Crop Recommendation System using Machine Learning & Deep Learning

MSc Computer Science and Data Science
Advanced Deep Learning
(MCSCIN5A1625)

Submitted by:
Divya JAYAPRAKASH
Jayasri DHANAPAL
Reshma KARTHIKEYAN NAIR

1. Project Overview

Agriculture plays a critical role in global food security, and selecting the right crop based on soil and climatic conditions is essential for maximizing yield and sustainability. This project presents the design and implementation of a **Crop Recommendation System** using both **Machine Learning (ML)** and **Deep Learning (DL)** techniques. The system predicts the most suitable crop for cultivation by analysing soil nutrients and environmental factors.

Beyond achieving high predictive accuracy, the project emphasizes a **rigorous and reproducible data science workflow**. The complete pipeline includes exploratory data analysis, preprocessing, feature engineering, feature selection, model training, hyperparameter tuning, explainability, and critical evaluation. Special attention is given to **model interpretability**, which is crucial for real-world agricultural decision-making, where transparency and trust in predictions are essential.

2. Problem Definition

The task is formulated as a **multi-class classification problem**, where the objective is to recommend one crop from a predefined set of crops based on input features describing soil and climate conditions.

- **Problem Type:** Multi-class Classification
- **Goal:** Predict the most suitable crop given soil nutrients and climatic features
- **Target Variable:** Crop label (22 distinct crop classes)

This problem is particularly challenging due to the multi-class nature of the target variable and the non-linear relationships between soil properties, climate conditions, and crop suitability.

3. Dataset

The dataset used in this project is a **publicly available agricultural crop recommendation dataset**, commonly used for benchmarking crop prediction models.

- **Number of Samples:** 2,200
- **Number of Classes:** 22 crop types (perfectly balanced)

- **Input Features:**
    - Nitrogen (N)
    - Phosphorus (P)
    - Potassium (K)
    - Temperature
    - Humidity
    - pH
    - Rainfall

The balanced nature of the dataset ensures that model performance is not biased toward any specific crop class. The features represent both **soil fertility indicators** and **climatic conditions**, making the dataset well-suited for learning complex interactions between environmental factors and crop suitability.

4. Project Pipeline

4.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to gain insights into the structure and characteristics of the dataset. This step included:

- Analysis of feature distributions to understand data ranges and variability
- Detection of potential outliers that could affect model performance
- Verification of class balance to ensure fair model training

EDA confirmed that the dataset is clean, balanced, and suitable for supervised learning, with no significant anomalies that require removal.

4.2 Preprocessing

To ensure reliable and reproducible results, the dataset underwent systematic preprocessing:

- **Train–test split** with stratification to preserve class distribution across both sets
- **Feature standardization** to normalize numerical features and improve model convergence

- Construction of **reproducible pipelines**, enabling consistent preprocessing across different models

This step ensures that models are trained on standardized data and that evaluation results are fair and comparable.

4.3 Feature Engineering

Feature engineering was performed to capture domain-specific knowledge and improve model performance. Newly engineered features include:

- Nutrient ratio features: **N/P, N/K, P/K**, reflecting relative nutrient availability
- **Soil fertility index**, combining multiple nutrient values into a single indicator
- **Climate interaction feature**, representing combined effects of temperature, humidity, and rainfall

These engineered features helped models better learn the complex relationships between soil composition, climate, and crop suitability.

4.4 Feature Selection

Feature relevance was evaluated using:

- **Correlation analysis** to identify redundant or highly correlated features
- **Random Forest feature importance**, which ranks features based on their contribution to predictions

This step ensured that the most informative features were retained, improving both performance and interpretability.

4.5 Machine Learning Models

Multiple classical machine learning models were implemented to compare performance:

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Decision Tree
- Random Forest

- Support Vector Machine (SVM)

These models were selected to represent a variety of learning paradigms, including linear, distance-based, tree-based, and margin-based approaches.

4.6 Hyperparameter Tuning

To optimize model performance:

- **GridSearchCV** was applied to tune Random Forest hyperparameters
- **RandomizedSearchCV** was used for efficient tuning of SVM
- Model performance before and after tuning was systematically compared

Hyperparameter tuning significantly improved accuracy, particularly for tree-based models.

4.7 Explainability

Explainability was addressed using **SHAP (SHapley Additive exPlanations)**:

- **SHAP summary plots** provided global explanations of feature importance
- **SHAP force plots** explained individual predictions
- Interpretation of results highlighted the dominance of climate and soil fertility features

This step enhances trust in the model and supports real-world adoption in agriculture.

4.8 Deep Learning

An **Artificial Neural Network (ANN)** was developed for tabular data:

- Fully connected dense layers with ReLU activation
- **Dropout regularization** to prevent overfitting
- **Early stopping** to improve generalization
- Loss curve analysis to validate training stability

Although deep learning achieved competitive results, it did not outperform the best classical machine learning models.

4.9 Critical Analysis

A detailed comparison of models revealed:

- Tree-based models (Random Forest, Decision Tree) achieved the highest accuracy
- Feature engineering played a crucial role in performance improvement
- Deep learning models require larger and more complex datasets to outperform classical approaches
- Interpretability remains a key advantage of tree-based models over neural networks

5. Final Model Performance (Accuracy)

Model	Accuracy
Random Forest	**98.86%**
Decision Tree	98.18%
Logistic Regression	97.73%
ANN	97.50%
KNN	96.36%
SVM	96.36%

6. Key Insights

- Tree-based models performed best on structured agricultural data.
- Feature engineering significantly improved model performance.
- ANN achieved competitive results but did not outperform classical ML models.
- SHAP explainability revealed climate and soil fertility as dominant factors.
- Model interpretability is critical for real-world agricultural decision support.

7. Limitations

- Dataset is static and does not capture temporal or seasonal variations.
- Geographic generalization may be limited.
- Real-world agricultural data may contain more noise and uncertainty.
- Deep learning interpretability remains limited compared to tree-based models.

8. Future Work

- Incorporate time-series weather data
- Integrate satellite or remote sensing information
- Expand dataset geographically
- Deploy as a farmer decision-support system

9. Conclusion

This project successfully implemented a Crop Recommendation System using Machine Learning and Deep Learning techniques to predict suitable crops based on soil nutrients and climatic conditions. A complete data science pipeline was followed, including preprocessing, feature engineering, model training, hyperparameter tuning, and explainability analysis. Experimental results showed that tree-based models, particularly Random Forest, achieved the highest accuracy, outperforming the deep learning model.

Feature engineering and SHAP-based explainability improved both performance and interpretability. Despite certain dataset limitations, the study demonstrates the effectiveness of machine learning–based decision-support systems in agriculture and provides a strong foundation for future enhancements using temporal and geospatial data.

10. Technologies Used

- Python
- NumPy, Pandas
- Matplotlib, Seaborn
- Scikit-learn

- TensorFlow / Keras
- SHAP

11. Project Structure

```
├── Data/
│   └── Crop_recommendation.csv
├── Notebooks/
│   └── Crop_Recommendation_ML_DL.ipynb
├── README.md
├── Crop Recommendation Report.pdf
```