JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

MACHINE LEARNING BASED CROP YIELD MANAGEMENT

Nadheer Ahmed B
Assistant Professor
Computer Science and Engineering
Aalim Muhammed Salegh College of Engineering
Chennai, India
b.nadheerahmed@aalimec.ac.in

Abdullah Syed Ismail[1], Adnan Shariq P[2], Hisham Ismail VA[3]. Mohamed Irfaan I[4]
Students
Computer Science and Engineering
Aalim Muhammed Salegh College of Engineering
Chennai, India
110119104002@aalimec.ac.in[1], 110119104004@aalimec.ac.in[2], 110119104021@aalimec.ac.in[3], 110119104038@aalimec.ac.in[4]

Abstract—Farmers face several challenges when growing crops like uncertain irrigation, poor crop selection, etc. Especially in India, a major fraction of farmers faces challenges to select appropriate crops and fertilizers. Moreover, crop failure due to disease causes a significant loss to the farmers, and has a huge impact on the economy. This paper aims to leverage the power of machine learning techniques to optimize crop yield and enhance agricultural productivity. To provide practical recommendations, the system incorporates a user-friendly interface that allows farmers to input their specific parameters and receive customized suggestions. These recommendations may include suitable crop varieties, fertilizer selection strategies, and cure for plant diseases, all tailored to maximize crop yield based on the machine learning models.

I. INTRODUCTION

Our paper contains a combination of multiple modules that will help farmers to yield better crops. Our system has three modules (Crop recommendation, Fertilizer recommendation and Disease detection).

Crop Recommendation – The paper aims to develop and implement machine learning algorithms to accurately predict crop yields based on real-time and historical data. The models will be trained using a comprehensive dataset comprising diverse crops, regions, and agricultural parameters. The trained models will then be validated and tested using independent datasets to evaluate their performance and accuracy. The ultimate goal is to create a reliable and user-friendly crop yield prediction system that can assist farmers, policymakers, and agricultural stakeholders in optimizing agricultural practices, ensuring sustainable production, and addressing food security challenges. By leveraging the power of machine learning algorithms and incorporating real-time data inputs, this paper endeavors to revolutionize crop yield prediction, enabling farmers to make data-driven decisions, and optimize resource allocation.

Fertilizer Recommendation – The efficient use of fertilizers is crucial for maximizing crop productivity while minimizing environmental impact. Traditional fertilizer recommendation methods often rely on general guidelines or expert knowledge, which may not consider the specific needs of individual crops, soil types, or environmental conditions. However, with the advancements in machine learning techniques, it is now possible to develop more accurate and personalized fertilizer recommendation systems. By adopting a machine learning-based fertilizer recommendation system, farmers can optimize nutrient management, minimize fertilizer waste, reduce environmental pollution, and improve crop yields. It recommends suitable fertilizer for the crop based on the following parameters: Nitrogen, Phosphorus and Potassium and it will suggest some good ways to use the fertilizers and cultivate crops. The system can adapt to changing soil conditions, crop varieties, and weather patterns, ensuring continuous improvement and responsiveness.

Disease Detection – By implementing deep learning, we detect the type of disease the crop is affected by and suggest a cure to that disease so that the farmers can benefit and enable them to implement targeted management strategies such as precise pesticide application or crop rotation. The system can aid in reducing the reliance on broad-spectrum pesticides, optimizing resource allocation, and improving overall crop health and productivity.

This paper aims to develop a reliable and user-friendly system that assists farmers in detecting and diagnosing crop diseases accurately and in a timely manner. By harnessing the power of machine learning, we can enhance crop health, increase yields, and contribute to sustainable and resilient agricultural practices.

II. Literature Survey

The growth of artificial intelligence has opened up new opportunities in the advancement of agriculture framework. Artificial neural network (ANN) is the most widely used machine learning algorithm for crop-yield prediction and agriculture planning.

