TITLE:"Soil Data Analysis Using Machine Learning Models"

ABSTRACT

This project proposes a machine learning framework to analyze and predict soil conditions, leveraging environmental parameters such as soil moisture, temperature, air humidity, and nutrient levels. Various statistical and machine learning models, including Decision Trees, K-Nearest Neighbors (KNN), and Random Forest, are employed to classify soil health and predict crop-related outcomes. The proposed methodology integrates data preprocessing techniques like scaling and encoding with model evaluation strategies to assess performance in real-world scenarios. Results indicate an accuracy of up to 71% with Decision Trees, demonstrating the model's effectiveness in capturing relationships between features. This analysis serves as a foundation for optimizing irrigation systems, enhancing yield prediction, and supporting sustainable agricultural practices.

INTRODUCTION

Background

Sustainable agriculture is essential to meet the growing demand for food while conserving resources. Soil conditions, including moisture levels, temperature, and nutrient availability, significantly influence crop health and yield. Traditional methods of monitoring soil conditions are time-consuming and labor-intensive. Recent advancements in machine learning (ML) have enabled efficient analysis and prediction of soil health, paving the way for precision agriculture.

Problem Statement

Despite the proliferation of smart farming technologies, many systems lack the ability to process large datasets effectively and provide actionable insights. Accurate prediction models can help optimize resource utilization, reduce waste, and improve crop productivity.

LITERATURE SURVEY

1. **Soil Moisture Analysis:** Studies have shown that soil moisture is a critical parameter affecting plant growth.Techniques like remote sensing and IoT-based sensors are commonly used but require advanced data processing for actionable insights.

2. **Machine Learning in Agriculture**: ML models such as Decision Trees, KNN, and Random Forest have beenwidely adopted for predictive tasks in agriculture, including crop classification, yield prediction, and disease detection.

3. **Statistical Analysis in Agriculture:** Statistical measures like kurtosis provide insights into data distribution,helping detect anomalies and understand feature importance.

4. **Research Gaps:** There is a need for models that are not only accurate but also interpretable and adaptableto diverse agricultural datasets.

This project aims to address these gaps by integrating statistical analysis and ML techniques for comprehensive soil condition evaluation.

METHODOLOGY

1. Data Preprocessing

- Data is read from a CSV file (TARP.csv).

- Label encoding is applied to categorical features like 'Status' (ON/OFF).

- Features such as soil moisture and temperature are standardized using StandardScaler.

2. Statistical Analysis

- Kurtosis is calculated to understand the distribution of soil moisture data.

- Visualization includes histograms and density plots for feature insights.

3. Machine Learning Models

- Proposed Method: Decision Tree

Decision Tree Classifier is implemented with a maximum depth of 4 to balance interpretability and performance. It achieves an accuracy of 71%, with detailed feature importance analysis.

- K-Nearest Neighbors (KNN)

The KNN algorithm classifies data points based on feature proximity. With n_neighbors=5, it achieves an accuracy of 66%.

