BITS, PILANI – HYDERABAD CAMPUS
BITS F464 MACHINE LEARNING
FIRST SEMESTER 2025-2026

Assignment 1
Exploratory Data Analysis and Regression Modeling for Agricultural Greenhouse Gas Emissions

Dataset Description

The dataset is a structured agricultural emissions dataset developed to support climate analysis and decision-making in sustainable agriculture. It captures the relationship between agricultural practices, environmental conditions, and greenhouse gas (GHG) emissions across diverse farming scenarios. Each record in the dataset represents an agricultural unit observed over a cultivation period and includes detailed information on crop type, fertilizer usage, irrigation patterns, and climate variables. Crop-related attributes describe the type of crop cultivated, while fertilizer-related features quantify nutrient application levels. Irrigation data reflect both the method and intensity of water usage. Climate variables include key atmospheric conditions such as temperature, rainfall, and humidity, which influence emission behavior and crop performance.

The dataset provides emission indicators for major greenhouse gases, including carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O), along with aggregated CO_2-equivalent values. In addition to baseline observations, the dataset contains explicitly labeled counterfactual scenarios representing alternative agricultural and climatic conditions. These scenarios enable comparative analysis of emission outcomes under different management and environmental settings.

Designed with scalability and interoperability in mind, the dataset supports advanced analytical workflows and transparent evaluation of agricultural emission patterns. The structured format, combined with scenario labeling, makes the dataset suitable for benchmarking emission studies, sustainability assessments, and policy-oriented climate analysis in agriculture.

Link for dataset https//drive.google.com/file/d/19jYSig3v1j_drsjygkkwmEuQdrl7hDq1/view?usp=sharing

Learning Objectives

By completing this assignment, students will be able to

- Preprocess and explore real-world data using statistical analysis and visualizations.
- Build and evaluate regression models with different optimization and regularization techniques.
- Analyze the effects of model complexity on performance and generalization.
- Reformulate a regression problem as a classification task and interpret the results.

Tasks to be done

1. **Data Understanding and Preprocessing**
    a. Load and inspect the dataset structure (size, variables, data types).
    b. Identify numerical and categorical features.
    c. Define a dependent feature as the regression target.
    d. Justify the selection of numerical input features.
2. **Exploratory Data Analysis (EDA)**
    a. Compute descriptive statistics for numerical variables.
    b. Visualize feature distributions using histograms and boxplots.
    c. Analyze relationships between inputs and the target using scatter plots.
    d. Compute and visualize a correlation matrix.

e. Include additional plots wherever useful.
f. Summarize the key observations and report at least three data-driven insights.

3. **Linear Regression**
   a. Formulate emission prediction as a regression problem.
   b. Implement Linear Regression using
      i. Batch Gradient Descent
      ii. Stochastic Gradient Descent
   c. Split data into training and testing sets.
   d. Evaluate models using MAE, MSE, and R^2.
   e. Compare optimization behavior and results using metrics and plots.
4. **Polynomial Regression and Regularization**
   a. Apply polynomial features of degree two.
   b. Train polynomial regression models with
      i. L1 regularization
      ii. L2 regularization
   c. Evaluate and compare results with linear regression.
   d. Summarize the performance differences and the effects of model complexity.
   e. Visualize the results
5. **Classification Reformulation**
   a. Convert the regression task into a classification problem by defining a **clear, well-justified labeling strategy** (e.g., low/medium/high emissions).
   b. Apply a linear classifier
      i. Logistic Regression
      ii. Naive Bayes
      iii. Perceptron
   c. Present results and discuss the suitability and limitations of the classification approach.
   d. Visualize the results

Detailed Task Description

1. **Data Loading and Preprocessing**

   Load the dataset into Python using the Pandas library and display the first few rows to understand its structure. Identify the number of rows and columns and examine the data types of all variables. Check for missing values in each column and report your findings. If missing values are present, explain their meaning and handle them appropriately without removing a large portion of the data.