In [1], the authors have studied the impact of various climatic factors on crop yield in Madhya Pradesh, India. They have developed a software tool named "Crop Advisor" that uses C4.5 algorithms to identify the most influencing parameters. In [2] the authors have used SVM and Naïve Bayes algorithms to design two ensemble methods AdaSVM and AdaNaive for rice yield prediction in Tamil Nadu, India. In [3] the authors have proposed a framework using SNM for crop selection and in [4] the authors have analyzed the use of machine learning for crop yield prediction in Jammu, India using soil parameters. In [5] the authors have analyzed the paddy yield using weather and soil parameters. In [6], the authors analyze the performance of deep and machine learning models by considering soil conditions and climate conditions to predict the yield of crops. In [7], the authors used traditional crop modeling with machine learning methods to create a generic crop yield forecaster. The authors in [8] have conducted their study in Karnataka, India and uses crop yield and weather data to predict crop yield using both machine and deep learning algorithms. In [9] a hybrid regression model using reinforcement and random forest algorithms has been proposed and in [10] random forest techniques have been used for cotton yield prediction in Maharashtra, India.

It has been observed that major machine learning algorithms that have been used for crop yield prediction are artificial neural network, support vector, machine, decision tree, random forest and regression. Secondly, apart from crop yield there are many other factors that affect crop yield. The most widely used parameters are soil parameters, climate parameters and solar parameters.

Since artificial neural networks are the most widely used machine learning algorithm, this article aims at investigating the performance of artificial neural networks with Logistic Regression, Decision Tree, Random Forest, XGBoost, Naive Bayes and Support Vector Machine.

III. Methodology

A. IMPORT LIBRARIES

Python modules can get access to code from another module by importing the file/function using import. The import statement is the most common way of invoking the import machinery, but it is not the only way. import module_name.

B. LOAD DATASET

Load Data With Built-In Python Functions. Dataset Description: This dataset is relatively simple with very few but useful features unlike the complicated features affecting the yield of the crop and has been taken from Kaggle.

It consists of 7 features namely - N: Ratio of Nitrogen content in the soil, P: Ratio of Phosphorus content in the soil, K: ratio of Potassium content in the soil, temperature: Temperature in degree Celsius, Humidity: relative humidity in %, ph: ph value of the soil, rainfall: rainfall in mm. The task is to predict the type of crop using these 7 features.

The number of samples is 2200, and the total number of class labels are 22, some of which are: rice, maize, coffee, muskmelon, etc. The number of samples per class is 100, which shows that the dataset is perfectly balanced and does not need any special imbalance handling technique.

C. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

D. DATA CLEANING

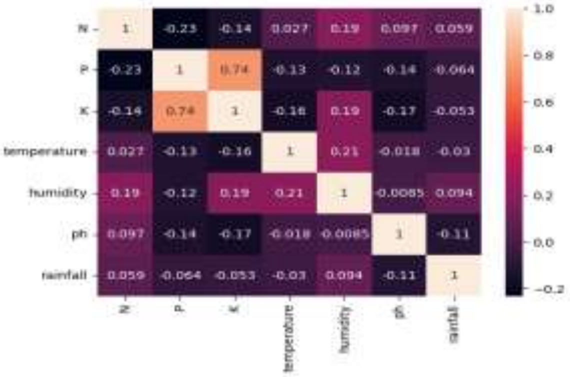
Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.

E. FEATURE ENGINEERING

Feature engineering refers to manipulation — addition, deletion, combination, mutation — of your data set to improve machine learning model training, leading to better performance and greater accuracy. Effective feature engineering is based on sound knowledge of the business problem and the available data sources.

F. DATA VISUALIZATION

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.



G. BUILDING A MODEL

The model building process involves setting up ways of collecting data, understanding and paying attention to what is important in the data to answer the questions you are asking, finding a statistical, mathematical or a simulation model to gain understanding and make predictions.

Approach: The dataset is split into 5-folds and cross-validation is performed on these folds. We test performance with six models:

- Decision Tree with entropy as the criteria and a max depth of 5.
- Naive Bayes.
- SVM with a 0-1 scaling on the input, polynomial kernel with degree 3, and the L2 regularization parameter C=3.
- Logistic Regression.
- Random Forest with 20 estimators
- XGBoost.

H. *TEST THE MODEL FOR FEW PROPERTIES*

To find the optimal procedure and parameters for the model, we will mostly employ K-fold Cross-Validation and the GridSearchCV approach. It turns out that the linear regression model produces the best results for our data, with a score of more than 80%, which is not terrible.

I. *EXPORT THE TESTED MODEL TO A PICKLE FILE*

We need to export our model as a pickle file (RandomForest.pickle), which transforms Python objects into a character stream. Also, to interact with the locations(columns) from the front end, we must export them into a JSON (columns.json) file.

