Data Preprocessing Documentation

# Introduction

Data preprocessing is a crucial step in the machine learning pipeline. It involves cleaning, transforming, and preparing raw data to make it suitable for analysis and model training. This document outlines the data preprocessing steps performed for the MLOps task.

## Data Source

- Source: Dawn and BBC Websites

- Description: Data is in raw form I have clean the data and maintain a CSV file.

## Data Preprocessing Steps

### 1. Data Loading:

- Loaded the raw data from the source into memory.

- Used pandas library in Python to read CSV files.

## 2. Data Cleaning:

- Handled missing values:

- Imputed missing numerical values using mean or median.

- Imputed missing categorical values using the mode.

- Removed duplicate records.

## 3. Feature Engineering:

- Extracted date/time features from timestamp columns.

- Created new features based on domain knowledge or insights.

## 4. Data Transformation:

- Standardized numerical features using z-score normalization.

- Encoded categorical features using one-hot encoding or label encoding.

- Applied log transformation to skewed numerical features.

## 5. Data Splitting:

- Split the preprocessed data into training and testing datasets.

- Used an appropriate split ratio, e.g., 80% training and 20% testing.

# DVC Setup Documentation

## Introduction

Data Version Control (DVC) is used to track changes to the data files and metadata. It ensures reproducibility and versioning of the data pipeline.

### DVC Configuration

## 1. Installation:

- Installed DVC using pip:

pip install dvc

## 2. Initialization:

- Initialized DVC in the project directory:

dvc init

## 3. Remote Storage Setup:

- Configured Google Drive as a remote storage for DVC:

dvc remote add -d myremote gdrive://<folder\_id>

- Replace `<folder\_id>` with the ID of the Google Drive folder.

## 4. Data Tracking:

- Tracked the preprocessed data files using DVC:

dvc add data/preprocessed\_data.csv

## 5. Commit and Push:

- Committed the changes to DVC:

dvc commit

- Pushed the data to the remote storage:

dvc push