Cleanest City Prediction ML Model

Team Name: Team Gradient

Team Members:

1. Vardhan Yadav, SRMIST

2. Apoorv Khanna, SRMIST

3. Atharva Singh, SRMIST

4. Chaitanya Agarwal, MUJ

Team Mentor: Dr. C. Muralidharan (Assistant Professor, SRMIST)

Problem Statement:

Urban cleanliness is a significant determinant of public health and quality of life, particularly in rapidly urbanizing countries like India. Despite ongoing efforts to improve sanitation and waste management, many cities continue to struggle with cleanliness issues, impacting residents' well-being and overall city livability. This project aims to develop a machine learning (ML) model that predicts the cleanliness score of Indian cities for the upcoming year based on historical data obtained from Kaggle.

The primary challenge lies in accurately forecasting cleanliness scores by analyzing various influencing factors, such as waste management practices, pollution levels, infrastructure quality, and citizen satisfaction. The dataset will include diverse features related to urban cleanliness, collected from multiple sources to ensure comprehensive coverage of the factors affecting cleanliness.

By employing advanced ML algorithms, this model seeks to identify patterns and relationships within the data that can inform city planners and policymakers about potential future cleanliness outcomes. The goal is to provide actionable insights that enable proactive measures for improving urban sanitation and enhancing the living conditions in Indian cities. Ultimately, this research will contribute to the broader understanding of urban cleanliness dynamics and support sustainable urban development initiatives in India.

Solution Description:

The code implements a web application that predicts the cleanliness scores of Indian cities for the year 2024 using a machine learning model. The application is built with Flask for the web framework and Gradio for creating an interactive user interface. Below is a detailed breakdown of the solution:

1. Data Loading and Preprocessing

The application begins by loading a dataset containing cleanliness scores of various Indian cities from a CSV file. This dataset includes historical scores from previous years.

The features used for prediction are defined as columns representing cleanliness scores from 2016 to 2023.

A target variable, 2024_Score_Predicted, is calculated based on the previous year's score and a simple linear extrapolation method.

2. Model Training

The dataset is split into training and testing sets using train_test_split to evaluate the model's performance effectively.

A StandardScaler is employed to standardize the feature values, ensuring that they contribute equally to the model training.

A GradientBoostingRegressor model is instantiated and trained on the scaled training data.

The R-squared metric is calculated to assess the model's performance on the test set, providing insights into how well the model can explain the variance in cleanliness scores.

3. City Data Retrieval

A function, get_city_data, retrieves data for a specific city based on user input. If the city is not found in the dataset, it returns None.

4. Prediction Functionality

The predict function serves as the core of the prediction process. It checks if the selected city exists in the dataset and retrieves its corresponding cleanliness scores.

The scores are scaled using the previously fitted scaler, and then fed into the trained model to predict the cleanliness score for 2024.

Additionally, it calculates and returns the overall accuracy (R^2) of the model based on predictions across all cities.

5. User Interface with Gradio

A Gradio interface is created to allow users to select a city from a dropdown menu and view its predicted cleanliness score along with model accuracy.

The Gradio interface runs in a separate thread, enabling simultaneous access to both Gradio and Flask functionalities.

6. Flask Web Application

The Flask app provides an endpoint (/) that renders an HTML template (index_gradio.html). This template displays prediction results when users submit their city selection via an HTML form.

Upon form submission, it processes user input, fetches relevant city data, performs predictions, and renders results back to the user.

Technical Architecture:

The technical architecture of the provided code outlines a web application designed for predicting the cleanliness scores of Indian cities for the upcoming year using machine learning. The architecture integrates Flask for web services and Gradio for creating an interactive user interface. Below is a detailed breakdown of the components and their interactions:

1. Frameworks and Libraries

Flask: A lightweight web framework used to handle HTTP requests, serve web pages, and manage routing.

Gradio: A library that simplifies the creation of user interfaces for machine learning models, allowing users to interact with the model through a web interface.