- Random Forest

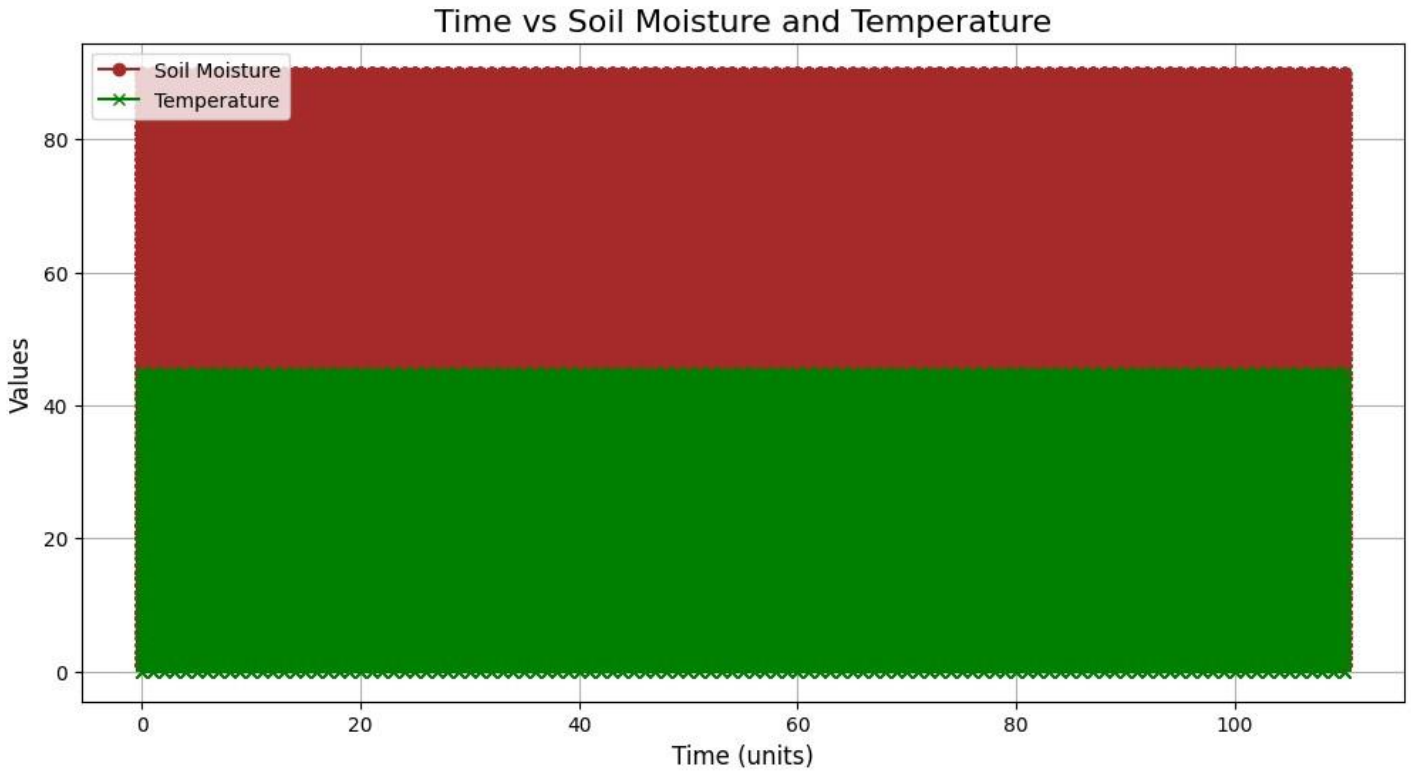
A Random Forest model is trained with 100 estimators to improve prediction reliability. It achieves an accuracy of 69% and highlights key features like soil moisture and nutrients.

4. Clustering Analysis

K-Means clustering segments data based on soil moisture and temperature, identifying clusters with shared characteristics and visualizing centroids.

