World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Crop recommendation and yield prediction using machine learning algorithms

Sundari V, Anusree M, Swetha U and Divya Lakshmi R *

Department of Computer Science and Engineering, Meenakshi Sundararajan Engineering College, Anna University, Chennai, India.

World Journal of Advanced Research and Reviews, 2022, 14(03), 452–459

Publication history: Received on 13 May 2022; revised on 16 June 2022; accepted on 18 June 2022

Article DOI: https://doi.org/10.30574/wjarr.2022.14.3.0581

Abstract

Agriculture is the foundation of many countries' economies, particularly in India and Tamil Nadu. The young generation who are new to farming may confront the challenge of not understanding what to sow and what to reap benefit from. This is a problem that has to be addressed, and it is one that we are addressing. Predicting the proper crop and production will aid in making better decisions, reducing losses and managing the risk of price fluctuations. The existing system is not deployed, unlike ours, which is done by applying classification and regression algorithms to calculate crop type recommendations and yield predictions. Agricultural industries must use machine learning algorithms to anticipate the crop from a given dataset. The supervised machine learning technique is used to analyse a dataset in order to capture information from multiple sources, such as variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments, and so on. A comparison of machine learning algorithms was conducted in order to identify which algorithm was more accurate in predicting the best harvest. The results show that the proposed machine learning algorithm technique has the best accuracy when comparing entropy calculation, precision, Recall, F1 Score, Sensitivity, Specificity, and Entropy.

We have ensured that our proposed system accomplishes its job effectively by projecting the yield of practically all types of crops grown in Tamil Nadu, relieving some of the burden from their shoulders as they enter a new business.

Keywords: Supervised Machine Learning Approach; Classification and Regression Models; Precision; Linear Regression; Logistic Regression; Decision Tree and Random Forest

1. Introduction

Agriculture is an important source of income for many people in underdeveloped countries. Several technologies, conditions, practices, and civilizations have all influenced agricultural expansion in recent years. Furthermore, the use of information technology may alter the state of decision-making, allowing farmers to produce the best results. Data mining techniques connected to agriculture are employed in the decision-making process. The process of extracting the most important and relevant information from a large number of datasets is known as data mining. As agriculture involves a variety of data, such as soil data, crop data, and weather data, we now employ a machine learning approach designed for crop or plant yield prediction. Machine learning techniques are efficiently used to propose a crop recommendation and yield prediction system.

* Corresponding author: Divya Lakshmi R
Department of Computer Science and Engineering, Meenakshi Sundararajan Engineering College (Affiliated to Anna University) Chennai, India.

2. Related work

2.1. Prediction of Land Suitability for Crop Cultivation

Machine learning techniques are used to identify the best crops for a given region, greatly benefiting farmers in crop prediction. The feature selection (FS) facet of machine learning is a critical component in the selection of relevant features for a given region, and it keeps the crop prediction process up to date. To choose acceptable features from a data collection for crop prediction, this paper introduces a novel FS approach called modified recursive feature elimination (MRFE). Using a ranking algorithm, the suggested MRFE technique chooses and ranks salient features. The MRFE method selects the most accurate features, while the bagging methodology aids in properly predicting an appropriate crop, according to the testing results. Various metrics like as accuracy (ACC), precision, recall, specificity, F1 score, area under the curve, mean absolute error, and log loss are used to assess the performance of the proposed MRFE approach. The MRFE methodology outperforms other FS methods with 95% accuracy, according to the performance analysis [2].

2.2. Ascertaining the Fluctuation of Rice Price using Machine Learning

The purpose of this research is to use a machine learning approach to anticipate rice prices in Bangladesh. The price was predicted using data from Bangladesh's Ministry of Agriculture website. Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naive Bayes, Decision Tree, and Random Forest were among the machine learning techniques used to make this prediction. All of these algorithms are compared to see which one delivers the greatest results. Despite the fact that all five algorithms performed similarly, the random forest method stood out as the best. As a result, they employed the Random Forest to forecast rice prices. This projection could lead to a situation in which the Bangladesh government and people know how much rice they need to plant in order to meet everyone's food needs. Based on the findings obtained by the mentioned algorithms, they forecasted the price of rice, whether it is reasonable, low, or exorbitant. The fundamental restriction of their work was a lack of real data, and instead of gathering data from the entire country, they simply collected data from a small portion of it. This study was hampered by a lack of real data, and rather than collecting data from across the country, they relied solely on data from Dhaka to create a dataset. The inclusion of new data in their dataset is a top priority for future work [3].

