Machine Learning-Based Crop Recommendation System for Mizoram

V.D. Ambeth Kumar
dept. of Computer Engineering
Mizoram University
Aizawl, India
ambeth@mzu.edu.in

Ajoy Kumar Khan
dept. of Computer Engineering
Mizoram University
Aizawl, India
mzut250@mzu.edu.in

Vanlalhruaia
dept. of Computer Engineering
Mizoram University
Aizawl, India
mzut160@mzu.edu.in

Saithantluanga
dept. of Computer Engineering
Mizoram University
Aizawl, India
saithantluangack@gmail.com

Ramesh Prabhakaran R
dept. of Computer Engineering
Mizoram University
Aizawl, India
rkrrameshsamy5@gmail.com

Zaitinkhuma
dept. of Computer Engineering
Mizoram University
Aizawl, India
zaiamay94@gmail.com

2025 3rd International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC) | 979-8-3315-2389-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/ISACC65211.2025.10969444

Abstract - Agriculture is an important sector of Mizoram domicile as more than half of its population relies on Agriculture as principle source of income and sustenance. Some farmers rely on the knowledge acquire from their parents through explicit explanation, and by observing and modelling their practices. But most farmer often struggle to understand the method and type of crops to cultivate for better crop yield. Even experienced farmer believed that using more fertilizer result in better crop yield in spite of that it damages the soil properties. To resolve this challenge, this paper presents a Crop Recommendation System using Machine Learning, tailored for Mizoram, enhancing Agriculture practices towards sustainable development. The Crop Recommendation System analyzes historical data, soil properties, weather pattern and crop performance to recommend the best crop for a specific region and its condition. The aim is to provide the Crop Recommendation System with information about the soil and the condition of the region. The study utilizes various Machine Learning Algorithms such as Random Forest, Decision Trees, Support Vector Machine and Logistic Regression to make optimal recommendation. The results indicate that Random Forest provides superior performance of 99% across all the evaluations metrics. Although many prevention measures need to be taken to avoid complications such as overfitting, etc. The overall results suggested that Random Forest achieved the best results as compared to all the other state of the art algorithms utilized with the same preprocessing steps. This approach enhances the crop and soil. After a long and often complicated process of farming method and selection of crop problem the Crop Recommendation System will aid Mizoram farmers to achieve better crops, yield and higher profit.

Keywords - Crop Recommendation (CRS) System, Machine Learning, Agriculture, Mizoram, Decision Trees, Random Forest, Support Vector Machines

I. INTRODUCTION

More than 60% domicile of Mizoram relies on farming for sustenance and for source of income. Mostly farmer practice “Jhum” cultivation with the assistance of their forefathers practices. Food crops are maize and rice, while crop such as sugarcane, tapioca, ginger take the part of cash crop that play significant roles according to production per area. The region of Mizoram around 83% of it is covered by Forest. Shifting cultivation involve deforestation and burning vegetation in order to make the soil enrich in nutrient for their crop yet leads to ill-treatment of environment and habitat. Additionally, experienced farmer dwelling in the hilly area practice terrace farming overuse fertilizers leading to increase cost and potential soil damage. Thus, traditional farming methods subsequently result in suboptimal yield in future and economic instability.

To resolve these challenges, this paper present a Crop Recommendation System (CRS) using Machine Learning (ML), tailored for Mizoram. The CRS analyses soil factor such as Nitrogen, Phosphorus, Potassium, pH, Moisture and Electrical Conductivity (salinity) levels, as well as environmental factor like rainfall, humidity and temperature. By providing all these accurate factors, the CRS provide data-driven recommendation. And helps to identify the best suitable crops for the information given, leading to better crop and yield for their soil and climatic condition.

Machine learning algorithm include Decision Tree, Support Vector Machine, Random Forest and Logistic Regression are utilized to create accurate and robust System. This CRS aims to minimize the effect of nutrient deficiency in crops and also enhance the quality of soil henceforth. This technology also increases productivity and profitability and offers a promising solution for sustainable development of agriculture in Mizoram.

