IJSRA

International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.
IJSRA
International Journal of Science and Research Archive
Research Journal Archive, INDIA

(Review Article)

Harnessing microbiomes and machine learning for sustainability efforts

Mahi Shah *

Research Workshop at Erik Jonsson School of Engineering and Computer Science, University of Texas, Dallas, TX, USA.

International Journal of Science and Research Archive, 2025, 14(02), 1476-1481

Publication history: Received on 20 December 2024; revised on 29 January 2025; accepted on 01 February 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.14.2.0311

Abstract

The Malthusian theory suggests that exponential population growth will eventually surpass the linear growth of food supply, leading to widespread land scarcity for urbanization. However, Malthus could not have anticipated the advancements in scientific knowledge and technology that now have the potential to revolutionize food production. With the global population projected to reach 9 billion by 2050 [4], ensuring sustainable food systems while minimizing environmental harm has become a critical challenge. Recent research reveals the role of the soil microbiome—complex communities of bacteria, fungi, and other microorganisms in enhancing plant growth. Notably, the correlation between soil bacteria and the growth of barley provides a method for identifying the most productive land.

This study tests the machine learning capability to analyze soil microbiome data to predict crop yields based on the role of bacteria. A dataset of 627 bacterial strain features from over 1,340 farms was analyzed using Python to build regression models, uncovering patterns that optimize smart farming practices and soil conservation [5]. By integrating artificial intelligence with microbiology, the research highlights innovative approaches to advancing sustainable agriculture by identifying the most suitable areas for crop cultivation. The paper demonstrates the effectiveness of KNN, Decision Tree, and MLP Regressor models in accurately predicting crop yields, outperforming random chances or any existing lottery-based systems.

Keywords: Sustainable Agriculture; Food Security; Machine Learning; Artificial Intelligence; Urbanization; Regressor Models

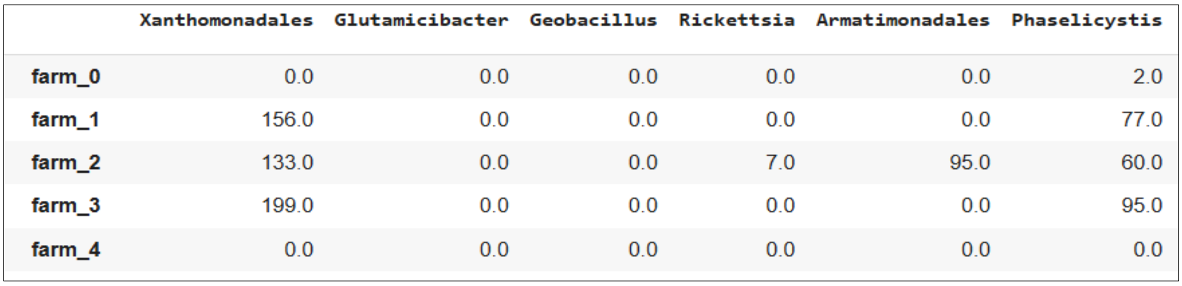
1. Introduction

Advances in microbiome research have revealed the critical role of bacteria in supporting ecosystems, including crop productivity. The 16S ribosomal RNA (rRNA) region, a unique genetic marker found in all bacteria, serves as a "barcode" for identifying bacterial species. Through 16S sequencing, researchers can decode the bacterial composition of a sample, linking specific bacterial communities to their ecological functions [3]. The soil microbiome, the diverse collection of bacteria and other microorganisms within soil ecosystems, correlates directly with agricultural productivity forming a symbiotic association using up minimal energy from plant species. Just as the human gut microbiome contributes to human health, the soil microbiome supports the vitality of plants [9]. This study leverages the relationship between microbial composition and plant growth to explore the potential of machine learning in predicting crop yields based on soil bacterial composition. Regression models are used to identify and analyze the connections between soil composition and crop yields. These models are well-suited for predicting continuous outcomes and identifying patterns in complex datasets.

[*] Corresponding author: Mahi Shah.

2. Data Set

For this study, a dataset of 627 distinct bacterial strains was identified within the soil of approximately 1,340 farmland sites, alongside the recorded crop yields, allowing for the correlation of soil microbiome data with agricultural productivity [5].