2.3. Forecasting Crop productivity

This research provides a realistic and user-friendly yield prediction system for Indian farmers, outlining the shortcomings of current methods and their practical use in yield prediction. A smartphone application is used to link farmers to the proposed system. Users may utilize a variety of tools in the mobile application to choose a crop. GPS assists in determining the user's position. As input, the user enters the area and soil type. Machine learning algorithms may be used to identify the most lucrative crop list or to forecast crop yields for a user-selected crop. The built-in predictor technology assists farmers in predicting crop yields. The built-in recommender system helps the user to explore the various crops and their yields in order to make better informed judgments. Various Machine Learning techniques, including SVM, Artificial Neural Network (ANN), Random Forest, Multivariate Linear Regression (MLR), and KNN, are applied and tested on datasets from Maharashtra and Karnataka to forecast crop productivity. With 95% accuracy, the random forest is the best among the set of typical algorithms used on the given datasets. In addition, the system recommends the optimal time to apply fertilizers to increase production. Future study will focus on periodically updating datasets in order to generate reliable predictions, and the processes can be automated. Another feature that will be implemented is the ability to offer the appropriate fertilizer for each crop [5].

2.4. Crop Yield Estimation

Machine learning (ML) is an important approach for obtaining practical and real-world solutions to crop yield difficulties. ML can predict a target/outcome from a set of predictors using Supervised Learning. To obtain the required results, an appropriate function based on a collection of variables must be created, which will map the input variable to the desired output. Crop yield prediction is predicting a crop's production based on previous data that includes parameters like temperature, humidity, pH, rainfall, and crop name. The Random Forest method is used to forecast the best crop yield as an output. In the agriculture field, the crop yield prediction is mostly appropriate. The more increase in accuracy results in more profit to the crop yield. It will achieve the most accurate crop prediction possible. The random forest technique is used to find the optimum crop yield model with the fewest number of models. The proposed method aids farmers in gaining an understanding of the demand for and pricing of various crops. This approach will cover the widest range of crops, making it extremely valuable for predicting crop yields in the agriculture industry. Accurate forecasting of different specified crops across different districts will benefit Indian farmers [1].

3. Proposed system

In our proposed crop recommendation and yield prediction system, we employed various machine learning techniques. Machine learning a branch that focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy, from that we have used classification and regression algorithms to predict the crop type and yield produced Quintal/Hectare respectively [1]. The inputs would be state name, district name, season, area, rainfall, average humidity, and mean temperature. After data preprocessing and data visualization methods, a separate dataset with past data, other than user input, is given to do the training and testing.

The classification algorithms and regression algorithms, three algorithms each are used to find the best accuracy calculated and the results are displayed to the user. The classification algorithms, such as logistic regression, decision tree, and random forest, weigh the input features so that the output separates one class into positive values and the other into negative values, while the regression algorithms, such as linear regression, decision tree, and random forest, predict the output values based on input features from the data fed into the system. The best accuracy is ranked based on the accuracy computed by each of those algorithms, and the results are displayed to the user.

The output is then displayed using flask, a small and lightweight Python web framework that provides essential tools and capabilities to make web application development simple. We prepared separate model files for the algorithms that produced the best results for crop type and yield prediction, and used those to display the results in Flask.

