CROP RECOMMENDATION AND YIELD PREDICTION SYSTEM USING MACHINE LEARNING MODELS AND AGRO-ECONOMIC FACTORS

[1]Anirban Sasmal, [2]Sachin Dubey, [3]Aman Pal Singh

[1]B.Tech Student, [2]B.Tech Student, [3]Associate Professor
[1]School Of Computer Science and Engineering,
[1]Lovely Professional University, Phagwara, India

Abstract : Agriculture plays a big role in India's economy, but farmers often struggle to choose the best crops to grow. This is because the conditions vary a lot—like the type of soil, the weather, and the local economy. Most traditional systems for recommending crops focus only on environmental factors and don't take into account economic or other important factors. This study presents a new system that helps farmers decide which crops to grow. It combines machine learning with simple rules to give better crop suggestions, especially for areas in Punjab. The system uses data about the soil, the climate, past harvests, and economic factors like the Minimum Support Price (MSP) and government subsidies.
To make the recommendations more accurate and trustworthy, it uses a mix of Random Forest and XGBoost models along with a rule-based tool that helps decide which crops are suitable. Also, it uses SHAP to explain the decisions in a way that farmers can easily understand. The goal is to make farming decisions easier, increase crop production, and support sustainable farming by giving clear and helpful advice based on real data.

IndexTerms **- Crop Recommendation, Crop Yield Prediction, Explainable AI, SHAP, Precision Agriculture, Random Forest, XGBoost, Hybrid Model**

I. INTRODUCTION

Agriculture is still the biggest part of India's economy. It provides jobs to many people, ensures that people have enough food, and helps develop rural areas. A lot of people, either directly or indirectly, rely on farming for their daily life. Even with new technology, farmers often struggle to make smart choices about what crops to grow and how to get the best results from their land. One of the major problems is that we don't know what will happen with the economy or the environment. The production of crops is greatly affected by things like changes in rainfall, temperature, soil quality, and climate change. Along with environmental issues, economic factors such as market prices, Minimum Support Price (MSP), the cost of farming inputs, and government support also influence whether a certain crop is profitable. Regrettably, many farmers still use old methods or traditional knowledge, which may not work well for today's farming conditions. In recent years, machine learning and data-based techniques have become more common in agriculture, especially for predicting crop yields and suggesting which crops to grow. Many of these models use data like rainfall, temperature, soil nutrients, and regional information to decide which crops will do well. As mentioned in earlier studies, simple features like district, rainfall, and temperature are often used to make these systems easier for farmers to understand. While these methods are more accessible, they still have some issues. First, most systems only look at environmental factors and don't consider important economic factors that affect farmers' decisions. Second, many machine learning models work like black boxes, giving predictions without explaining why they were made. Because of this lack of transparency, farmers don't trust the system as much and are not very willing to use it.

II. NEED OF THE STUDY

The Indian economy is heavily dependent on agriculture as over 50% of the population relies on it and it is a major contributor to the GDP. Nevertheless, Indian farmers, particularly those in areas such as Punjab still have major problems when it comes to crop choice and yield maximization. Poor choice of crops usually results in low productivity, high cost of farming, soil erosion, and massive losses.
As climatic conditions change, so do soil conditions, rainfall is unpredictable, and market prices fluctuate, the traditional knowledge and the conventional practices of farmers are no longer adequate to ensure profitable and sustainable farming. Modern precision agriculture practices have become necessary, and the majority of farmers are yet to obtain scientific, data-oriented decision support systems. It has been found that many farmers have been experiencing crop failure or low yields in the past because of the lack of compatibility between the soil-climatic condition and the crop needs. Moreover, economic principles like Minimum Support Price (MSP), input prices, subsidies, and the market demand are also significant in determining a crop profitability, but they are hardly incorporated in current recommendation systems.

The intelligent, explainable and region-specific Crop Recommendation and Yield Prediction System is urgently required to help farmers make scientifically correct and economically viable decisions. The proposed study will address this gap by creating a hybrid machine learning model with rule-based expert knowledge and explainable AI methods; in the specific context of agro-climatic conditions of Punjab.

2.1 Population and Sample

The study population is major crops which are planted in the state of Punjab, India. Punjab is a major producer of food grain in India especially wheat, rice, maize and other horticultural crops. The universe of the study comprises all major crops grown in various districts of Punjab in the years 20002023.

