# **Replication Package README**

## **Overview**

This replication package provides all necessary files and scripts to reproduce the results and analyses presented in the paper **"Into the ML-universe: An Improved Classification and Characterization of Machine-Learning Projects"**. The package is structured into multiple directories, each containing relevant datasets, scripts, and documentation.

## **Contents**

### **1. Dataset**

This directory contains the datasets used in the study:

* Final\_Dataset.csv: The final dataset used for analysis.
* Baseline.xlsx: Baseline data in Excel format.
* Baseline.csv: Baseline data in CSV format.
* NicheValidation.xlsx: Data used for classifying the **NICHE dataset** into ML-Libraries & Toolkits and ML-Applied categories.
* Filtering Dataset/: Contains preprocessing scripts and additional filtering steps applied to the dataset.

### **2. Pipeline**

Contains the detailed process used in the study:

* Pipeline.xlsx: Detailed breakdown of the pipeline methodology.

### **3. Automatic Classification of ML Projects**

This directory includes scripts and modules used for the automatic classification of machine learning projects.

* Refer to README.md inside this directory for installation and usage instructions.

### **4. Research Questions (RQs)**

This section is divided into directories corresponding to different research questions (RQs) explored in the study.

#### **RQ1 - Validation of ML Project Classification**

* validation/: Contains validation data for RQ1.
* selected\_projects.csv: The selected sample for validation.
* RQ1.xlsx: Contains data on the evaluation process of the tool’s detection performance.

#### **RQ2 - Evaluation of Metrics in ML Projects**

* final\_validated\_repos/: The set of projects deemed eligible for the study.
* metrics/: Contains metric files related to RQ2.
* test\_results/: Includes results from the **Friedman and Nemenyi tests** applied to assess classification performance.

### **5. MARK Approach**

The MARK approach is divided into two key folders:

#### **MARK-Configuration Study**

* Allows the execution of **MARK** with the inclusion/exclusion of different components.
* Contains the **MLTaxonomy-AblationStudy** directory with scripts and configuration files.
* MLTaxonomy-AblationStudy/Categorizer/: Source code and modules related to the categorization process.
* MLTaxonomy-AblationStudy/repos/: Contains repository-related analysis and classification output.
* MLTaxonomy-AblationStudy/cloner/: Scripts for repository cloning and preprocessing.

#### **MARK-Tool**

* Contains the original **MARK** version and detailed execution guidelines.
* A specific README.md is provided inside this folder with detailed instructions for setup and usage.

### **6. Execution Overview**

* Provides an overview of the experimental setup and execution steps followed in the study.
* ExecutionOverview/: Contains logs, step-by-step execution details, and supplementary documentation.

### **7. Libraries and API Methods**

* Libraries\_API\_Methods/: Contains relevant libraries and API methods used in the classification and evaluation process.

## **Reproducibility**

To reproduce the results, follow the instructions provided in the README.md files within each relevant directory. These instructions include:

* Required dependencies and installation steps.
* Data preprocessing and execution steps.
* How to interpret the generated results.

For further inquiries, please refer to the paper or contact the authors.