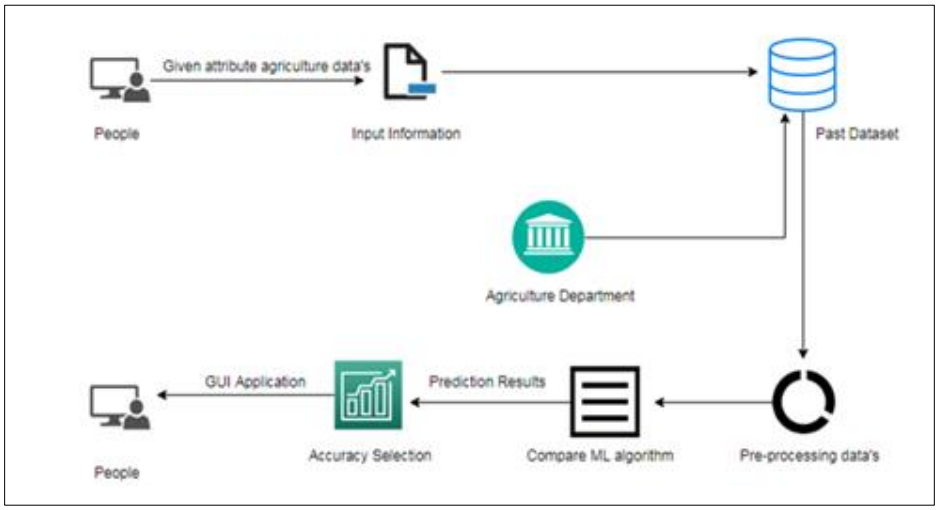

Figure 1 System Architecture

4. Module description

Python programming and a variety of libraries is used to build the system. The flask library is used to create the frontend. The model is trained and tested using a variety of machine learning models. We utilize a supervised machine learning approach because we have a labeled dataset with a set of crop cultivation features.

The dataset is first gathered. The data is then preprocessed to make it easier to use in training. The data is divided into two categories: training and testing. Each machine learning model is trained using training data. The accuracy of each model is then calculated and compared using testing data to discover the best model for the dataset. A user interface is built which will display the prediction using the stored training data.

4.1. Data pre-processing

The dataset is collected in this initial module with information such as state name, district name, season, crop, area, rainfall, and so on. Data preparation is carried out in order to convert the raw data into a readable and acceptable format. This is done to improve the data's quality.

The first step in data preprocessing is data cleaning, which involves locating duplicate values in the dataset, deleting null values, and removing unwanted values. The data is then categorized before being encoded into numbers so that the

model can understand and extract valuable information, as machine learning algorithms only take numeric variables. A categorical data encoding approach is employed for label encoding in this operation.

4.2. Data visualization

Data visualization is used to create a graphical representation of data in order to have a better understanding of the dataset. The tabular data is converted into a visual representation, such as a scatter plot, bar chart, heatmap plot, and so on. Hidden data patterns can be discovered by visualizing the data. The data is then examined to see if there are any clusters, and if so, whether they are linearly separable/overlapped, and so on. We can simply rule out models that aren't fit for such data based on this preliminary study, and we will implement only such models that are, saving time and money. This method enables us to quickly comprehend the dataset and construct machine learning models.

A test harness data is created before selecting an appropriate machine learning model. The test harness is the data against which an algorithm is trained and tested, as well as the performance metric that will be used to evaluate its performance. The dataset is separated into training and testing datasets for this purpose. 70% of the time is used for training and 30% for testing. The model will be trained using both input and output data. The input and output of testing data will be determined by prediction.

4.3. Choosing machine learning models

In this module, Model selection is done, it is the process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset. Before choosing a machine learning model, it is found which type of the dataset we are having. For our problem statement we are using labeled data so supervised machine learning algorithms are chosen for training the dataset. Supervised learning uses a training set to teach models to yield the desired output [1]. This training dataset includes inputs and correct outputs, which allow the model to learn over time.

For crop recommendation classification supervised machine learning algorithms are used as we can categorize the output into classes. Classification models have the task of approximating the mapping function from input variables to discrete output variables. The main goal is to identify which class/category the new data will fall into. For crop yield prediction regression supervised machine learning algorithms are used as the prediction gives us an output value which is continuous [4]. Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables.

