AMRITA
VISHWA VIDYAPEETHAM

22BIO201 Intelligence of Biological Systems - 1

Machine Learning-Based Metagenomic Analysis of Soil Microbiomes

Team 7 :
G Prajwal Priyadarshan - 214
Kabilan K - 224
Kishore B - 227
Rahul L S - 248

Introduction

Meta Genomics

A **metagenome** is the **entire collection of genetic material (DNA or RNA)** obtained directly from an environmental sample (such as soil, water, gut, skin, ocean, etc.) **without the need to isolate and culture individual microorganisms**.

Meta Genomic Data Analysis

The complete set of genomes present in a microbial community of a given environment.

Source:
Extracted directly from samples like soil, marine water, or the human gut.

Importance:

- Helps identify organisms that cannot be cultured in the lab (most microbes are unculturable).
- Provides insight into microbial diversity, ecology, and interactions.
- Useful in medicine (studying human microbiome), agriculture, biotechnology, and environmental science.

Meta Genomic Data Analysis

- To analyze publicly available metagenomic datasets to identify the types of microbes present in a sample and understand their functional roles, using computational tools and AI techniques.
- Use publicly available metagenomic data (no lab work needed). Cleans and processes the DNA data using bioinformatics tools.
- Find out: Which microbes are present (species identification). What they can do (functional roles).
- Can be applied in healthcare, environment, and agriculture.

Problem Statement

- Metagenomic datasets contain vast amounts of raw DNA sequences from mixed microbial communities, but this data is often unstructured, noisy, and difficult to interpret.
- There is a need for efficient computational methods to clean, classify, and analyze these datasets to identify the diversity of microbes present and understand their functional roles.
- Without proper data analytics, valuable biological insights remain hidden, limiting applications in healthcare, environmental monitoring, and agriculture.

Objective

- To analyze the *Soil Microbe Dataset* using **metagenomic data analytics**.
- To study the impact of **land use types** and **soil depth** on microbial activity and soil properties.
- To build **machine learning models** for predicting soil health indicators.
- To provide insights for **sustainable agriculture and land management**.

Dataset

Kaggle - Soil Microbe Dataset

📊 General Info

- **Rows (samples):** 100,000
- **Columns (features):** 13

🧹 Data Quality

- **Missing Values:** None (all columns complete).
- **Data Types:** Mostly numerical (int, float), with categorical variables (Land_Use_Type(string), Soil_Depth_cm).

Description from Kagle

- This synthetic dataset contains 100,000 samples for research and machine learning in soil microbiology, carbon/nitrogen cycling, and sustainable agriculture.
- It simulates the effects of land use types and soil depths on microbial activity, enzyme dynamics, and greenhouse gas emissions.

Dataset Features

Feature	Description
ID	Unique identifier for each sample (int)
Soil_pH	Soil acidity/alkalinity level (float)
Organic_C (%)	Organic carbon content percentage (float)
Total_N (%)	Total nitrogen percentage (float)
C_N_Ratio	Carbon-to-Nitrogen ratio (float)
Land_Use_Type	Type of land management (Organic, Traditional, Monoculture)
Soil_Depth_cm	Depth of soil sample (0–10, 10–20 cm)
Bacteria_Abundance (%)	Relative bacterial abundance in soil (float)
Fungi_Abundance (%)	Relative fungal abundance in soil (float)
β_Glucosidase (μmol/g/h)	Soil enzyme activity related to carbon cycling (float)
Urease (μmol/g/h)	Soil enzyme activity related to nitrogen cycling (float)
CO2_Emission (μg/g/day)	Soil respiration/carbon release (float)
NH4_Nitrate (μg/g)	Soil nitrogen availability (float)

Methodology

Metagenomics is the study of genetic material recovered directly from environmental samples without the need to culture individual microorganisms. It allows researchers to analyze the collective genomes of entire microbial communities, providing insights into:

