Manuel Sánchez Mc Sweeney
Machine Learning Master's Project

Water Pipelines in Tanzania

This Machine Learning project made for my class of Machine Learning in my master's degree in Big Data, Data Science, and Business Analytics, focuses on predicting the operational status of water pipelines (taps) in Tanzania. The main goal is to determine whether a water point is functional, non-functional, or in need of repair, which is crucial for managing water resources and ensuring access to clean water in different regions.

Here's a general overview of the project's workflow:

- **Data Used:** The project utilizes several datasets containing detailed information about water points in Tanzania. Key features include:
    - Geographical data (longitude, latitude, GPS height, basin, region, district, sub-village, ward).
    - Administrative details (funder, installer, scheme management, scheme name).
    - Operational characteristics (amount of water, date recorded, construction year, water quality, quantity, extraction type, management type, payment details, waterpoint type, population served).
    - The target variable, *status_group*, indicates whether a water point is 'functional', 'non functional', or 'functional needs repair'.
- **Data Preprocessing** Before training the model, the raw data underwent several cleaning and preparation steps:
    - **Handling Missing Values**: Empty values in various categorical columns (e.g., 'funder', 'installer', 'scheme_name') were replaced with 'Unknown' to ensure completeness.
    - **Date Feature Engineering**: The *date_recorded* column was converted into a *datetime* format, and new numerical features like *month_recorded* and *day_recorded* were extracted. The *construction_year* column had zero values replaced with the median *construction year*, and a new *age_in_years* feature was calculated to represent the age of each water point.
    - **Feature Selection**: Columns with too many unique values or those considered less relevant, specifically '*recorded_by*', '*wpt_name'*, and '*subvillage'*, were removed to streamline the dataset.
    - **Categorical Encoding**: All remaining categorical features were converted into a numerical format using one-hot encoding, a necessary step for machine learning models to process the data.
- **Machine Learning Model** A *RandomForestClassifier* model was chosen for this predictive task. This ensemble learning method is well-suited for classification problems and can handle a large number of features.

- **Model Performance** The model was trained on a portion of the processed data and validated on a separate set. It achieved an accuracy of approximately 80.60% on the validation set, indicating that it can predict the operational status of water points with good reliability.
- **Key Insights** Based on the analysis, the project highlighted regional differences in water point functionality:
    - Lindi region was identified as having the highest proportion of non-functional water points.
    - Kinga region showed the highest proportion of functional water points.
    - Kigoma region indicated the greatest need for water point repairs.

This project provides a valuable tool for understanding and addressing the challenges related to water infrastructure in Tanzania.

You can see the code of the project on my GitHub profile.