4.4. Comparing machine learning algorithms

The chosen machine learning algorithms are applied to the prepared data to solve the problem. The training dataset is applied for each of the algorithms one by one and trained. Using each of these algorithms, crop yield prediction and prediction of crop name is done. For crop name prediction classification algorithms are used and for crop yield prediction regression algorithms are used. Then, using testing data evaluation of each model is done. It is important to compare the performance of multiple different machine learning algorithms consistently. Each model will have different performance characteristics.

Using resampling methods like cross validation, it is estimated for how accurate each model may be on unseen data. Then, using various performance metrics the performance of each algorithm is evaluated. A classification report is generated by finding precision, recall, F1 score [2]. By comparing the results the best algorithm that can be used for this model is found.

4.5. Storing the trained data

The best machine learning model which gives a higher accuracy for the dataset is selected. Using this model, the dataset is applied and training is made. The trained data is exported and stored in file for using during the prediction.

4.6. Deployment using flask

A user interface is built which displays crop prediction results. During the prediction the trained data is loaded. Users can input agricultural parameters. The model which we built will take the input and make predictions. The predicted result is displayed to the user through the user interface that is built.

5. Machine learning models

We use supervised machine learning models to build our system. The types of supervised machine learning models used are:

- Classification algorithms – for Crop recommendation
- Regression algorithms – for Yield Prediction

5.1. Logistic regression

When the goal variable is categorical, a supervised learning approach called logistic regression is applied. The target variable in Logistic Regression is categorical, therefore we must restrict the range of predicted values. Multiclass logistic regression is also known as multinomial logistic regression. We used this algorithm to predict more than two classes of crop types.

5.2. Decision tree

Decision Tree algorithm is a type of machine learning algorithm in which the data is continually split according to a parameter. The decision tree has an if-else structure, with each node splitting into two or more branches using only one independent variable. It doesn't matter if the independent variable is categorical or continuous. For categorical variables, the categories are used to determine the node split and for continuous variables, the algorithm generates various threshold values that serve as the decision-maker. We applied Decision tree classifier for crop name prediction and Decision tree regressor yield prediction.

5.3. Random forest

A random forest is a supervised machine learning system that uses decision tree algorithms to construct it. This algorithm determines the outcome based on decision tree predictions. It predicts by averaging the output of various trees. The precision of the result improves as the number of trees grows. It reduces dataset overfitting and improves precision. We applied Random forest classifier for crop name prediction and Random forest regressor yield prediction.

5.4. Linear regression

Linear regression analysis is a statistical technique for predicting the value of one variable based on the value of another. The dependent variable is the variable you want to predict. The independent variable is the one we are using to predict the value of the other variable. We applied linear regression classifier for crop name prediction and Linear regression yield prediction.