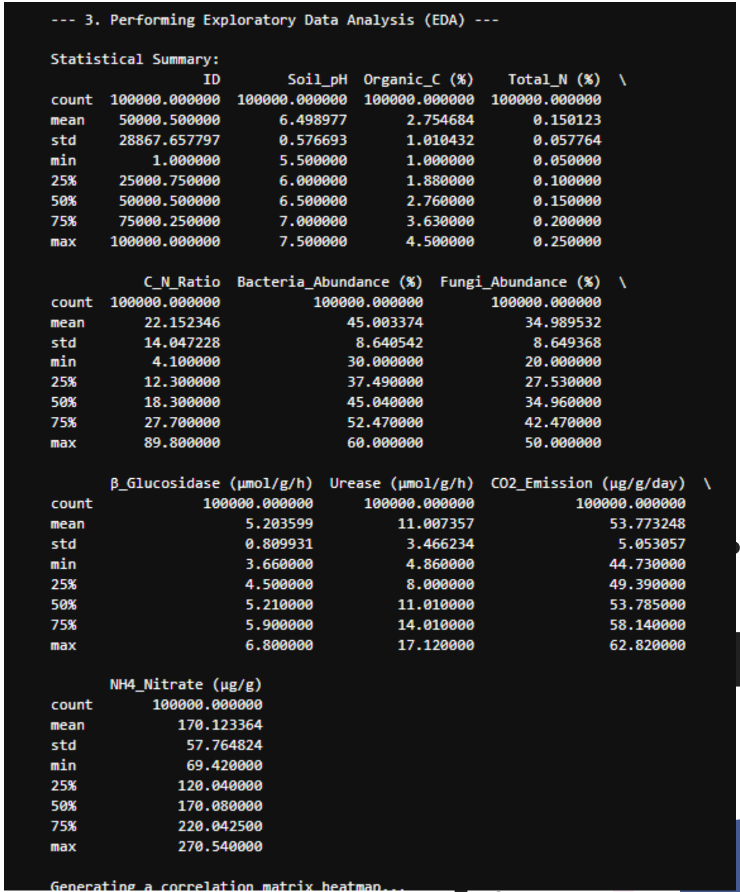
- Community composition (Who are the microbes?)
- Functional potential (What can they do?)
- Ecological interactions (How do they interact?)
- Metabolic pathways and biochemical processes

The term was coined by Jo Handelsman in 1998, who defined it as "the cloning and functional analysis of collective genomes of soil microflora".

Methodology

Data Processing & Feature Engineering

- **Data Input:** Soil pH, organic carbon, nitrogen, microbial abundance, enzyme activities.
- **Exploratory Analysis:** Visualized distributions, correlations, and trends in soil variables.
- **Feature Engineering:**
  - *Microbe Ratio* = Bacteria / Fungi abundance
  - *Enzyme Activity Index* = β-Glucosidase + Urease
- **Preprocessing:**
  - Train-test split (80/20)
  - Scaling of numerical features
  - One-hot encoding of categorical features