This research has selected a representative sample of 12 major crops with reference to the economic significance, the area of cultivation and availability of data. Such crops are Wheat, Rice, Maize, Cotton, Sugarcane, Potato, Mustard, Barley, Gram, Lentil, Soybean and Sunflower. This was selected based on both Kharif and Rabi season to cover all the aspects of the cropping system in Punjab. Historical data on key agricultural areas in Punjab (district-wis) have been used to ensure geographical representation.

2.2 Data and Sources of Data

In this research, the secondary data has been gathered through trusted government and public resources. The primary sources of data are:

Indian Crop Production Data (1950-2020).
Punjab Government Agriculture Portal and Department of Agriculture, Punjab.
Soil Health Card information and farm surveys.
Indian Meteorological Department Meteorology (IMD) data.
Economic data such as Minimum Support Price (MSP) of Commission for Agricultural Costs and Prices (CACP).
Information on market price and subsidies of different government reports.

The data includes a time-series and cross-sectional data on the parameters of soil, climatic variables, district data, and economic indicators of the chosen crops during the period of over twenty years. Each and every data is pre-processed and integrated to come up with a complete dataset that can be used in the development of machine learning models.

2.3 Theoretical framework

Precision Agriculture, Machine Learning, and Agro-Economic Decision Theory are the theoretical frameworks underpinning this study.

A dependent variable in the research is the Recommended Crop (Crop Type) and its predicted Yield. Independent variables are put into four broad categories:

Soil Parameters- pH, Nitrogen (N), Phosphorus (P), Potassium (K), soil texture.
Climatic Conditions - Rain, Temperature, Humidity.
Geographical Factors — District/Region
Economic Factors Minimum Support Price (MSP), Market Demand, Subsidies, Cost of inputs.

The proposed Hybrid Model in the study utilizes the predictive capabilities of ensemble machine learning algorithms (Random Forest and XGBoost) in addition to a Rule-Based Crop Suitability Index (CSI) based on domain knowledge. The system is also transparent and makes the final recommendation trustworthy to the farmers as SHAP (SHapley Additive Explanations) values explain how each factor contributes to the final recommendation.

This framework makes sure that the recommendation is not only agronomically viable but also economically viable to the farmer.

III. RESEARCH METHODOLOGY

The methodology section is a systematic plan and scientific approaches that were used in conducting this research. It outlines the study research design, universe and sample of the research, data gathering sources, research variables, proposed framework, model development and the methods of evaluation. The information is listed as follows:

3.1 Population and Sample

The study population will be made up of all the major crops cultivated in various agro-climatic regions of Punjab, India. Punjab is the state of India which is the most developed in terms of agriculture and produces a lot to the national food grain basket. Kharif and Rabi seasons in the state are characterized by a large number of crops grown.

In this study, a sample of 12 significant crops has been selected cautiously regarding the economic importance, cultivated area and availability of data. Crops that have been chosen include: Wheat, Rice, Maize, Cotton, Sugarcane, Potato, Mustard, Barley, Gram, Lentil, Soybean and Sunflower. Such crops constitute over 85 percent of the total cropped area in Punjab. To cover a sufficient geographical area and variety of soil and climatic conditions, district-level data of large agricultural districts have been incorporated.

3.2 Data and Sources of Data

The paper utilizes mainly secondary data sources which have been obtained through various credible sources. Primary sources are:

Indian Crop Production Data (1950-2020).
Punjab State Agriculture Department and Punjab Government Agriculture Portal.
Soil Health Card Scheme database.
India Meteorological Department (IMD) of rainfall and temperature.
Minimum Support Price (MSP) Commission of Agricultural Costs and Prices (CACP).
Market trends directorate of Marketing and Inspection.
Government reports and research papers.

The time-series and cross-sectional data have been gathered over the period 2000-2023. The dataset incorporates soil parameters, climatic variables, economic variables as well as historical yield data of the selected crops.

3.3 Theoretical framework

The study variables include dependent and independent variables. The research has employed a pre-specified approach to selecting variables utilizing large bodies of literature and practical applicability in precision agriculture.
The Recommended Crop with its Predicted Yield is the dependent variable of the study. Recommended crop is the most appropriate crop in a given set of conditions and predicted yield is useful in estimating the potential productivity of the farmers.