II. RELATED WORK

The integration of machine learning (ML) techniques in agriculture has emerged as a transformative approach to enhance productivity and sustainability. This literature review synthesizes findings from key studies to understand the advancements in crop recommendation systems and yield prediction methods using machine learning, focusing on their applicability to regions like Mizoram, India.

Chen, Liaw, and Breiman (2004) [1] highlighted the robustness of Random Forest algorithms in handling imbalanced and high-dimensional datasets, which are common in agricultural data. This technique's ability to process complex data effectively makes it a suitable choice for crop recommendation systems. Quinlan (1986) [2] introduced the Induction of Decision Trees, a foundational work that has since become integral to agricultural data analysis due to its simplicity and interpretability. Cortes and Vapnik (1995) [3] established Support Vector Machines (SVM) as a powerful classification method, particularly influential in scenarios requiring high accuracy in crop classification and yield prediction. The Government of Mizoram (2020) [4] provided comprehensive agricultural statistics, crucial for understanding regional agricultural patterns. This data is essential for developing localized crop recommendation systems tailored to Mizoram's unique topography and climatic conditions. Traditional agricultural practices in Mizoram, such as "Jhum" cultivation, and the cultivation of key crops like maize, rice, sugarcane, tapioca, ginger, and cotton, underline the need for modern, data-driven approaches to improve productivity and sustainability.

Kumar and Singh (2015) [5] proposed a crop recommendation system for precision agriculture, leveraging machine learning algorithms to provide tailored crop recommendations. This system demonstrated significant potential in enhancing productivity by offering data-driven insights. Patil and Biradar (2016) [6] applied a decision tree approach to crop selection in Karnataka, India, showing the practical applicability of decision tree algorithms in real-world agricultural settings. Murthy and Prasad (2017) [7] explored various machine learning techniques for crop yield prediction and selection, emphasizing their potential to optimize agricultural practices. Zhang, Wang, and Huang (2018)[8] developed a hybrid model combining SVM and decision trees for crop yield prediction, showcasing the benefits of integrating multiple machine learning techniques to improve prediction accuracy.

Yadav and Yadav (2019) [9] provided a comprehensive review of machine learning applications in agriculture, summarizing advancements and highlighting challenges. Their review underscores the need for continuous research and development to overcome existing barriers and fully realize the benefits of machine learning in agriculture. Sharma and Singh (2020) [10] focused specifically on crop recommendation systems, providing insights into the latest methodologies and their effectiveness in enhancing agricultural decision-making processes.

The reviewed studies collectively underscore the transformative impact of machine learning on agriculture. Techniques such as Random Forest, decision trees, and SVM have proven effective in crop recommendation and yield prediction. The integration of these methods with regional agricultural statistics, as demonstrated in the studies, can lead to more accurate and tailored agricultural solutions. These advancements ultimately contribute to increased productivity and sustainability, enabling farmers to make more informed decisions, optimize resource use, and achieve better yields and higher profits. The ongoing development and application of machine learning techniques in agriculture hold significant promise for addressing future agricultural challenges and meeting the growing global food demand.

III. PROPOSED WORK

The main architecture of the proposed system is shown in Figure 1. It consist of the following phase a. Data Collection b. Data Preprocessing c. Model Training d. Evaluation e. Deployment