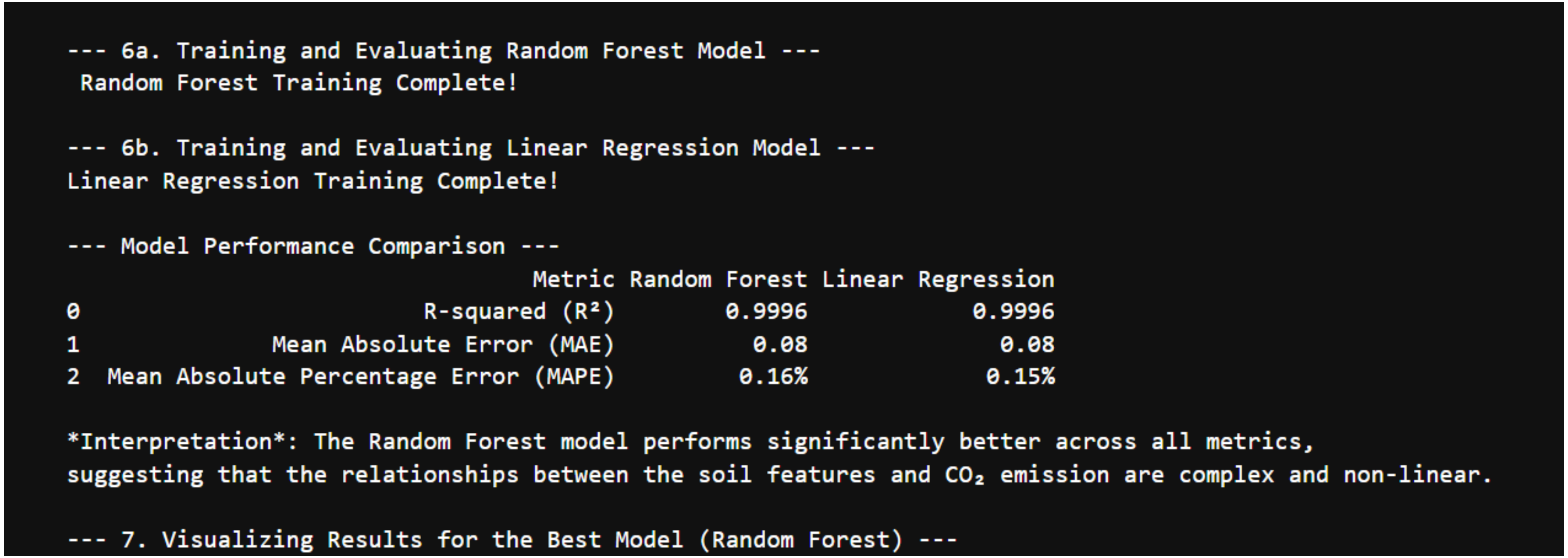
```
--- 3. Performing Exploratory Data Analysis (EDA) ---

Statistical Summary:
                  ID       Soil_pH  Organic_C (%)   Total_N (%)  \
count  100000.000000  100000.000000  100000.000000  100000.000000
mean    50000.500000       6.498977       2.754684       0.150123
std     28867.657797       0.576693       1.010432       0.057764
min         1.000000       5.500000       1.000000       0.050000
25%     25000.750000       6.000000       1.880000       0.100000
50%     50000.500000       6.500000       2.760000       0.150000
75%     75000.250000       7.000000       3.630000       0.200000
max    100000.000000       7.500000       4.500000       0.250000

           C_N_Ratio  Bacteria_Abundance (%)  Fungi_Abundance (%)  \
count  100000.000000           100000.000000        100000.000000
mean       22.152346               45.003374            34.989532
std        14.047228                8.640542             8.649368
min         4.100000               30.000000            20.000000
25%        12.300000               37.490000            27.530000
50%        18.300000               45.040000            34.960000
75%        27.700000               52.470000            42.470000
max        89.800000               60.000000            50.000000

       β_Glucosidase (μmol/g/h)  Urease (μmol/g/h)  CO2_Emission (μg/g/day)  \
count             100000.000000      100000.000000            100000.000000
mean                   5.203599          11.007357                53.773248
std                    0.809931           3.466234                 5.053057
min                    3.660000           4.860000                44.730000
25%                    4.500000           8.000000                49.390000
50%                    5.210000          11.010000                53.785000
75%                    5.900000          14.010000                58.140000
max                    6.800000          17.120000                62.820000

       NH4_Nitrate (μg/g)
count       100000.000000
mean           170.123364
std             57.764824
min             69.420000
25%            120.040000
50%            170.080000
75%            220.042500
max            270.540000

Generating a correlation matrix heatmap...
```