	Xanthomonadales	Glutamicibacter	Geobacillus	Rickettsia	Armatimonadales	Phaselicystis
farm_0	0.0	0.0	0.0	0.0	0.0	2.0
farm_1	156.0	0.0	0.0	0.0	0.0	77.0
farm_2	133.0	0.0	0.0	7.0	95.0	60.0
farm_3	199.0	0.0	0.0	0.0	0.0	95.0
farm_4	0.0	0.0	0.0	0.0	0.0	0.0

Figure 1 Subset of real dataset of 1340 farms' 627 distinct bacterial composition

Before initializing the machine learning model, data cleaning removed low-prevalence bacteria-strains appearing in fewer than 10 samples, reducing noise and preventing the model from overfitting on rare instances, thereby enhancing dataset robustness. Additionally, the data was log normalized to emphasize relative changes over absolute counts, which helps mitigate the impact of outliers and provides a more meaningful comparison of bacterial abundance across samples. This transformation aims to stabilize variance and bring the data closer to a normal distribution, ultimately leading to more accurate and reliable analysis.

To better understand the dataset and guide model creation, hierarchical clustering, an unsupervised learning technique, was applied to group bacterial strains with similar abundance patterns and cluster farms with comparable microbiome compositions using Seaborn, a Python data visualization library capable of statistical graphics [8].

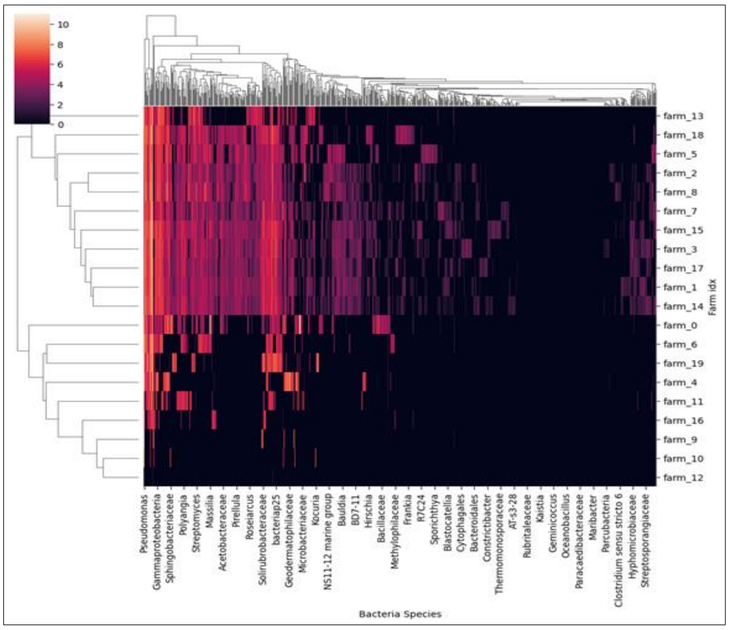


Figure 2 Heatmap color-coded visualization of bacterial species in farmlands

The heatmap visualization shows bacterial species on the x-axis and farmland sites on the y-axis, with color intensity representing bacterial abundance. Dendrograms on both axes highlight clusters, revealing patterns in bacterial communities across farms [2]. This process is valuable for identifying if features are correlated or if there are certain outliers in the dataset, improving the model's performance in the future.

These preprocessing steps ensure that the input data is cleaner, more representative, and suitably scaled, enhancing the machine learning model's performance in the future and leading to more accurate predictions while also providing deeper insights into the data and revealing which bacterial strains are correlated.

3. Computing environment

11th Gen Intel(R) Core (TM) i7-11390H @ 3.40GHz 2.92 GHz, 64-bit OS, Windows 11 Home Version 24H2 with access to internet. Python code was created on a Google Colab environment with access to 100 GPU compute units.

4. Architecture

When determining the most suitable model architecture for this study, three machine learning models were considered: Decision Tree, K-Nearest Neighbors (KNN), and Multilayer Perceptron (MLP). Decision Trees works by recursively splitting the dataset based on feature values to create a tree-like structure of decisions. Each internal node of the tree represents a feature test, each branch represents the outcome of the test, and each leaf node represents a class label or decision outcome. The tree is built by selecting the feature that best separates the data at each step, using criteria such as Gini impurity [1] and MAE (Mean Absolute Error). It is a prospective model as this method allows for precise and interpretable decision-making processes, making it particularly well-suited for problems where model transparency is critical. However, they are prone to overfitting, especially with noisy datasets. Training complexity is $(m \cdot n \cdot \log(n))$, where m is the number of features, and n is the number of training samples.