   Identify the numerical input columns to be used for modeling, such as fertilizer usage, irrigation water, temperature, rainfall, and humidity. **Clearly define a dependent feature as the regression target.**

2. **Exploratory Data Analysis (EDA)**

   Generate descriptive statistics for all numerical variables using pandas. Report common measures such as mean, median, minimum, maximum, and standard deviation, and briefly explain what these statistics reveal about the data. Create simple visualizations to support your analysis. Plot histograms and boxplots for numerical variables to study their distributions and identify possible outliers. Develop visualizations that highlight descriptive statistics for each numerical feature, such as a bar chart showing the minimum, maximum, mean, median, and standard deviation.

   Analyze relationships between key input variables and total greenhouse gas emissions using scatter plots. Compute a correlation matrix for numerical features and visualize it using a heatmap.

Summarize at least three important insights obtained from the exploratory analysis.

3. **Linear Regression**

   Formulate a regression problem using numerical agricultural and climate variables to predict total greenhouse gas emissions. Split the dataset into training and testing sets and train a linear regression model. Evaluate the model using Mean Absolute Error, Mean Squared Error, and R^2 score. Briefly interpret the results and explain which variables appear to have the strongest influence on emissions.

4. **Polynomial Regression**

   Extend the linear regression model by applying polynomial features of degree two to the numerical input variables. Train a polynomial regression model and evaluate it using the same metrics as before. Compare the performance of linear and polynomial regression and comment on whether the polynomial model improves prediction accuracy.

5. **Classification Reformulation**

   Convert the greenhouse gas emissions prediction problem into a classification task by selecting a discrete feature. Train and evaluate linear classifiers, including Logistic Regression, Naive Bayes, and Perceptron. Assess model performance using metrics such as Accuracy, Precision, Recall, F1-score, and Confusion Matrix. Briefly discuss the results, highlighting the strengths and limitations of the classification approach, including information loss from discretization and its suitability for categorical decision-making scenarios.

Submission Requirements

- Submit **one Jupyter Notebook per team**.
- The notebook must include
  - clean, readable code
  - labeled plots
  - concise explanations in Markdown cells
- The notebook must run end-to-end without errors.

Submission Format

- File name TeamXX_Assignment1.ipynb
- Include team details at the top of the notebook.
- Submit a single file named TeamXX_Assignment1.ipynb

Before Submission

- Execute the entire notebook before submission.
- Outputs should be visible for all plots and results.

Sample Jupyter notebook ML_Assignment1.ipynb

Timelines

- Final submission (Jupyter notebook on Google classroom) – March 1st, 2026 1159 PM
- Final Demos (in class) - Week beginning Mar 2nd, 2026

Evaluation Rubric (30 Marks)

1. Data Loading and Preprocessing **(4 Marks)**
2. Exploratory Data Analysis (EDA) **(4 Marks)**
3. Linear Regression **(4 Marks)**
4. Polynomial Regression **(4 Marks)**
5. Classification Reformulation **(4 Marks)**
6. Demo and Viva **(10 Marks - Individual Assessment)**

Academic Integrity

- This is a team-based assignment, and all submitted work must be the original effort of the team members.
- No external machine learning libraries (e.g., scikit-learn, TensorFlow, PyTorch) may be used; all algorithms must be implemented from scratch.
- Use of standard numerical and plotting libraries (e.g., NumPy, Pandas, Matplotlib) is permitted unless otherwise specified.
- Any external references (textbooks, papers, documentation) must be clearly cited.
- Plagiarism, code sharing between teams, or reuse of online implementations will result in disciplinary action as per the institute policy.

Contact Information

For any queries, please contact the course Teaching Assistants (TAs) via email communications through any other channels will not be considered. Please include all the listed TAs in your email for clarification.

- Shivram (p20230075@hyderabad.bits-pilani.ac.in)
- Harsha (p20200437@hyderabad.bits-pilani.ac.in)