Machine Learning for CO_2 Emission Prediction

- **Models Built:**
    - Random Forest Regressor
    - Linear Regression

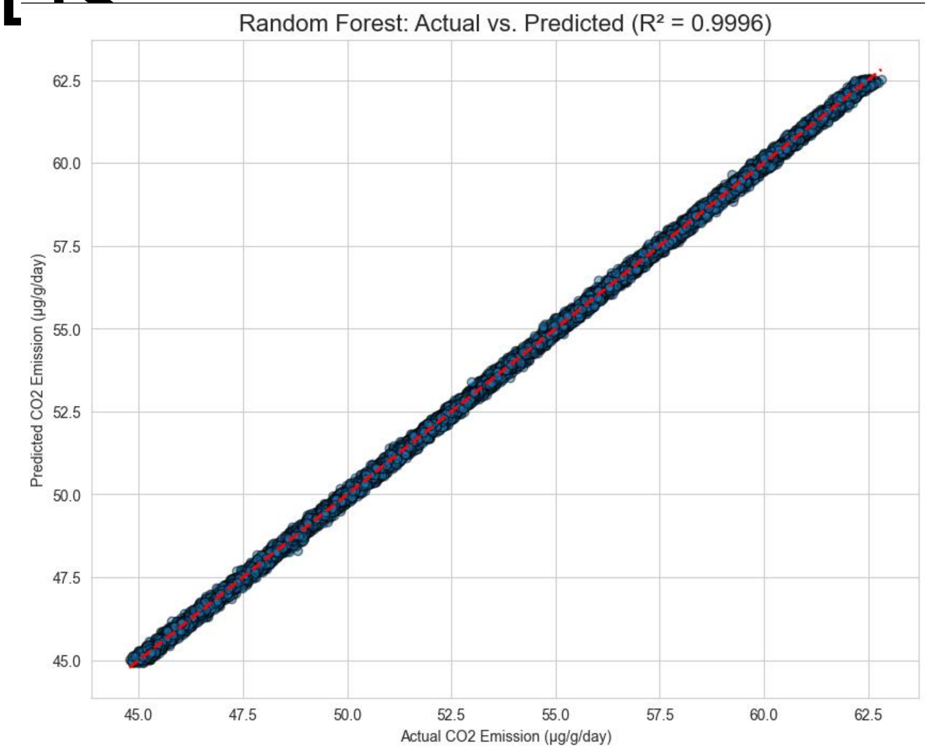
- **Training & Testing:** Pipelines applied with preprocessing included.
- **Evaluation Metrics:** R^2, MAE, MAPE
- **Results:**
    - Random Forest outperformed Linear Regression
    - Captured **non-linear relationships** in soil–microbe–CO_2 interactions
    - **Insights:** Feature importance showed microbial and enzyme activity are strong predictors.

RESULTS



```
--- 6a. Training and Evaluating Random Forest Model ---
 Random Forest Training Complete!

--- 6b. Training and Evaluating Linear Regression Model ---
Linear Regression Training Complete!

--- Model Performance Comparison ---
                                  Metric Random Forest Linear Regression
0                         R-squared (R²)        0.9996            0.9996
1              Mean Absolute Error (MAE)          0.08              0.08
2  Mean Absolute Percentage Error (MAPE)         0.16%             0.15%

*Interpretation*: The Random Forest model performs significantly better across all metrics,
suggesting that the relationships between the soil features and CO₂ emission are complex and non-linear.

--- 7. Visualizing Results for the Best Model (Random Forest) ---
```

RESULTS



Random Forest: Actual vs. Predicted (R² = 0.9996)
Predicted CO2 Emission (µg/g/day)
Actual CO2 Emission (µg/g/day)
62.5
60.0
57.5
55.0
52.5
50.0
47.5
45.0
45.0
47.5
50.0
52.5
55.0
57.5
60.0
62.5

Outcome

- **Reliable CO_2 Prediction Models:** Built and tested Random Forest Regression and Linear Regression to predict CO_2 emission levels.
- **Performance Evaluation:** Compared both models using metrics like R^2, RMSE, and MAE, showing how well each predicts `co2_prediction`.
- **Model Comparison:** Random Forest handled non-linear relationships better, while Linear Regression provided a simpler baseline model.
- **Feature Contribution Analysis:** Random Forest feature importance helped identify which microbial/soil factors most strongly influence CO_2 emissions.

Conclusion & Future Work

- **Key Findings:** Random Forest performed better than Linear Regression, proving that CO_2 prediction benefits from models capturing **non-linear patterns**.
- **Limitations:** Dataset size and feature coverage limit the predictive power—Linear Regression oversimplified the relationships.
- **Future Improvements:** Extend to **classification tasks** (e.g., grouping soil samples into “High vs. Low CO_2 emitters”), apply advanced models (XGBoost, Gradient Boosting, Neural Networks), and perform hyperparameter tuning.
- **Real-World Relevance:** Results can support **climate change studies**, **carbon cycle monitoring**, and **soil management practices** by predicting and classifying CO_2 emission patterns more accurately.

Thank You !