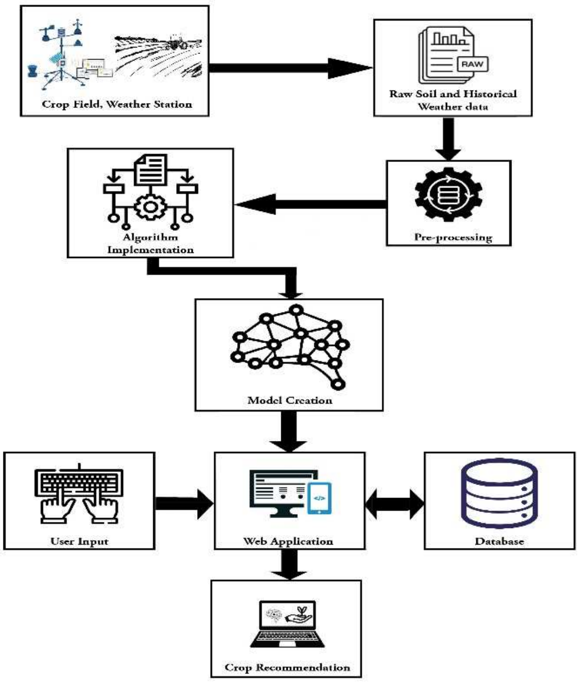

Figure 1: Architecture of CRS

A. Data Collection

Data collection is a critical step in developing a CRS, as the accuracy and reliability of the recommendations depend on the quality and completeness of the data. For this study since crops of the same type required the same amount of soil properties and climatic conditions all over the world, we are looking for a dataset that can provide us such that. The data that is being used is taken from Kaggle and comprises the following variables as shown in Figure 2 and Figure 3.

Soil Properties: Soil pH, nutrient levels (nitrogen, phosphorus, potassium).

Climatic Conditions: Temperature, rainfall, humidity.

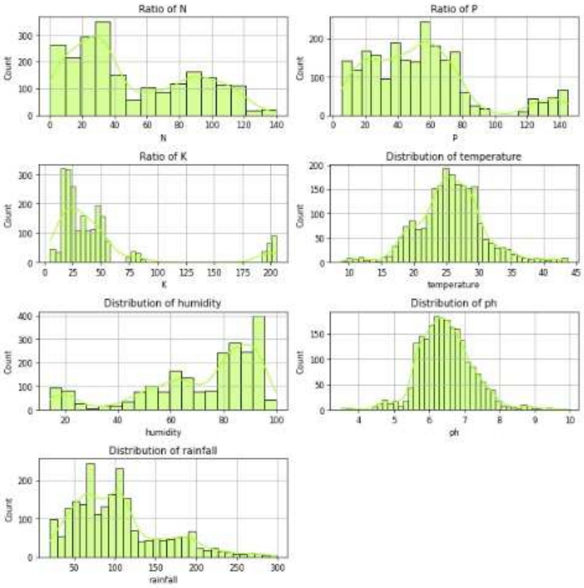

Figure 2: Histogram diagram of different features on the dataset

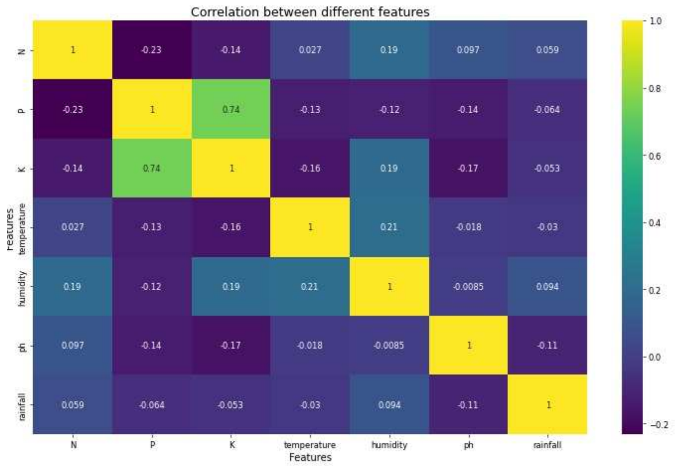

Figure 3: Correlation matrix of different features

Also the dataset have a 22 different categories of crops. Since we are also interested in recommending new crops to our farmers, we did not exclude any crops so that our farmers will try to grow different kinds of crops that they have not tried cultivating before.

B. Data Pre-Processing

The collected data underwent preprocessing to handle missing values, normalize features, and encode categorical variables. Data preprocessing is essential to ensure that the data is clean, consistent, and suitable for analysis. The following steps were performed during data preprocessing:

Handling Missing Values: Missing values in the dataset were handled using various techniques, including mean/mode imputation, interpolation, and removal of records with excessive missing values.