The independent variables are categorized into four main groups:

1. Soil Parameters
Crop production is based on soil. The soil parameters that are important in this study are soil Ph, Nitrogen (N), Phosphorus (P) and Potassium (K). The pH of soil influences the nutrient availability to plants and the NPK values directly influence the growth and yield of crops. A number of studies have found that there is a close correlation between nutrient status of soils and suitability of crops.

2. Climatic Factors
Crop growth and productivity are dependent on climate. The average rainfall, temperature, and humidity are the main climatic variables included in this study. In rain-fed and irrigated regions of Punjab, rainfall is an issue of particular concern. The growth cycle and phenological stages of crops are influenced by temperature. Numerous scientists have pointed out that variations in climatic parameters greatly influence crop production and appropriateness.

3. Geographical Factors
Geographical information, which is in form of district information, describes the regional differences in soil type, availability of water and local climate. Punjab comprises a varied agro-climatic zones; hence, location-specific recommendations require a district-wise analysis.

4. Economic Factors
Ignoring economic factors is one of the biggest shortcomings of the currently existing crop recommendations systems. Minimum Support Price (MSP), market demand, and government subsidies are some of the economic variables in this study. It is presumed that although a crop can be agronomically viable, it can not be profitable to the farmer when the market price is low or when the cost of inputs is high. Sustainable farming decisions are dependent on economic viability.
The increase in MSP and enhanced market demand is likely to have a positive impact on the ultimate score on the recommendation because farmers will eventually strive towards profitable crops. Past research has demonstrated that incorporation of economic aspects can greatly enhance the practical value of recommendation systems.

The current research suggests a Hybrid Theoretical Framework, which combines:

Pattern recognition and prediction algorithms (Random Forest and XGBoost) based on machine learning.
Practical agricultural feasibility rule-based expert knowledge.
Model transparency with explainable AI (SHAP).

This combination strategy overcomes the shortcomings of either pure machine learning models (black-box nature) and pure rule-based systems (lack of learning capability). The framework will make the end result of the recommendation accurate and interpretable in addition to being trustworthy and economical to the Punjab farmers.
The conceptual model presumes that environmental suitability and economic profitability are functions of suitability of a crop. Incorporating all these dimensions, the proposed system will assist farmers in making data-driven and holistic decisions on the selection of crops.

3.4 Statistical tools and econometric models

This paragraph expounds on the correct statistical methods and machine learning algorithms that have been used to transform raw agricultural data into accurate crop advice and yield forecasts. The analytical paradigm is developed in a systematic way starting with simple statistical exploration to more sophisticated modeling, hybridization, interpretability and rigorous validation. The information is displayed in the following way:

3.4.1 Descriptive Statistics

The descriptive statistics have been very much applied in order to obtain preliminary information about the data. The study calculated important values of all the variables (such as minimum value, maximum value, mean, median, standard deviation, variance, skewness, and kurtosis) of all the variables (soil pH, Nitrogen (N), Phosphorus (P), Potassium (K), rainfall, temperature, MSP, and historical crop yield).

With these statistics, it is possible to know the central tendency, dispersion and shape of the data distribution. To see an example, when the standard deviation in rainfall is high this means that there is a lot of variability in the amount of rainfall received in different years and in different districts which poses a significant challenge in Punjab agriculture. In the same manner, MSP values can be analyzed to indicate the economic appeal of various crops.

Histograms, box plots and scatter plots were also created to determine outliers, skewness of the data and potential relationships amongst the variables. These statistical properties are imperative as the performance of machine learning models is very sensitive to the underlying data distribution. The descriptive statistics were also used to form a basis on the decisions that were to be made in data preprocessing.

3.4.2 Data Preprocessing and Feature Engineering

Agricultural data are usually rough, incomplete and not consistent. Thus, a number of preprocessing steps were carried out. Any missed values were addressed with the help of suitable imputation methods (mean imputation of normally distributed variables and median of skewed variables). The Interquartile Range (IQR) method was used to detect outliers which were treated with a lot of care in order to prevent loss of valuable information.

All numerical variables were Min-Max scaled and standardized by means of Z-score normalization to have equal contribution of each variable in the process of model training. Labels Encoding and One-Hot Encoding methods were used to transform categorical variables (district names and soil types) into numerical expressions.

The engineering of features was conducted to build new meaningful variables like rainfall deviation to normal, soil nutrient balance index and economic profitability ratio. Correlation Analysis, Mutual Information, Recursive Feature Elimination (RFE) methods were used to select the most influential predictors, and to simplify the model.