RESULTS & DISCUSSION

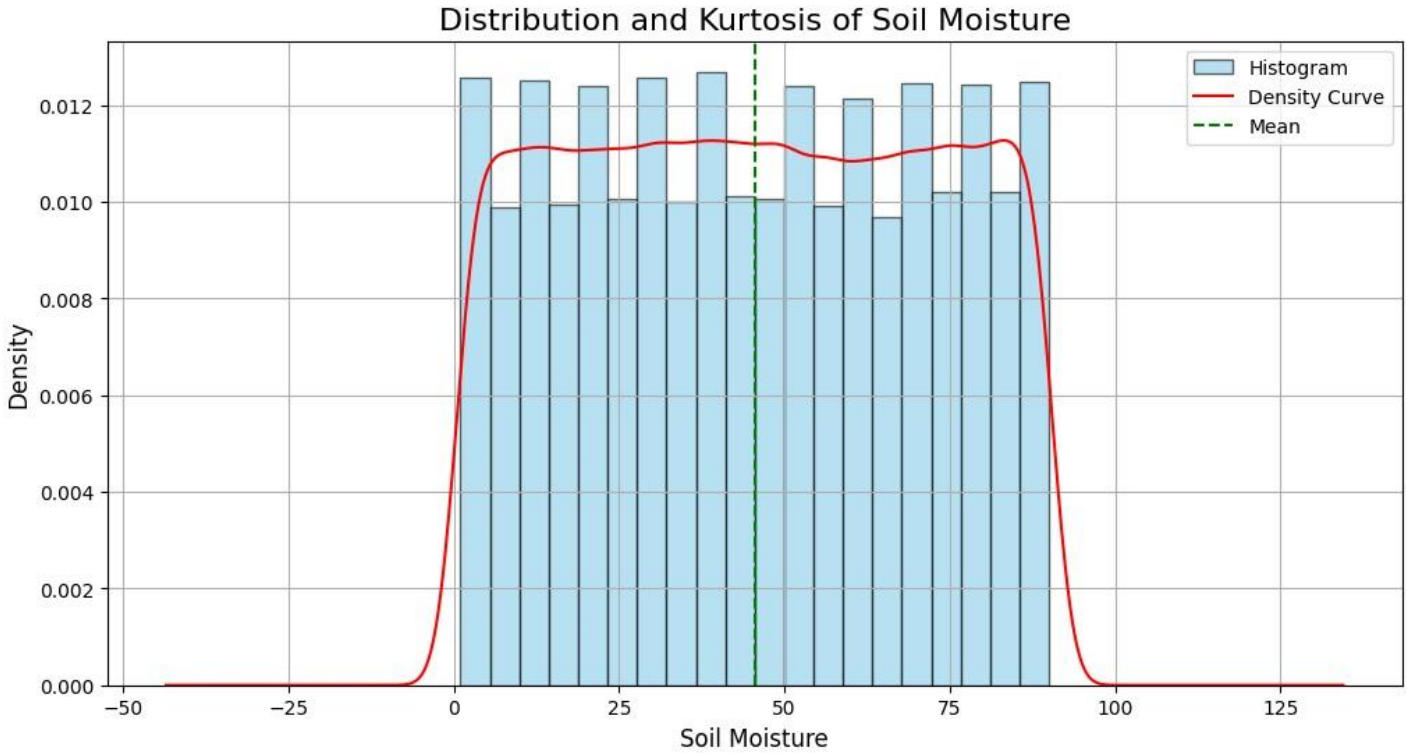
1:



Time vs. Soil Moisture and Temperature

- Explanation: Trends over time show soil moisture levels fluctuating alongside temperature changes, highlighting environmental impacts.