Algorithm 1: Handling Missing Values

```
Input: Raw dataset $\boldsymbol{X}$ with missing values
Output: Dataset $\boldsymbol{X'}$ with handled missing values
for each numerical feature $\boldsymbol{X_i}$ in $\boldsymbol{X}$ do
    if missing values are found then
        if percentage of missing values is low then
            impute missing values using mean:
                $$X'_{ij} = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$$
            or median:
                $X'_{ij} = \text{median}(X_{ij})$
            else
                Remove the feature $X_i$
        end if
    end if
end for
for each categorical feature $X_i$ in $\boldsymbol{X}$ do
    if missing values are found then
        Impute using the mode:
            $$X'_{ij} = \text{mode}(X_i)$$
    end if
    if number of missing values exceeds a threshold then
        Remove the corresponding records
    end if
end for
Return dataset $\boldsymbol{X'}$ with missing values handled
```

Normalization: Features were normalized to ensure that they are on the same scale, which is essential for ML algorithms to perform effectively. Normalization techniques, such as Min-Max scaling and Z-score normalization, were used.

Algorithm 2: Normalization

```
Input: Dataset $\boldsymbol{X}$ with numerical features
Output: Normalized dataset $\boldsymbol{X'}$
for each numerical feature $\boldsymbol{X_i}$ in $\boldsymbol{X}$ do
    Apply one-hot encoding:
        $$X'_{ij} = \frac{X_{ij} - \min(X_i)}{\max(X_i) - \min(X_i)}$$
    or Z-score normalization:
        $$X'_{ij} = \frac{X_{ij} - \mu_i}{\sigma_i}$$
  where $\mu_i$ is the mean and $\sigma_i$ is the standard deviation of $\boldsymbol{X_i}$
end for
Return normalized dataset $\boldsymbol{X'}$
```

Encoding Categorical Variables: Categorical variables were encoded using techniques such as one-hot encoding and label encoding to convert them into numerical values suitable for ML algorithms.

Algorithm 3: Encoding Categorical Variables

```
Input: Dataset $\boldsymbol{X}$ with categorical features
Output: Encoded dataset $\boldsymbol{X'}$
for each categorical feature $\boldsymbol{X_i}$ in $\boldsymbol{X}$ do
    Determine number of unique categories $k_i$
        if $k_i$ is small then
            Apply one-hot encoding:
                $$X'_{ij} = \begin{cases} 1, & \text{if } X_{ij} \text{ is the } j^{th} \text{ category} \\ 0, & \text{otherwise} \end{cases}$$
        else
            Apply label encoding:
                $X'_{ij} = f(X_{ij})$
    Where $f(X_{ij})$ is the encoding function

        end if
end for
Return encoded dataset $\boldsymbol{X'}$
```

Feature Selection: Feature selection techniques, such as correlation analysis and feature importance scores, were applied to identify the most relevant attributes for crop recommendation. Irrelevant or redundant features were removed to improve the model's performance.

Algorithm 4: Feature Selection

```
Input: Preprocessed dataset $\boldsymbol{X}$ with numerical and
          categorical features
Output: Dataset $\boldsymbol{X'}$ with selected features
for each feature $\boldsymbol{X_i}$ in $\boldsymbol{X}$ do
    Compute corelation with target variables $\boldsymbol{Y}$:
        $r_i = \text{corr}\ (\boldsymbol{X_i}, \boldsymbol{Y})$
    Compute feature importance score $s_i$ using a model
```

```
        if $r_i$ or $s_i$ is below threshold then
            Remove feature $\boldsymbol{X_i}$
        end if
end for
Return encoded dataset $\boldsymbol{X'}$
```

C. Model Training

Several ML algorithms were evaluated for their performance in crop recommendation. The chosen algorithms include:

Decision Trees: A tree-based model that splits the data into subsets based on feature values, creating a tree structure for decision-making. Decision Trees are simple and interpretable, making them suitable for crop recommendation.

Random Forest: An ensemble method that constructs multiple decision trees and aggregates their predictions for improved accuracy and robustness. Random Forest is known for its high accuracy and ability to handle large datasets with multiple features as shown in Figure 4.

