Report on Preprocessing the Wildfire History Dataset

"2.3 Million US Wildfires (1992-2020)" Dataset

Overview

This report outlines the preprocessing steps for the "2.3 Million US Wildfires (1992-2020)" dataset, which contains a comprehensive record of wildfire occurrences across the United States from 1992 to 2020. The dataset supports analyses related to fire management and risk assessment.

Data Description

The dataset includes detailed records of wildfires sourced from federal, state, and local fire reporting systems. Key data points include the date of discovery, fire size, location, cause of the wildfire, and various administrative identifiers.

Preprocessing Steps

Step 1: Import Libraries

The preprocessing begins with the importation of essential Python libraries for data manipulation and numerical operations.

Step 2: Load the Dataset

The dataset is read into a DataFrame from a CSV file, facilitating easy access and manipulation of the data.

Step 3: Standardize Mixed-Type Columns

To prevent inconsistencies, columns identified as having mixed data types are standardized to string format. This uniformity is crucial for avoiding type-related errors in subsequent processing.

Step 4: Remove Irrelevant Columns

Columns that are redundant or irrelevant to the analysis goals are removed. This includes various local identifiers and reporting details that do not contribute significantly to wildfire risk prediction. Removing these columns helps reduce the complexity and size of the dataset.

Step 5: Handle Date Columns

Date columns are converted into a consistent datetime format. This conversion facilitates easier manipulation of dates and times for calculations and time-series analysis.

Step 6: Calculate Fire Duration

The duration of each fire, from discovery to containment, is calculated. This metric is important for understanding the severity and management effectiveness of each wildfire incident.

Step 7: Encode Categorical Variables

Categorical variables are transformed into a format suitable for statistical modeling, expanding the dataset to include a column for each category, thereby enhancing the dataset's utility for machine learning models.

Step 8: Clean Numeric Columns

Numeric columns are inspected and cleaned to ensure they contain valid numerical data. Any non-numeric entries are converted or removed to maintain the dataset's integrity.

Step 9: Impute Missing Values

Missing values in numeric columns are filled with the median value of each column. This approach helps preserve data integrity without introducing significant bias.

Step 10: Save Processed Data

The processed data is saved to a new CSV file, ensuring that the cleaned and optimized dataset is readily available for further analysis and modeling.