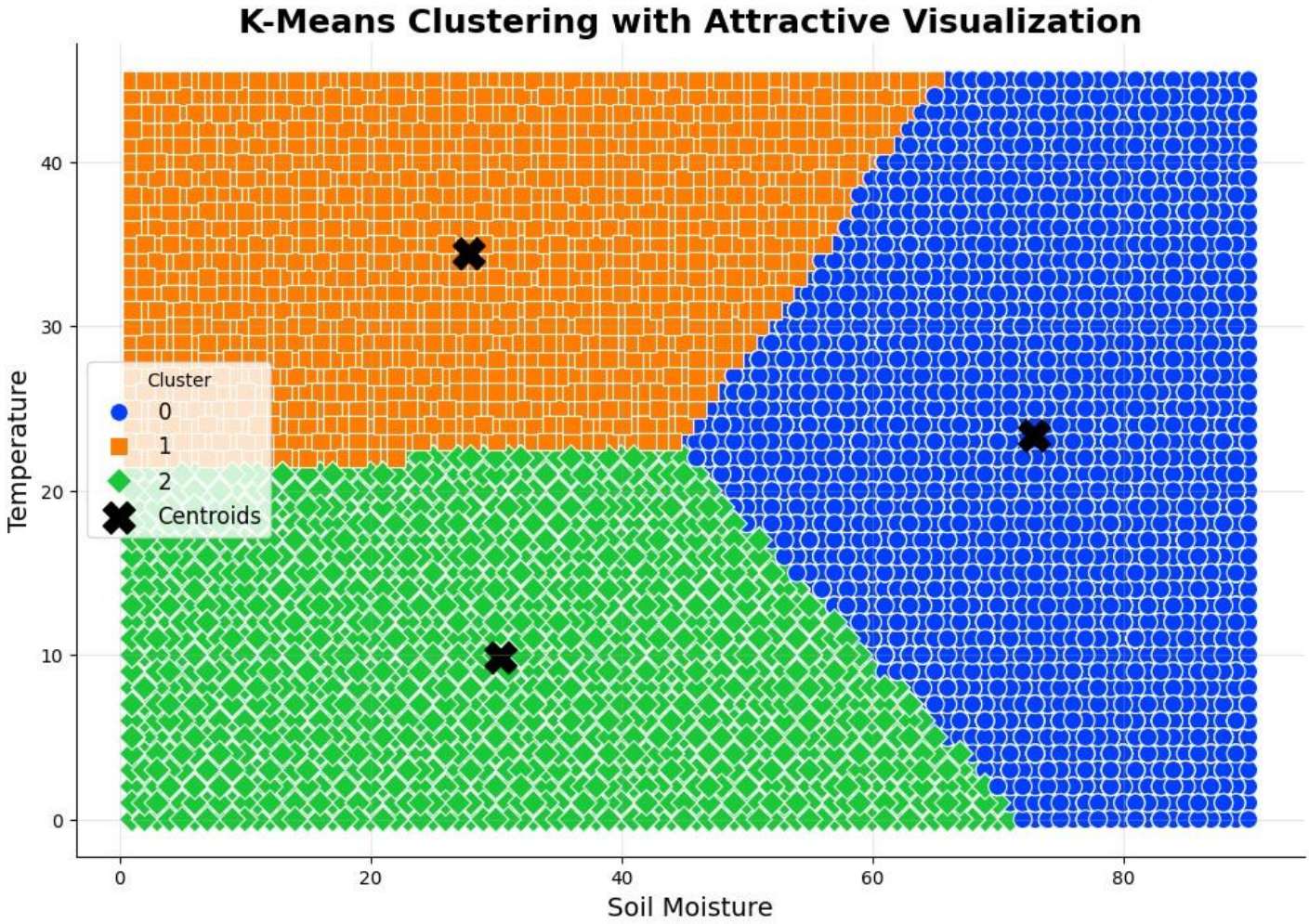
2.