Algorithm 5: Handling Missing Values

```
Input:
    Training dataset $\boldsymbol{D} \leftarrow \{(X_i, y_i)\}_{i=1}^{N}$
    $n_{trees}$ ← number of trees
    $d_{max}$ ← maximum depth
    $s_{min}$ ← minimum sample to split
    $f_{max}$ ← maximum features
Output: Trained Random Forest model
Initialize an empty list of trees   $\boldsymbol{T} \leftarrow []$
    $n_{samples}$ ← number of samples in $\boldsymbol{X}$
    $n_{features}$ ← number of features in $\boldsymbol{X}$
if $f_{max}$ = None then
    $f_{max} \leftarrow n_{features}$
end if
for $i$ – 1 to $n_{trees}$ do
    Create new Decision Tree $\boldsymbol{T_i}$
    Perform bootstrap sampling to create $\boldsymbol{X}^{(i)}$ and $\boldsymbol{y}^{(i)}$
from $\boldsymbol{D}$
    Fit $\boldsymbol{T_i}$ on $\boldsymbol{X}^{(i)}$ and $\boldsymbol{y}^{(i)}$
    Add the trained tree to the list of trees $\boldsymbol{T} \leftarrow \boldsymbol{T} \cup \boldsymbol{T_i}$
end for
function PREDICTFOREST($\boldsymbol{X_{test}}$)
Initialize an empty list of prediction AllPreds ← []
for each $\boldsymbol{T_i}$ in $\boldsymbol{T}$ do
    $pred_i \leftarrow \boldsymbol{T_i}$ . predict($\boldsymbol{X_{test}}$)
    Add $pred_i$ to AllPreds
end for
final_preds ← Major of AllPreds
Return final_preds

Return List of trained trees $\boldsymbol{T}$
```

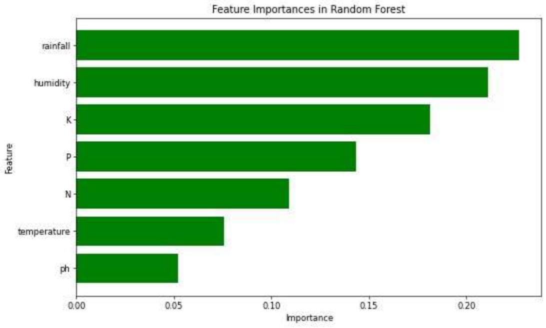

Figure 4: Feature importances in Random Forest

Support Vector Machines (SVM): A classification algorithm that finds the optimal hyperplane to separate different classes in the feature space. SVM is effective in high-dimensional spaces and is known for its robustness and accuracy.

D. Evaluation

The dataset was split into training and testing sets, with 80% of the data used for training and 20% for testing. The models were trained using the training set and evaluated using various performance metrics, including accuracy, precision, recall, and F1-score. Cross-validation techniques, such as k-fold cross-validation, were used to ensure that the models are robust and generalizable.

Training: The training process involved feeding the training data into the ML algorithms, adjusting the model parameters, and optimizing the model to minimize errors.

Algorithm 6: Training

```
Input: Training dataset $\boldsymbol{X_{train}}$, Training labels $\boldsymbol{Y_{train}}$,
       ML Algorithm
Output: Train model $\boldsymbol{M}$
Initialize model parameters
for each epoch
Feed $\boldsymbol{X_{train}}$ and $\boldsymbol{Y_{train}}$ into the ML algorithm
Preform forward pass
Compute loss
Perform backpropagation
Update model parameters using optimization algorithm

Return trained model $\boldsymbol{M}$
```

Evaluation: The trained models were evaluated on the testing set to assess their performance. Performance metrics, such as accuracy, precision, recall, and F1-score, were calculated to determine the models' effectiveness in crop recommendation.

Algorithm 7: Evaluation

```
Input: Testing dataset $\boldsymbol{X_{test}}$, Testing label $\boldsymbol{Y_{test}}$, Trained
       model $\boldsymbol{M}$
Output: Performance metrics (Accuracy, Precision,
       Recall, F1-score)
Predict $\boldsymbol{Y_{pred}}$ using $\boldsymbol{M}$ on $\boldsymbol{X_{test}}$
Calculate Accuracy
```

```
    $\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n}(Y_i = Y_{pred,i})$
Calculate Precision:
    $\text{Precision} = \frac{TP}{TP + FP}$
Calculate Recall:
    $\text{Recall} = \frac{TP}{TP + FN}$
Calculate F1-score:
    $\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$
**Return** Performance metrics
```

Hyperparameter Tuning: Hyperparameter tuning techniques, such as grid search and random search, were used to find the optimal hyperparameters for each ML algorithm, further improving the models' performance.