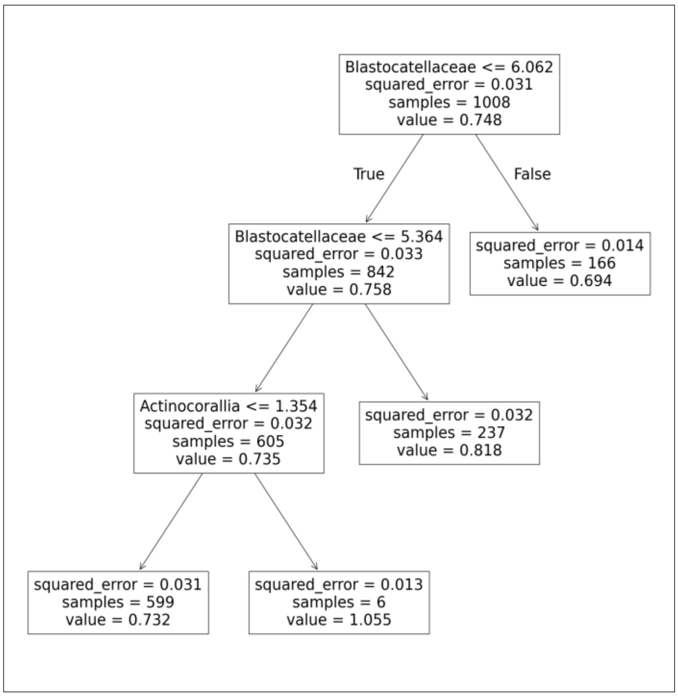


Figure 3 Tree-based Visualization of Regression Decision Tree ML-model evaluation for crop-yield prediction

Each "box" is a node, representing a decision or a final prediction (leaf). Lines indicate the flow of data based on conditions in the parent node. The first line in each box specifies the splitting condition. For instance, "Feature <= Value" directs the data into branches based on the condition's truth. MSE (Mean Squared Error) measures the prediction error at that node, "samples" denotes the number of data points, and "value" indicates the predicted outcome for those samples. The leaves (bottom row) show the final predictions, MSE, sample count, and predicted values.

The KNN model classifies data points based on the majority class among the k-nearest neighbors in the feature space. This model works by calculating the distance between data points, typically using Euclidean distance, and then assigning the class of the nearest neighbors to the new data point [6]. KNN is effective in handling non-linear relationships and smaller data sets. However, it is computationally expensive during prediction due to its $(n \cdot d)$ complexity per query, where n is the number of training samples and d is the dimensionality of the data. Additionally, KNN is sensitive to irrelevant or redundant features, which can degrade its performance in high-dimensional spaces.

The MLP model, a type of neural network, consists of multiple layers of nodes, each connected to nodes in the previous and subsequent layers. It employs non-linear activation functions to capture complex patterns and relationships in the data. MLPs are trained using backpropagation, where the error is propagated backward through the network to update the weights, improving the model's predictions [7]. This model is highly flexible as it can handle large datasets with high dimensionality and learn intricate data structures, but it is sensitive to hyperparameter tuning. Training complexity depends on the architecture, typically $(k{\cdot}m{\cdot}n)$, where k is the number of epochs or times the dataset is passed through the model, m is the model size, and n is the training.

After assigning the variables X and y to their respective data frames, the dataset was split into training and testing subsets using the train_test_split() function from the sklearn.model_selection package in Python. The training set, which contains most of the data, has been used to train the models using the fit method with X-train and Y-train. After training, the model's performance was assessed by predicting the labels of the testing data using the prediction method. These predictions were then plotted on a graph to visually evaluate the models' accuracy and effectiveness on unseen data.

To further experiment the effectiveness of the model, the importance of each feature in predicting crop yield was determined by examining the feature importances provided by the model. These importances indicate how much each feature contributes to the model's decision-making process. For the decision tree model, the feature importance is as follows: Xanthomonas (0.060571), Chryseobacterium (0.055838), Hungatei Clostridiaceae (0.055538), Acidobacteriaceae (Subgroup 1) (0.053313), and Salinispora (0.052732). It is determined by the decrease in impurity brought by the feature in splits across all nodes. Features that result in more significant reductions in impurity are considered more important. This metric helps identify which features contribute most to the model's predictive accuracy and are most influential in predicting the outcome.

Usually, the Y-axis shows the number of bacterial strains while the X-axis of the graph represents the feature importance values of bacterial strains in predicting crop yield. In decision tree models, feature importance quantifies each feature's contribution to reducing impurity and making accurate predictions. Higher values on the X-axis indicate that certain bacterial strains significantly influence crop yield predictions. This typically ranges from 0 to 1, where a value closer to 1 indicates higher importance, and a value closer to 0 indicates lower importance. These values help to quantify how significantly each bacterial strain contributes to the prediction of crop yields.

5. Evaluation

In evaluating the performance of the models, two key metrics were utilized: Coefficient of Determination (R^2) and Mean Absolute Error (MAE).

