**Intelligent Crop Recommendation System: A Reproduction and Validation**

**Name:** Soham S. Mishra

**Roll No:** 16014223082

**1. Introduction**

Agriculture is a critical sector of the Indian economy, with a vast portion of the population depending on it for their livelihood. A significant challenge faced by farmers is selecting the most suitable crop for their specific land and environmental conditions. An incorrect choice can lead to poor yields, financial losses, and soil degradation.

To address this issue, modern data-driven approaches using machine learning can provide powerful tools for decision-making. This project aims to reproduce and validate the methodology presented in the research paper "Smart Farming: Crop Recommendation Using Machine Learning with Challenges and Future Ideas" by Dahiphale et al. (2025). The goal is to build an intelligent system that recommends the most suitable crop from 22 different options based on key soil and climate features.

**2. Methodology**

The methodology of the source paper was carefully reproduced and applied.

* **Dataset:** The project utilizes the publicly available **Crop Recommendation Dataset from Kaggle**. This dataset contains 2200 instances and was created by augmenting data on rainfall, climate, and fertilizers available for India.
* **Features:** The prediction is based on seven key environmental and soil features:
  1. **N:** Nitrogen content in the soil (kg/ha)
  2. **P:** Phosphorus content in the soil (kg/ha)
  3. **K:** Potassium content in the soil (kg/ha)
  4. **Temperature:** Temperature in Celsius
  5. **Humidity:** Relative humidity in %
  6. **pH:** pH value of the soil
  7. **Rainfall:** Rainfall in mm
* **Data Preprocessing:** The dataset was loaded and split into a **70% training set and a 30% testing set** as specified in the paper. For the Neural Network model, the categorical crop labels were converted to a one-hot encoded format.
* **Models Implemented:** Following the paper's comparative analysis, three primary machine learning models were implemented :
  1. Random Forest
  2. Naive Bayes
  3. A 4-Layer Neural Network

**3. Results and Comparison**

The implemented models were evaluated and their performance was compared against the results published in the source paper.

| Model Name | Paper's Validation Accuracy | My Validation Accuracy |
| --- | --- | --- |
| Random Forest | 99.5% | **99.45%** |
| Naive Bayes | 99.5% | **99.50%** |
| Neural Network | 97.73% | **96.10%** |

**Analysis of Results**

The reproduction of the research paper was highly successful. The performance of the

**Random Forest (99.45%)** and **Naive Bayes (99.50%)** models were nearly identical to the paper's reported accuracy of 99.5%, which confirms their exceptional effectiveness for this dataset .

Initially, the Neural Network's performance was lower. However, by increasing the training from 100 to

**1000 epochs**, to match the methodology in the paper, the validation accuracy significantly improved from 92.86% to **96.10%**. This result is now very close to the paper's reported 97.73%. The small remaining difference can be attributed to minor variations in hyperparameter tuning or the random initialization of the network weights.

**4. Feature Importance Analysis**

To better understand the model's decisions, a feature importance analysis was conducted on the best-performing model (Random Forest).

The analysis clearly indicates that environmental conditions, specifically **rainfall** and **humidity**, are the two most dominant factors in determining a suitable crop. Following these climatic factors, the primary soil nutrients—**Potassium (K)**, **Phosphorus (P)**, and **Nitrogen (N)**—play the next most significant role. This analysis makes the model's logic interpretable and aligns with real-world agricultural principles.

**5. Challenges Faced & Learnings**

During the project, several technical challenges were encountered, which provided valuable learning opportunities:

* **Environment Setup:** Initial ModuleNotFoundError issues highlighted the importance of properly setting up a project's Python environment and installing all required libraries from a requirements.txt file.
* **Jupyter Notebook Workflow:** A series of NameError issues emphasized the "Golden Rule of Notebooks": cells must be run sequentially from top to bottom, as the kernel's memory is not persistent and all setup cells (imports, data loading) must be executed before the main logic.

**6. Conclusion**

This project successfully reproduced the work of Dahiphale et al. (2025) and validated their key findings on crop recommendation using machine learning . The high accuracy achieved across all implemented models demonstrates that machine learning is a valuable and reliable tool for smart farming. Furthermore, the feature importance analysis provided insight into the model's decision-making process, confirming that environmental factors like rainfall and humidity are the most critical predictors. The project provides a scalable and accurate technique that can empower farmers to make informed, data-driven decisions.