Algorithm 8: Hyperparameter Tuning

```
**Input:** Training dataset $\boldsymbol{X_{train}}$, Training labels $\boldsymbol{Y_{train}}$,
        ML algorithm, Hyperparameter space
**Output:** Optimal hyperparameter $\boldsymbol{\theta}^*$, Best model $\boldsymbol{M}^*$
Initialize best score to a very low value
**for** each combination of hyperparameters $\theta$ in
    hyperparameter space **do**
    Train model $\boldsymbol{M}$ with $\boldsymbol{\theta}$ on $\boldsymbol{X_{train}}$ and $\boldsymbol{Y_{train}}$
    Evaluate $\boldsymbol{M}$ using cross validation
    Compute score (e.g., validation accuracy)
    **if** score !best score **then**
        Update best score
        Set $\boldsymbol{M}^*$ to $\boldsymbol{M}$
        Set $\boldsymbol{\theta}^*$ to $\boldsymbol{\theta}$
    **end if**
**end for**
**Return** Optimal hyperparameter $\boldsymbol{\theta}^*$and best model $\boldsymbol{M}^*$
```

E. Deployment

The deployment of the Crop Recommendation System (CRS) was carried out using the MERN stack (MongoDB, Express.js, React.js, Node.js) along with Flask to integrate machine learning models and ensure seamless interaction with the user interface. This combination of technologies provides a robust and scalable solution for delivering real-time crop recommendations as shown in Figure 5, to farmers in Mizoram.

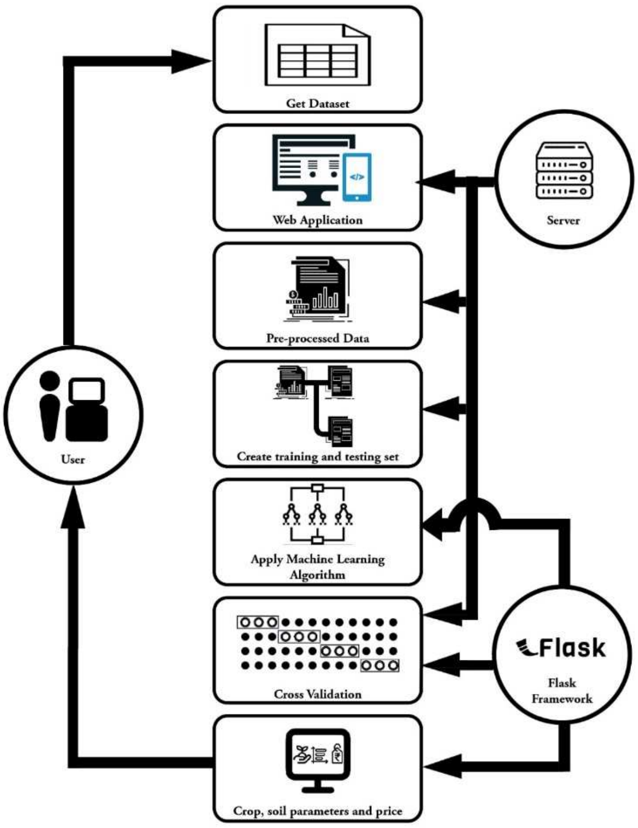
Get Dataset
Web Application
Server
Pre-processed Data
User
Create training and testing set
Apply Machine Learning Algorithm
Cross Validation
Flask
Flask Framework
Crop, soil parameters and price

Figure 5: Web Architecture of CRS

IV. RESULT AND DISCUSSION

The performance of the ML algorithms was evaluated based on the accuracy and robustness of their crop recommendations. The Random Forest algorithm outperformed other models, achieving an accuracy of 99% on the testing set. Decision Trees and SVM also showed promising results, with accuracies of 98% and 96%, respectively.

