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

**Name:** Soham 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 by following these steps:

* **Dataset:** The project utilizes the publicly available **Crop Recommendation Dataset from Kaggle**. This dataset was created by augmenting data on rainfall, climate, and fertilizers in India and contains 2200 instances.
* **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 prepared for model training. The data was 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 using 5-fold cross-validation to ensure robustness, and the results were compared against those 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% | **92.86%** |

Export to Sheets

**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.

My **Neural Network** model achieved a final validation accuracy of 92.86%. While this is a strong result, it is slightly lower than the 97.73% reported in the paper. This difference is likely because the paper mentions training for up to 1,000 epochs for optimal performance, whereas my model was trained for 100 epochs to ensure rapid execution. It is expected that increasing the training epochs would close this performance gap.

**4. Challenges Faced & Learnings**

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

* **Environment Setup:** The initial ModuleNotFoundError for sklearn highlighted the importance of properly setting up the Python environment and ensuring all required libraries from the requirements.txt file are installed before starting.
* **Jupyter Notebook Workflow:** A series of NameError issues (for pd, df, X\_train) emphasized the "Golden Rule of Notebooks": cells must be run sequentially from top to bottom. This taught me that a kernel's memory is not persistent and that all setup cells (imports, data loading) must be executed before the main logic.

**5. 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 by the implemented models, particularly Random Forest and Naive Bayes, demonstrates that machine learning is a valuable and reliable tool for smart farming. The project provides a scalable and accurate technique that can empower farmers to make informed, data-driven decisions, ultimately leading to improved agricultural productivity.