3.4.3 Machine Learning Models

Both crop classification (recommendation) and regression (yield prediction) were done using supervised machine learning techniques.

3.4.3.1 Random Forest

Random Forest is an ensemble learning algorithm built on bagging that constructs multiple decision trees and combines the predictions of the trees. The reason behind its choice is its resistance to overfitting, capability to deal with high-dimensional data, and intrinsic feature ranking of importance. Random Forest in this study was tuned with the help of hyperparameters which were number of trees, maximum depth and minimum samples per split via Grid Search Cross-Validation.

3.4.3.2 XGBoost (Extreme Gradient Boosting)

XGBoost is a boosting algorithm that has excellent predictive power and computational efficiency. It progressively trains weak learners and optimizes a loss function by gradient descent. XGBoost was selected due to its great ability to handle missing values, the use of regularization to avoid overfitting, and speediness. Figure 3: Hyperparameter optimization in the grid search and the Bayesian optimization techniques was done.

The two models were trained using the preprocessed dataset where crop type was used as the target variable to be recommended and yield was used as the target variable to be predicted.

3.4.4 Rule-Based Expert System

Possibly, pure machine learning models provide impractically implementable recommendations. To solve this a deep-rooted rule-based expert system was formulated with consultation of knowledge of the agricultural domain. Critical factors were considered to develop a Crop Suitability Index (CSI) that includes the pH range, minimum water requirements, temperature tolerance, and the growing degree days of each crop.

As an example, rice cannot be cultivated in areas with a very low rainfall and less irrigation facilities, and wheat grows well in neutral pH soils. These regulations serve as a safety net and part of the old farming expertise is integrated into the new AI framework.

3.4.5 Hybrid Model Approach

A hybrid approach was designed to integrate the advantages of the machine learning-driven and knowledge-driven rule-based systems. The weighted ensemble approach is defined as:

Final Recommendation Score = 0.8 ML Model Output + (1 -0.8) Rule-Based CSI Score.

and α is an optimism weighting value, which is established by validation experiments (usually between 0.6 and 0.8). Such an integrated approach enhances accuracy and practicality of the system.

3.4.6 Explainable AI (XAI) - SHAP Analysis.

To overcome the black-box nature of ensemble models, SHAP (SHapley Additive Explanations), a game-theoretic method has been applied. SHAP values indicate the amount of each input feature to the output of the model to each individual prediction.
This method enables the system to produce human comprehensible explanations like: Wheat is suggested due to a proper level of rainfall (+0.32) and proper soil Nitrogen content (+0.25), and high MSP (+0.18). These kinds of explanations are essential in the development of trust between farmers and extension workers that might not be technical in nature.

3.4.7 Evaluation and validation techniques of models.

The evaluation metrics that were used to evaluate the performance of the developed models were multiple evaluation metrics that are applicable to the classification and regression tasks: Accuracy, Precision, Recall, F1-Score, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE).
To have a sound assessment, the following validation strategies were used:

K-Fold Cross-Validation
Stratified K-Fold Cross-Validation
Spatial Cross-Validation (training on part of the districts and testing on unknown districts)
Temporal Validation (training using past and testing using recent years)

These methods guarantee that the proposed system will be reliable in various geographical areas and in the fluctuating climatic conditions over time.

IV. RESULTS AND DISCUSSION

4.1 Findings of Descriptive Statistics of Study Variables

Descriptive statistics was calculated to get familiar with the common features and distribution of the variables applied in the research. These are soil parameters, climatic, economic variables and yield of crops.

Table 4.1: Descriptive Statistics of Study Variables

Variable	Minimum	Maximum	Mean	Std. Deviation	Jarque-Bera Test	Sig.
Soil pH	5.80	8.90	7.45	0.68	4.872	0.087
Nitrogen (kg/ha)	85	285	168.45	52.34	3.214	0.201
Phosphorus (kg/ha)	12	68	34.76	14.82	5.678	0.059
Potassium (kg/ha)	95	320	198.67	61.45	2.945	0.229
Rainfall (mm)	320	1250	682.45	218.76	6.124	0.047
Temperature (°C)	12.5	38.4	24.85	6.78	3.876	0.144
MSP (₹/quintal)	1250	3200	1985.67	542.30	4.567	0.102
Crop Yield (kg/ha)	850	6850	3124.56	1245.78	5.892	0.053