Scikit-learn: A machine learning library that provides tools for data preprocessing, model training, and evaluation metrics.

Pandas: A data manipulation library used to load and process the dataset.

2. Data Pipeline

Data Loading: The application begins by loading a CSV file containing cleanliness scores for various cities in India. This data is processed using Pandas to create a DataFrame.

Feature Engineering: The relevant features (historical cleanliness scores) are extracted from the DataFrame, and a target variable (`2024_Score_Predicted`) is calculated based on historical trends.

3. Machine Learning Model

Data Preprocessing: The features are standardized using `StandardScaler` to ensure uniformity in the input data, which is crucial for model performance.

Model Training: A `GradientBoostingRegressor` is trained on the scaled features to predict future cleanliness scores. The model's performance is evaluated using R-squared metrics to determine its accuracy.

4. Prediction Logic

City Data Retrieval: A function (`get_city_data`) fetches data specific to a selected city from the dataset. If the city does not exist, it returns an appropriate message.

Prediction Functionality: The `predict` function handles user input from the Gradio interface, scales the city's historical scores, and uses the trained model to predict the cleanliness score for 2024. It also calculates and returns the overall accuracy of the model.

5.User Interface

Gradio Interface Creation: A Gradio interface is created that allows users to select a city from a dropdown menu. Upon selection, it displays the predicted cleanliness score and model accuracy.

Flask Web Page Rendering: The Flask application serves an HTML page (`index_gradio.html`) that allows users to submit their city selection via a form. Upon submission, it processes the input and displays prediction results.

6.Concurrency Management

Threading: To enable simultaneous access to both Gradio and Flask functionalities, the Gradio interface is launched in a separate thread. This allows users to interact with the prediction interface while also accessing other features of the Flask application.

7. Deployment Considerations

The application is designed to run locally with debugging enabled (`debug=True`), making it suitable for development and testing purposes. For production deployment, considerations such as security, scalability, and performance optimization need to be addressed.

The development of a machine learning (ML) model to predict the cleanliness index of Indian cities using a Kaggle dataset has significant social, economic, and ecological impacts. This model not only serves as a predictive tool but also influences various aspects of urban living and governance. Below are the detailed impacts categorized into social, economic, and ecological dimensions.

Social Impact

1. Enhanced Public Awareness:

The model provides citizens with insights into their city's cleanliness score, fostering greater awareness and engagement in local sanitation efforts. As residents become more informed about cleanliness levels, they may be motivated to participate in community clean-up initiatives or advocate for better waste management practices.

2. Improved Quality of Life:

By predicting cleanliness scores, the model helps identify cities that may require urgent attention to sanitation issues. This proactive approach can lead to improvements in public health, as cleaner environments reduce the risk of disease transmission and enhance overall well-being.

3. Empowerment of Local Governance:

Local authorities can use the predictions to prioritize resources and interventions in areas with lower cleanliness scores. This data-driven approach empowers governments to make informed decisions that directly impact community health and safety.

4. Community Engagement:

The interactive nature of the Gradio interface allows users to select their city and view predicted scores, promoting transparency and encouraging community dialogue about cleanliness standards and local governance.

Economic Impact

1. Resource Allocation:

The model aids in optimizing resource allocation for waste management and sanitation services. By predicting which cities are likely to have lower cleanliness scores, municipalities can allocate funds more efficiently, ensuring that resources are directed where they are most needed.

2. Attracting Investments:

Cities with higher cleanliness scores may attract more tourism and investment opportunities. Clean urban environments are often perceived as more desirable for living and business operations, potentially boosting local economies.

3. Cost Savings:

Predictive analytics can lead to cost savings by preventing sanitation crises before they occur. Early interventions based on predicted scores can mitigate the need for expensive emergency clean-up operations or health care costs associated with pollution-related illnesses.

4. Job Creation:

As cities strive to improve their cleanliness scores, there may be an increase in jobs related to waste management, environmental services, and urban planning. This can contribute positively to local employment rates.