Kurtosis Analysis

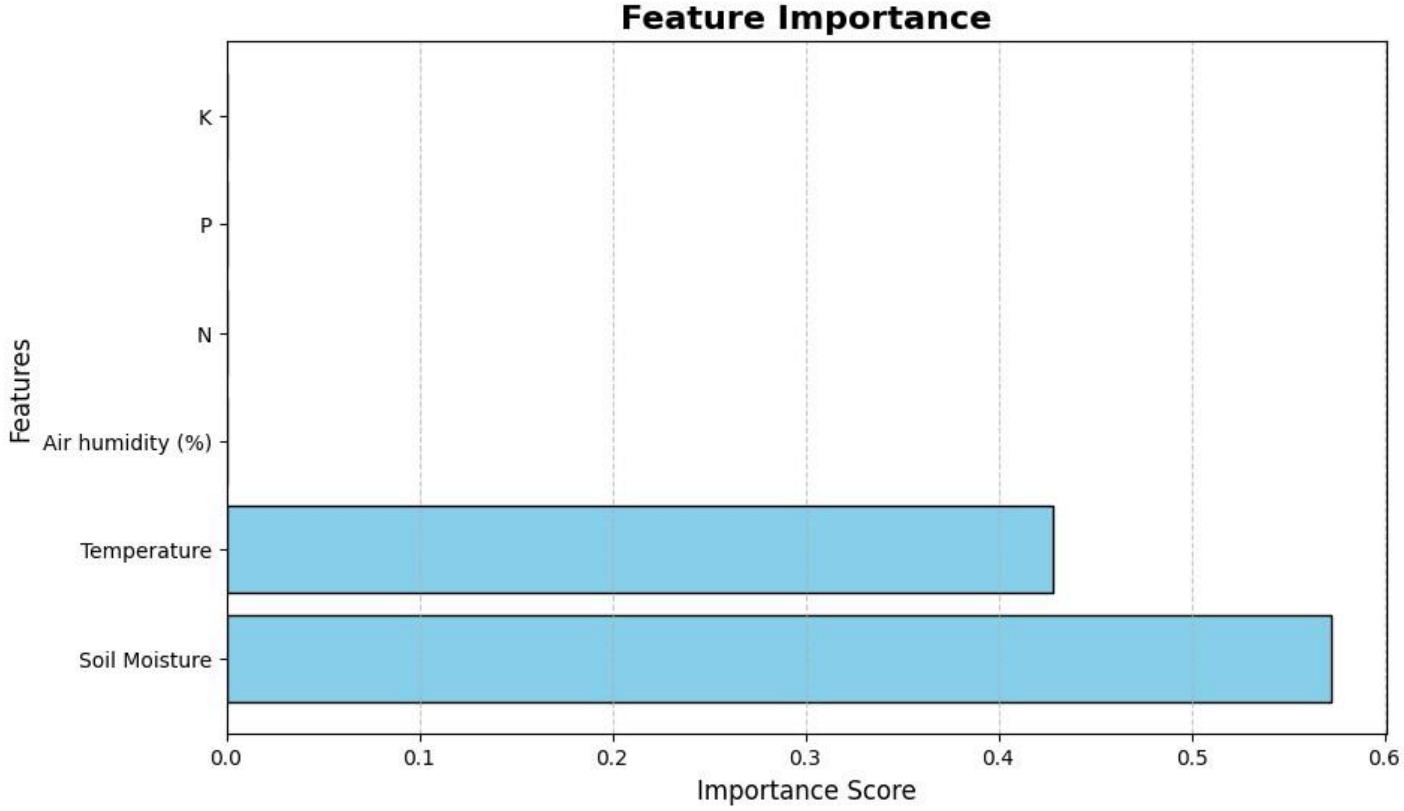
- Explanation: The kurtosis value (-1.2) indicates a platykurtic distribution, suggesting less extreme outliers.