Table 1 Comparison of R^2 and MAE metrics for evaluating model’s performance

Aspect	**R^2**	**MAE**
Type of Metric	Relative (measures fit relative to variance)	Absolute (measures average error directly)
Range of Values	[−1, 1], with 1 indicating perfect fit in this case	[0, ∞), with 0 indicating perfect accuracy
Interpretation	Indicates how much of the variance is explained by the model	Indicates the average magnitude of errors
Sensitivity to Scale	Scaled by variance; less interpretable in terms of raw errors	Directly interpretable in the same units as the data
Effect of Outliers	Can be influenced by outliers (as variance increases)	Moderately robust to outliers (does not square differences)
Comparison to Baseline	Assesses improvement over predicting the mean	Does not compare to a baseline model explicitly

The Coefficient of Determination denoted as R^2 measures how well the variability in the predicted values is explained by the variability in the actual values. As R represents the correlation between the predicted and actual test values, an R^2 score closer to 1 indicates a model that accurately captures the data trend, akin to a linear relationship.

On the other hand, Mean Absolute Error (MAE) is the average absolute difference between the predicted and actual values. It is calculated using the mean_absolute_error function from the sklearn.metrics module in Python. A lower MAE value signifies that the model's predictions are closer to the actual values, indicating better performance.

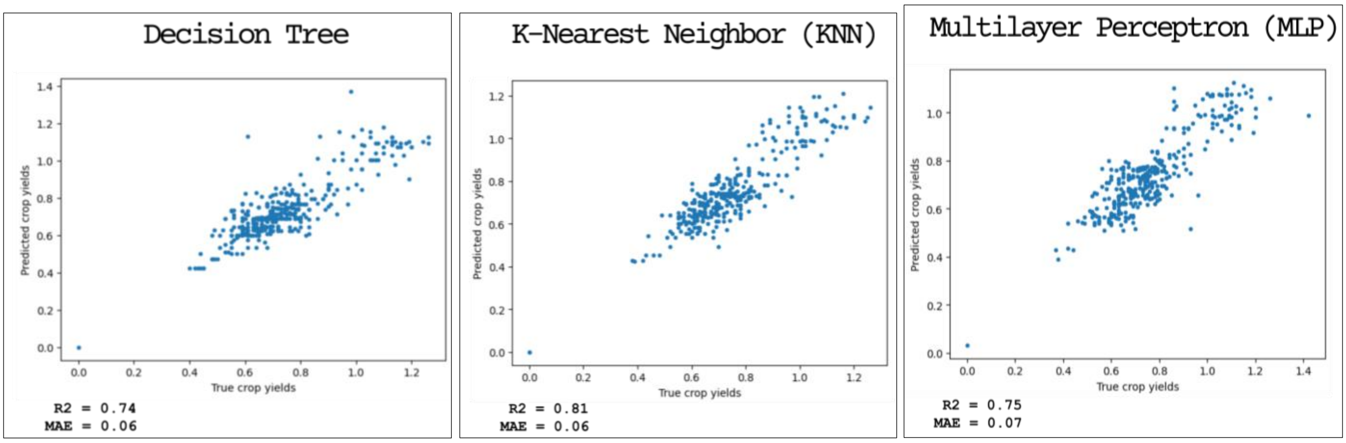


Figure 4 Visual Analysis of 3 Machine Leaning Models' Crop-Prediction against Actual-Crop-Yield

Graphs were generated using the Matplotlib Python library to visualize the comparison between testing and training data.

The KNN model, evaluated with n_neighbors=3, achieved an R^2 score of 0.8052 and a Mean Absolute Error (MAE) of 0.0645. The Decision Tree model, with parameters max_depth=30 and max_leaf_nodes=150, obtained an R^2 score of 0.7404 and an MAE of 0.0671. The MLP model, configured with hidden layer sizes (14, 12, 15), resulted in an R^2 score of 0.7517 and an MAE of 0.0741. Among the three models, the K-Nearest Neighbor (KNN) model was the most effective, achieving the highest R^2 score of 0.8052 and the lowest Mean Absolute Error (MAE) of 0.0645. These metrics indicate that the KNN model demonstrated superior predictive accuracy and reliability compared to the Decision Tree and Multilayer Perceptron (MLP) models for predicting crop yields in this study.