Ecological Impact

1. Sustainable Urban Practices:

By highlighting areas where cleanliness is lacking, the model encourages cities to adopt more sustainable waste management practices. This could lead to reduced landfill use, improved recycling rates, and better overall environmental stewardship.

2. Pollution Reduction:

- A focus on improving cleanliness can lead to initiatives aimed at reducing pollution sources within urban areas. Cleaner cities often correlate with lower emissions from waste disposal processes and improved air quality.

3. Biodiversity Protection:

Cleaner urban environments contribute to healthier ecosystems within city limits. Reducing litter and pollution can protect local flora and fauna, promoting biodiversity even in densely populated areas.

4. Climate Resilience:

The insights gained from the model can help cities become more resilient to climate change impacts by promoting practices that enhance urban green spaces and reduce heat islands through better waste management strategies.

Intel oneAPI DataAnalytics Library on the the dataset:

Intel oneAPI Data Analytics Library is an open-source, cross-platform library designed to optimize data analytics and machine learning processes. It provides highly optimized algorithmic building blocks that enhance performance across various stages of the data analytics pipeline.

To implement oneDAL on the Kaggle dataset for clean city predictions:

1. **Dataset Preparation**:
    - Load and preprocess the dataset using oneDAL's preprocessing functions to ensure data quality.
2. **Model Training**:
    - Utilize the gradient-boosting regression capabilities provided by oneDAL to train the model on the prepared dataset.
3. **Performance Evaluation**:
    - Validate the model's performance using oneDAL's validation tools to assess accuracy and make necessary adjustments.
4. **Deployment**:

- Deploy the trained model for real-time predictions or further analysis using oneDAL's deployment options.

By integrating Intel oneAPI Data Analytics Library into the machine learning workflow, organizations can enhance their predictive analytics capabilities while ensuring efficient use of computational resources. This approach not only improves prediction accuracy but also accelerates the overall data analysis process, making it a valuable tool in urban planning and management initiatives.

Intel oneAPI Deep Neural Network Library on the Dataset:

Intel oneAPI Deep Neural Network Library (oneDNN) is an open-source, cross-platform library designed to optimize deep learning applications. It provides highly optimized implementations of various deep learning building blocks, allowing developers to leverage both CPU and GPU architectures seamlessly. This library is particularly beneficial for enhancing the performance of existing frameworks such as TensorFlow and PyTorch, making it suitable for diverse machine learning tasks.

To implement oneDNN on the Kaggle dataset for clean city predictions:

Dataset Preparation:

Load and preprocess the dataset using oneDNN's optimized functions to ensure data quality and readiness for model training.

Model Training:

Adapt the gradient-boosting regression model to utilize oneDNN's optimized primitives, which can significantly enhance training speed and accuracy.

Performance Evaluation:

Validate the model's performance using built-in evaluation tools within oneDNN to ensure that it meets the desired accuracy metrics.

Deployment:

Deploy the trained model for real-time predictions or further analysis using oneDNN's deployment capabilities.

By leveraging Intel oneAPI Deep Neural Network Library in conjunction with gradient-boosting regression techniques, organizations can achieve superior

predictive analytics capabilities while optimizing resource usage across various hardware platforms. This integration not only enhances prediction accuracy but also accelerates the overall machine learning workflow, making it an invaluable asset in urban management initiatives focused on cleanliness and sustainability.

Conclusion

The creation of a machine learning model that predicts the cleanliness index of Indian cities using a Kaggle dataset has far-reaching implications across social, economic, and ecological dimensions. By fostering greater awareness among citizens, optimizing resource allocation for local governments, promoting sustainable practices, and ultimately improving public health outcomes, this initiative represents a significant step toward enhancing urban living conditions in India. The integration of data-driven decision-making processes will not only benefit current residents but also pave the way for future generations to enjoy cleaner, healthier cities.