3



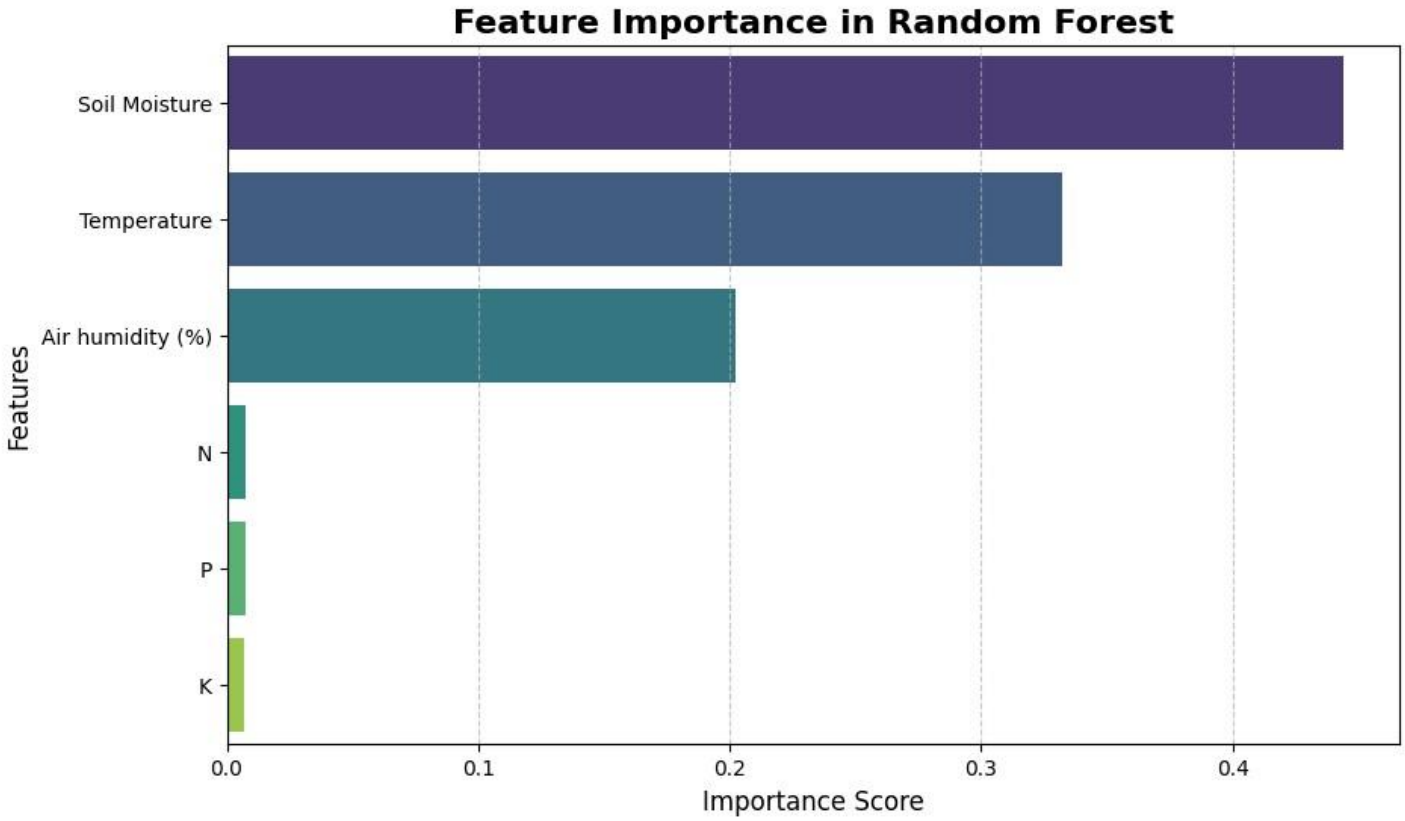
- Explanation: Three clusters are identified based on soil moisture and temperature. The centroids provide a reference for typical conditions.