6. Observation

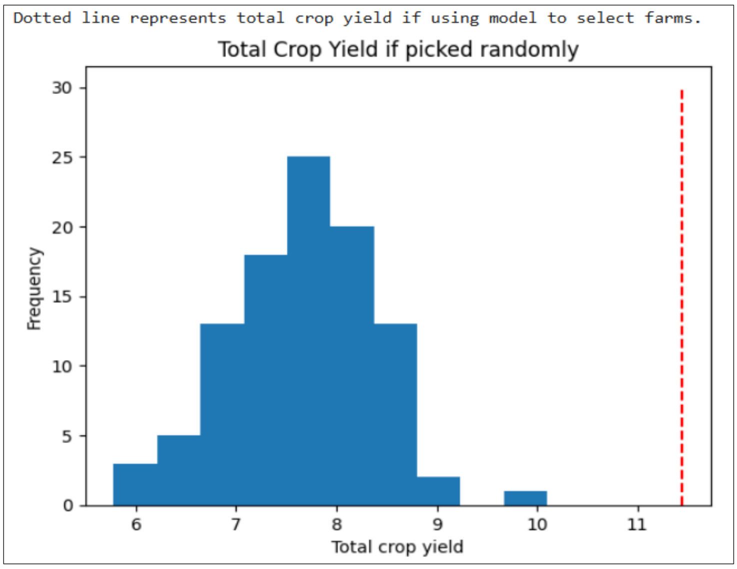


Figure 5 Comparison of crop-yield for randomly-picked-farmlands VS model-suggested-farmlands

The model's performance surpasses random chances, making it a valuable tool for government organizations to optimize land use. By accurately predicting crop yields based on soil microbiome data, the model identifies high-yield farmlands for agricultural purposes. This enables efficient resource allocation and allows less productive land designated for urbanization, balancing sustainable agriculture with urban development. It also helps in saving valuable resources like water, electricity, manpower by allocating it to the right farmlands for the most optimal production of crop.

The bar graph illustrates the results of 100 unique test cases, each simulating the random selection of 10 farmlands to evaluate their total crop yield. None of the test cases outperformed the total crop yield achieved using the machine learning model (dotted line), which identifies optimal farmland selections. This highlights the model's superiority over random selection and some land allocation systems.

7. Conclusion

In summary, the code exemplifies a structured approach to building and evaluating a model for crop yield. It demonstrates fundamental concepts in machine learning, including data preprocessing, model training, hyperparameters tuning, evaluation metrics computation, and visual representation of results, all crucial for developing robust predictive models in practice. A dataset of 627 distinct bacterial strains was identified within the soil composition of approximately 1,340 farmland sites, correlating soil microbiome data with agricultural productivity of Barley as crop. Data cleaning removed low-prevalence bacteria, while log normalization emphasized relative changes, enhancing model accuracy. The three considered machine learning models were Decision Tree, K-Nearest Neighbors (KNN), and Multilayer Perceptron (MLP). The dataset was split into training and testing subsets from Python packages and performance was assessed through Coefficient of Determination (R^2) and Mean Absolute Error (MAE), with KNN achieving the highest R^2 score of 0.8052 and the lowest MAE of 0.0645. The model's performance indicates its potential as a valuable tool for government organizations to optimize land use by identifying high-yield farmlands for agricultural purposes, thus balancing sustainable agriculture with urban development.

References

[1] Decision Trees. (2025). Retrieved November 10, 2024, from scikit-learn website: https://scikit-learn.org/stable/modules/tree.html

[2] 2.3. Clustering. (2019). Retrieved January 21, 2025, from scikit-learn website: https://scikit-learn.org/1.5/modules/clustering.html

[3] 16S and ITS rRNA Sequencing | Identify bacteria & fungi with NGS. (2020). Retrieved November 8, 2024, from Illumina.com website: https://www.illumina.com/areas-of-interest/microbiology/microbial-sequencing-methods/16s-rrna-sequencing.html

[4] Feeding 9 billion - National Geographic. (2019). Retrieved September 29, 2024, from Feeding 9 billion - National Geographic website: https://www.nationalgeographic.com/foodfeatures/feeding-9-billion/

[5] Index of /courses/MVE510. (2022). Retrieved February 14, 2024, from Chalmers.se website: https://bioinformatics.math.chalmers.se/courses/MVE510/

[6] KNeighborsClassifier. (2024). Retrieved November 10, 2025, from scikit-learn website: https://scikit-learn.org/1.6/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

[7] MLPClassifier. (2024). Retrieved November 10, 2025, from scikit-learn website: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

[8] seaborn: statistical data visualization — seaborn 0.13.2 documentation. (2024). Retrieved January 22, 2025, from Pydata.org website: https://seaborn.pydata.org/

[9] Suman, J., Amitava Rakshit, Siva Devika Ogireddy, Singh, S., Gupta, C., & J. Chandrakala. (2022). Microbiome as a Key Player in Sustainable Agriculture and Human Health. Frontiers in Soil Science, 2. https://doi.org/10.3389/fsoil.2022.821589.