6. Results

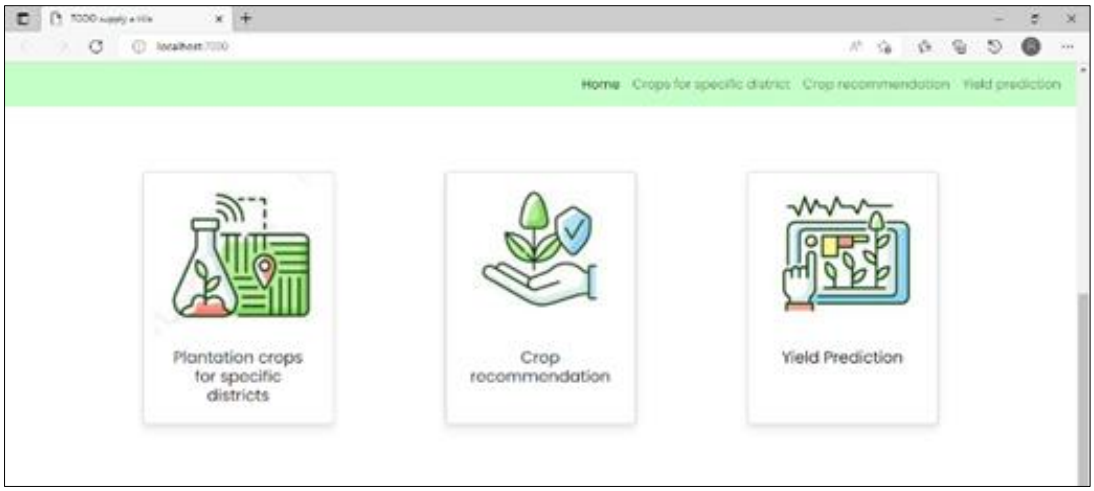

Figure 2 Home page

Figure 3 Plantation Crops for specific District

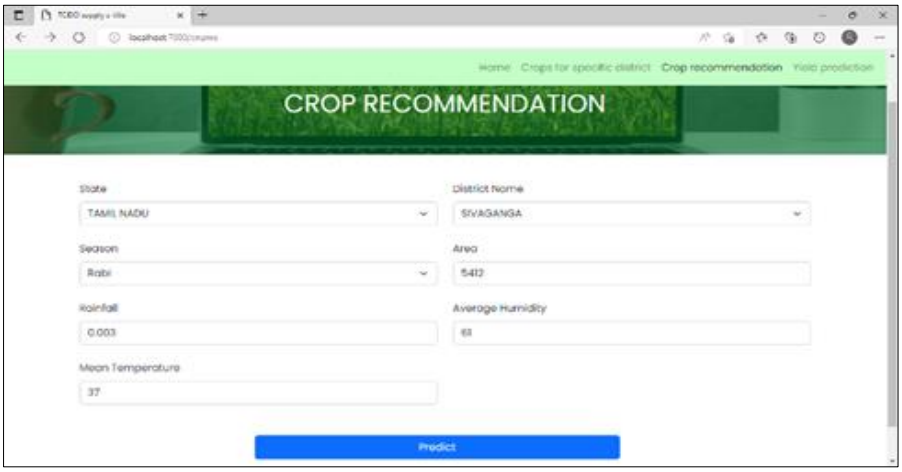

Figure 4 Crop Recommendation Form

Figure 5 Crop Recommendation Result

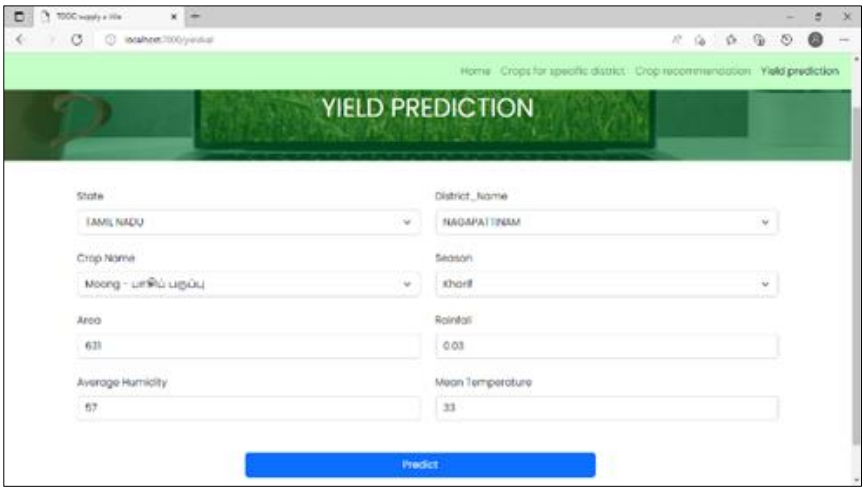

Figure 6 Yield Prediction form

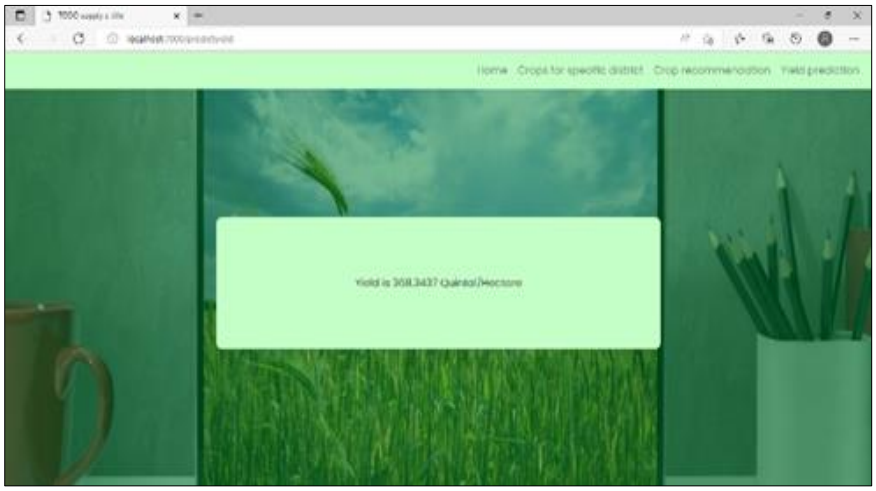

Figure 7 Yield Prediction result

7. Conclusion

A crop recommendation and yield prediction system has been developed successfully using Supervised Machine Learning Approach. The analysis began with data preprocessing and cleaning, followed by exploratory analysis using an agricultural dataset. After that, we used the dataset to train multiple machine learning models, and made various evaluation processes to find the best algorithm. As a result of evaluation, we use Decision tree classifier for crop recommendation and Random Forest regressor for yield prediction as they give best accuracies. Our trained algorithm can predict crops based on the specified characteristics, as well as crop yield rate. As this system will cover the widest range of crops, farmers will be able to learn about crops that have never been farmed before and will be able to see a list of all possible crops, which will aid them in deciding which crop to plant. We intent to help people who are planning to invest in farming without any prior knowledge about farming and how much they can make a profit out of it. It is also to help people who are new to farming and try to make their way in learning the practice that has been followed for generations. Thus, our approach can assist farmers in Tamil Nadu, particularly newcomers, in deciding which crop to produce by predicting crop and yield based on local climatic circumstances.

The proposed system is developed as a website. In future we may try to develop our system as a mobile application which makes the user use this application even more user friendly. As a future enhancement we may try to train the model using neural network algorithms. In our system we are not using neural networks as with the dataset now we are having good accuracy results with machine learning algorithms. We expect when increasing the features of input

