep 2: Saving Your Code on GitHub**

1. **Create a GitHub Repository:**
   * Go to GitHub and create a new repository for your project.
2. **Push Your Code:**
   * Initialize a Git repository in your project folder and push your code to GitHub.

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

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

**Step 3: Project Submission**

1. **Go to the Project Submission Part 3 Section:**
   * Navigate to the section where you need to provide the link to your GitHub repository.
2. **Upload the File:**
   * Upload any required files, which may include the preprocessed dataset or other project-related files.
3. **Provide GitHub Link:**
   * Copy and paste the link to your GitHub repository in the provided space.
4. **Click on Submit:**
   * After completing all the necessary steps, click on the submit button.

Remember to replace placeholders like **<your\_dataset.csv>** and **<your\_github\_repository\_url>** with actual file names or URLs.

1. **Load Dataset**:
   * At the beginning of Phase 3, you start by loading your dataset into your programming environment. This could be in the form of CSV files, text data, images, or any other relevant format.
2. **Data Exploration**:
   * Once the data is loaded, you explore it to understand its structure, features, and content. This step helps you get a better grasp of what you're working with.
3. **Data Preprocessing**:
   * In this step, you perform various operations on the data to prepare it for modeling. This includes handling missing data, encoding categorical variables, normalizing numerical features, and any other necessary transformations.
4. **Data Splitting**:
   * After preprocessing, you split the data into different subsets. Common splits include training, validation, and test sets. This ensures that you have separate datasets for training, tuning hyperparameters, and evaluating your model.
5. **End of Phase 3**:
   * This marks the completion of Phase 3, where your dataset is now loaded, explored, and preprocessed, ready for the next steps in your project.

This diagram provides a high-level overview of the process. Depending on the specifics of your project and the nature of your dataset, there may be additional steps or variations in the process.