Deforestation Issue Analysis Using Support Vector Machine (SVM)

Dataset

The dataset provided for this assignment is titled **deforestation_dataset.csv**. It contains data related to deforestation across multiple countries for different years. The key variables in the dataset include tree cover loss, CO2 emissions, rainfall, population, GDP, and several other factors contributing to deforestation.

Phase 1 - Data Preprocessing

1. **Loading the Data**
    - Load the dataset **(deforestation_dataset.csv)** into your programming environment.
    - Inspect the dataset for any missing values, inconsistent data, or any other anomalies.
2. **Data Cleaning**
    - Handle missing values (e.g., impute, remove, or replace them as necessary).
    - Convert categorical columns (such as `Deforestation_Policy_Strictness`, `Corruption_Index`) into numerical values using encoding methods (e.g., label encoding or one-hot encoding).
3. **Feature Scaling**
    - Normalize or standardize numerical features such as `CO2_Emission_mt`, `Population`, `GDP_Billion_USD`, etc., to make sure the SVM model works effectively.
    - Determine the importance of each feature in the model and decide if any feature selection is needed.
4. **Split Data into Training and Testing Sets**
    - Split the dataset into training (80%) and testing (20%) sets.
    - Use the `train_test_split` function from a relevant library to make sure the data is shuffled and the split is random.

Phase 2 - Model Building and Evaluation

1. **Train the Support Vector Machine (SVM) Model**
    - Train an SVM model to predict deforestation levels. The target variable can be either `Forest_Loss_Area_km2` or `Tree_Cover_Loss_percent`, depending on your analysis.
    - Implement a linear kernel for the initial model, and evaluate its performance on the training data.
2. **Model Evaluation**
    - Evaluate the model's performance using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared.
    - Visualize the SVM decision boundaries (if applicable) or evaluate feature importance using the trained model.
3. **Hyperparameter Tuning**
    - Experiment with different SVM kernels (e.g., polynomial, RBF) to improve model performance.
    - Tune hyperparameters such as `C`, `gamma`, and kernel types using grid search or random search methods.
    - Apply cross-validation techniques to validate the model's robustness.

Phase 3 - Feature Analysis and Interpretation

1. **Analyze Feature Importance**
    - Investigate which features have the most significant impact on deforestation based on the model's output.
    - Generate a report on how each feature (e.g., `Rainfall_mm`, `GDP_Billion_USD`, `Population`) influences deforestation levels.
2. **Interpretation of Results**
    - Provide insights based on the SVM model about the key contributors to deforestation.
    - Discuss how factors like **GDP**, **policy strictness**, and **illegal lumbering incidents** influence deforestation in the countries studied.

Phase 4 - Reporting and Recommendations

1. **Visualize the Results**

    - Create visualizations (graphs, charts) that illustrate key relationships between the features and deforestation levels.
    - Include scatter plots, bar charts, or heatmaps to better understand the feature impact.
2. **Write a Comprehensive Report**

    - Provide a detailed analysis of the findings from your model, including:
        - How deforestation is influenced by variables like population, economic factors, and policy strictness.
        - Possible regions or countries where deforestation can be mitigated using specific interventions.
3. **Recommendations for Deforestation Mitigation**

    - Based on your findings, make recommendations on policies or actions that could help reduce deforestation, considering economic, social, and environmental aspects.

Final Deliverables

- A Jupyter notebook or Python script that includes the entire SVM model development process, from data preprocessing to final model evaluation.

- A **report** summarizing your analysis, model evaluation, and deforestation trends, supported by visualizations.

Evaluation Criteria

- **Data Preprocessing and Cleaning** Correct handling of missing values, appropriate encoding of categorical variables, and effective normalization of data.
- **Model Accuracy and Performance** The SVM model's effectiveness in predicting deforestation, demonstrated through evaluation metrics like MAE, MSE, RMSE, and R-squared.
- **Feature Engineering and Importance** Clear explanation of the most influential factors on deforestation.
- **Quality of Analysis and Interpretation** Depth of insights provided based on model outputs.
- **Visualizations and Reporting** Clarity and effectiveness of the visualizations, and quality of the written report.

Note The dataset used in this assignment is **deforestation_dataset.csv**, which contains data related to deforestation in various countries and years. Make sure to use the correct target variable and model evaluation techniques to derive actionable insights on the deforestation issue.s