4. **DecisionTreeStructure**



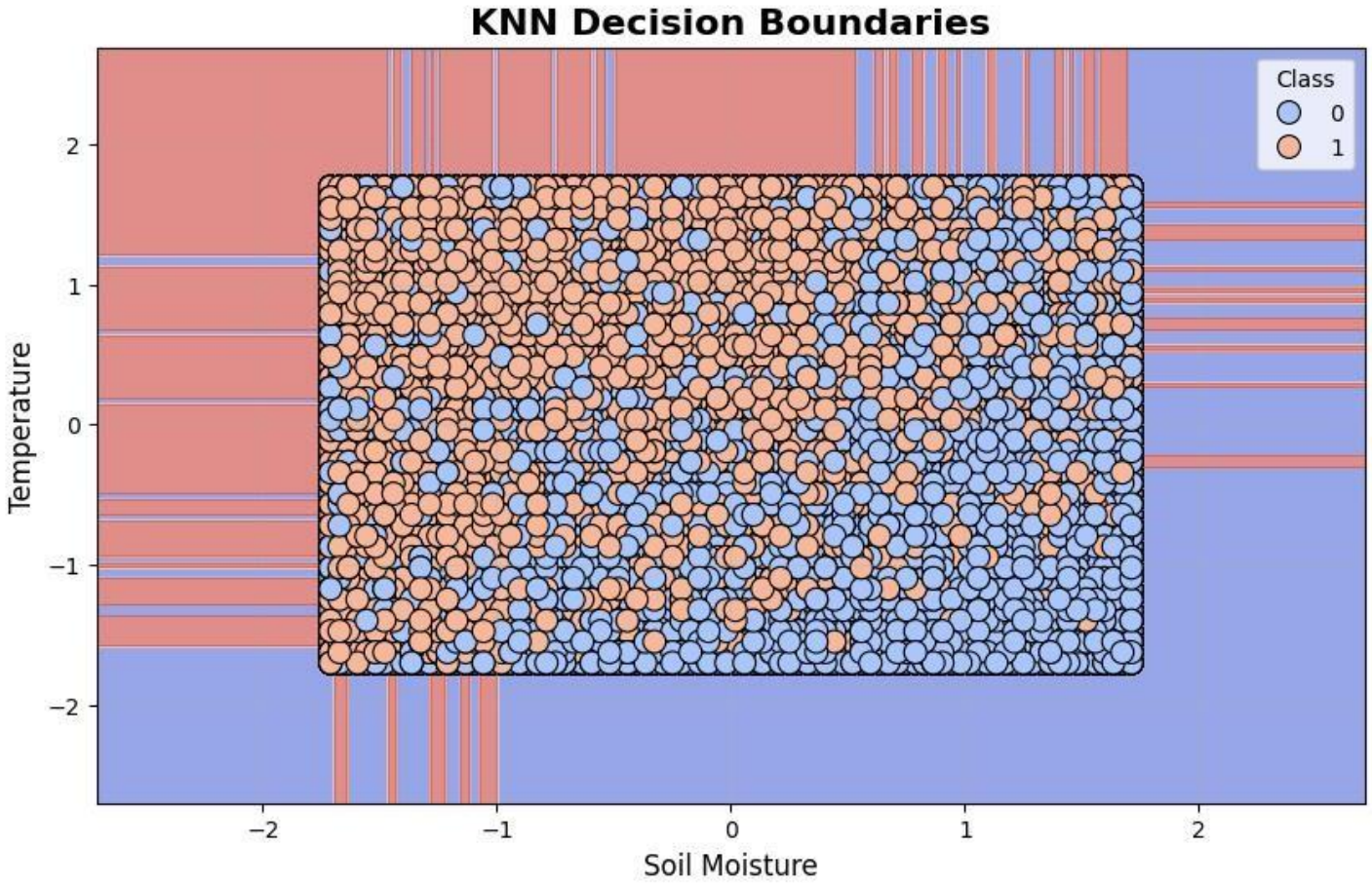
- Explanation: The tree highlights decision nodes based on features, showcasing interpretability. Features like soil moisture and nutrients dominate.

5.



- Explanation: Bar plots indicate soil moisture and temperature as the most influential features, aligning with domain knowledge.

6:



- Explanation: The boundary plot shows how KNN classifies soil conditions, with visible clusters matching observed patterns.

CONCLUSION

This project demonstrates the effective application of machine learning in soil condition analysis. Decision Trees emerged as the most accurate model, achieving a balance between interpretability and performance. Clustering and statistical techniques provide additional insights into data structure and feature relationships. The integration of these methods offers a comprehensive approach to optimizing agricultural processes, such as irrigation and nutrient management. Future work could focus on expanding datasets, incorporating real-time sensor data, and exploring advanced deep learning models to further enhance predictions.