Table 4.1: Descriptive Statistics

Table 4.1 shows the values of mean, standard deviation, minimum and maximum values, Jarque-Bra test statistic, and its value of significance of major variables of the study. The descriptive statistics shows that the mean soil pH of the sampled districts of Punjab is 7.45, which is within the neutral range and is usually favorable to most crops. The mean of nitrogen content is 168.45 kg/ha with standard deviation of 52.34, which implies that the soil fertility among the various districts is highly varied.
The mean precipitation is 682.45 mm and the standard deviation of 218.76 mm indicating a significant range of variation in precipitation which is a major limitation of crop scheduling in Punjab. The average Minimum Support Price (MSP) of the sampled crops is 1985.67 per quintal with a large deviation in the price based on the nature of the crop. The mean crop yield is 3124.56 kg/ha and the standard deviation is high at 1245.78 kg/ha indicating that there is a significant variation in the productivity, as a result of different soil, climate and management.

The Jarque Bera test that is used to test normality of data is indicated in column 6 in Table 4.1. The normal distribution hypotheses are:
H 0: The data is normally distributed.
H 0: The data is not normally distributed.
Most variables can not reject the null hypothesis of normality at a 5% level of significance. It would indicate that most of the variables are normally distributed. Mild deviation of rainfall and crop yield is however out of the norm, which is quite common in agricultural data owing to the external environmental factors.

The descriptive statistics indicate that the data is spread extensively around their respective means most notably in the example of rainfall, nutrient levels, and crop yield. Such variability explains why a strong, data-driven hybrid model should be developed, which will be able to manage these complexities and offer sound crop advice. The existence of plausible variability in the economic parameters such as MSP also depicts the significance of incorporating economic parameters in the recommendation system in achieving profitable decisions.

V. ACKNOWLEDGMENT

The authors would like to thank everyone who supported and helped to complete this research work to be a successful one.
To begin with, we wish to express our extreme gratitude to our supervisors and faculty members at Lovely Professional University, Jalandhar, Punjab, who have been of great help, accompanied with encouragement and helpful advice during the entire research. Their knowledge in agriculture and machine learning assisted us in fine-tuning the aims and methodology of this research.
We also owe the Department of Agriculture, Government of Punjab, and other agricultural research institutions the credit of giving us access to necessary datasets on the health of the soil, production of crops, and weather data. A special mention should be done to the Commission of Agricultural Costs and Prices (CACP) which has made Minimum Support Price (MSP) data easily accessible.
We would also like to acknowledge the creators and contributors of open-source machine learning frameworks like Scikit-learn, XGBoost, and SHAP, which have played a crucial role in developing and explaining our hybrid model.
We are particularly indebted to the farmers and the agricultural extension officers in Punjab who, having been the indirect source of the inspiration of this work, have been its inspiration in real-life situations and through direct feedback when we held informal discussions with them. Their real-world experience assisted us in creating more farmer-focused and practical recommendation system.
Lastly, we agree that we have always had moral support and encouragement of family members and friends without which this research would not have been possible.

VI. REFERNCES

[1] D. Israni, M. Tolani, K. Masalia, T. Khasgiwal, and M. R. Edinburgh,
"Crop Yield Prediction and Crop Recommendation System,"
SSRN Electronic Journal, 2022.

[2] S. Babu et al.,
"A software model for precision agriculture for small and marginal farmers,"
IEEE Global Humanitarian Technology Conference, 2013.

[3] S. Veenadhari, B. Misra, and C. D. Singh,
"Machine learning approach for forecasting crop yield based on climatic parameters,"
International Conference on Computer Communication and Informatics, 2014.

[4] P. S. Nishant, P. Sai Venkat, B. L. Avinash, and B. Jabber,
"Crop yield prediction based on Indian agriculture using machine learning,"
International Conference for Emerging Technology (INCET), 2020.

[5] S. Pudumalar et al.,
"Crop recommendation system for precision agriculture,"
International Conference on Advanced Computing (ICoAC), 2017.

[6] R. Kumar, M. P. Singh, P. Kumar, and J. P. Singh,
"Crop selection method to maximize crop yield rate using machine learning technique,"
ICSTM, 2015.

[7] A. Savla et al.,
"Survey of classification algorithms for yield prediction in precision agriculture,"
ICIIECS, 2015.