Decision Trees: The Decision Tree model provided interpretable and straightforward recommendations, making it easy for farmers to understand the rationale behind the recommendations. However, the model's performance was slightly lower compared to the Random Forest algorithm due to its tendency to overfit the training data.

Random Forest: The Random Forest algorithm demonstrated high accuracy and robustness, making it the most suitable model for crop recommendation in this study. The ensemble nature of Random Forest, which aggregates multiple decision trees, contributed to its superior performance.

Support Vector Machines (SVM): The SVM model showed good performance, especially in handling high-dimensional data. However, its complexity and computational requirements were higher compared to Decision Trees and Random Forest.

Table 1: Performance Metrics of different Models

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.99	0.99	0.99	0.99
Decision Tree	0.94	0.94	0.94	0.94
Support Vector Machines	0.96	0.96	0.96	0.95
Logistic Regression	0.94	0.94	0.94	0.94

Table 2: Performance metrics of different model after 5-fold Cross Validations

Model	Mean Accuracy	Mean Precision	Mean Recall	Mean F1-score
Random Forest	0.99	0.99	0.99	0.99
Decision Tree	0.98	0.98	0.98	0.98
Support Vector Machines	0.97	0.98	0.97	0.97
Logistic Regression	0.96	0.96	0.95	0.95

Impact on Agriculture Practices

The implementation of the CRS demonstrated its potential to significantly improve crop selection accuracy in Mizoram. By providing data-driven recommendations, the system can help farmers make informed decisions, leading to better crop yields and economic stability. The CRS promotes sustainable farming practices by recommending crops that are well-suited to the region's specific soil and climatic conditions.

The CRS can also serve as a valuable tool for agricultural extension officers and policymakers in Mizoram. By analyzing the recommendations and insights provided by the CRS, they can develop targeted interventions and support programs to assist farmers in optimizing their crop selection and improving productivity.

V. CONCLUSIONS

This study presents a Machine Learning-based Crop Recommendation System designed for the state of Mizoram. By leveraging soil properties, climatic conditions, and historical crop yield data, the CRS provides accurate and robust crop recommendations, promoting sustainable agricultural practices and improving farmers' livelihoods. The Random Forest algorithm is seen to demonstrate performance of 99% across all evaluation metrics. The results suggested that although the algorithm is suitable and provides superior performances, it can be seen that without any measures to avoid complications such as overfitting, the results may not generalize well in the future. But from what can be seen here, the results indicate that Random Forest is the best in providing crop recommendations, making it the most suitable model for this study.

The implementation of the CRS in Mizoram has the potential to revolutionize agricultural practices, increase productivity, and improve the economic stability of farmers. Future work will focus on addressing data limitations, incorporating additional features, and developing a user-friendly interface to enhance the system's effectiveness and accessibility.

Funding

Tribal Area Sub plan (TSP) scheme of IBITF- DST (Ref No: - IBITF/Note/TSP/SanctionLetter/2023-24/ 0252)

References

[1] Chen, J., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. University of California, Berkeley.

[2] Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.

[3] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.

[4] Government of Mizoram (2020). Agricultural Statistics. Department of Agriculture, Mizoram.

[5] Kumar, V., & Singh, P. (2015). Crop recommendation system for precision agriculture. Journal of Computer Science and Technology, 15(4), 345-352.

[6] Patil, P. B., & Biradar, B. (2016). A decision tree approach for crop selection in Karnataka, India. International Journal of Agricultural Science and Research, 6(3), 45-54.

[7] Murthy, S. R., & Prasad, R. (2017). Machine learning techniques for crop yield prediction and crop selection. International Journal of Engineering Research and Applications, 7(2), 67-72.

[8] Zhang, L., Wang, X., & Huang, Y. (2018). A hybrid model of support vector machine and decision tree for crop yield prediction. Agricultural Systems, 164, 101-110.

[9] Yadav, A. K., & Yadav, R. (2019). Machine learning applications in agriculture: A review. International Journal of Advanced Research in Computer Science, 10(1), 101-110.

[10] Sharma, P., & Singh, S. (2020). Crop recommendation using machine learning: A review. International Journal of Recent Technology and Engineering, 9(2), 234-240.