and increasing the size of the dataset, training the model with neural networks would produce even more efficient results.

Compliance with ethical standards

Acknowledgments

This research work is done by Anusree M, Divya Lakshmi R and Swetha U under the supervision of Sundari V, under the Department of Computer Science and Engineering, Meenakshi Sundararajan Engineering College, Chennai, India.

Disclosure of conflict of interest

The authors declare no conflict of interest.

References

[1] Jeevan Nagendra Kumar Y, V Spandana, VS Vaishnavi, K Neha, VGRR Devi. Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector. 5th International Conference on Communication and Electronics Systems. 2020;736-741

[2] Mariammal G, A Suruliandi, SP Raja, E Poongothai. Prediction of Land Suitability for Crop Cultivation Based on Soil and Environmental Characteristics Using Modified Recursive Feature Elimination Technique With Various Classifiers. IEEE Transactions on Computational Social Systems. 2021;8(5):1132-1142

[3] Mehedi Hasan Md, Muslima Tuz Zahara, Mahamudunnobi Sykot, Arafat Ullah Nur, Mohd Saifuzzaman, Rubaiya Hafiz. Ascertaining the Fluctuation of Rice Price in Bangladesh Using Machine Learning Approach. 11th International Conference on Computing, Communication and Networking Technologies. 2020;1-5

[4] Rakesh Kumar, MP Singh, Prabhat Kumar, JP Singh. Crop Selection Method to maximize crop yield rate using machine learning technique. International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials. 2015;138-145.

[5] Shilpa Mangesh Pande, Prem Kumar Ramesh, Anmol Anmol, BR Aishwarya, Karuna Rohilla, Kumar Shaurya. Crop Recommender System Using Machine Learning Approach. 5th International Conference on Computing Methodologies and Communication. 2021;1066-1071.