IV. IMPLEMENTATION DETAILS

A. *TOOLS USED*

For the current paper, Python is chosen as the programming language for all the implementations, starting from extracting the data, to evaluating the model. It has a huge library support for applications in the field of Machine Learning and Artificial Intelligence and this makes Python more suitable for solving problems in real world scenarios. Jupyter notebooks were used to write the code for the paper. All the exploratory data analysis was done using Python libraries like NumPy, Pandas, Matplotlib, and Seaborn. The selection, training and evaluating the model was done using the Scikit-Learn library and its classes. No specific operating system is required as Python is a portable language.

B. *PERFORMANCE PARAMETERS*

The performance parameters used in this study is cross validation score.

Cross validation is a technique for assessing how the statistical analysis generalizes to an independent data set. It is a technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data.

For the *k*th part (third above), we fit the model to the other *K* - 1 parts of the data, and calculate the prediction error of the fitted model when predicting the kth part of the data. We do this for *k* = 1, 2, …, *K* and combine the *K* estimates of prediction error.

Here are more details. Let *K:* {1, ..., *N* } → {1, ..., *K*} be an indexing function that indicates the partition to which observation *i* is allocated by the randomization. Denote by $f^{-k}(x)$ the fitted function, computed with the k^{th} part of the data removed. Then the cross-validation estimate of prediction error is

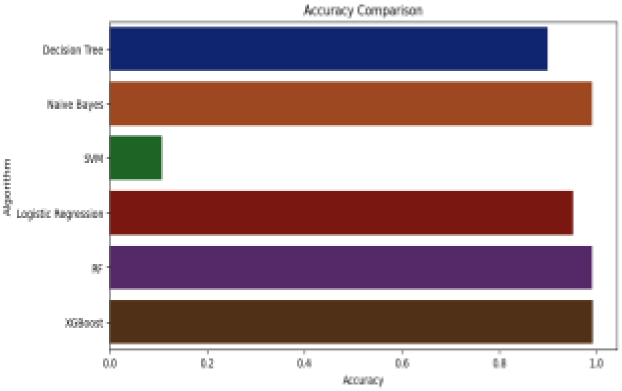
$$\mathrm{CV}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}^{-\kappa(i)}(x_i)).$$

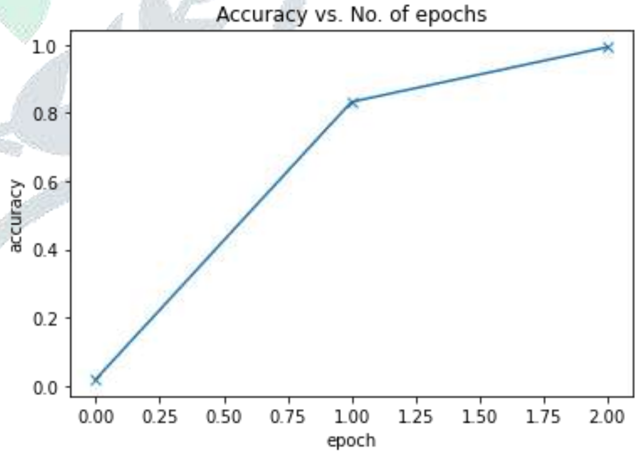
Typical choices of *K* are 5 or 10. The case *K* = *N* is known as leave-one-out cross-validation. In this case *k(i)* = *i,* and for the *i*th observation the fit is computed using all the data except the *i*th.

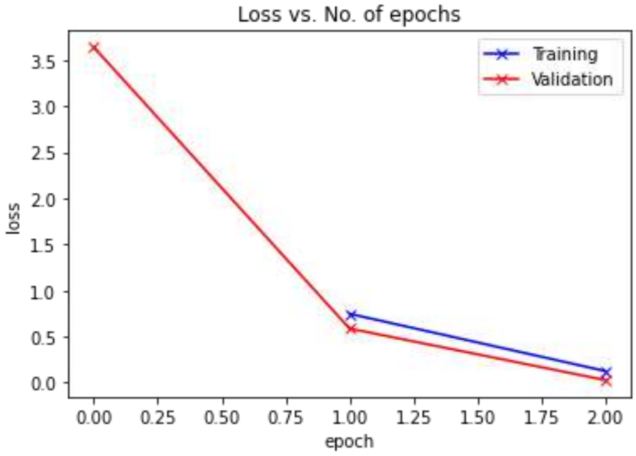
V. RESULTS

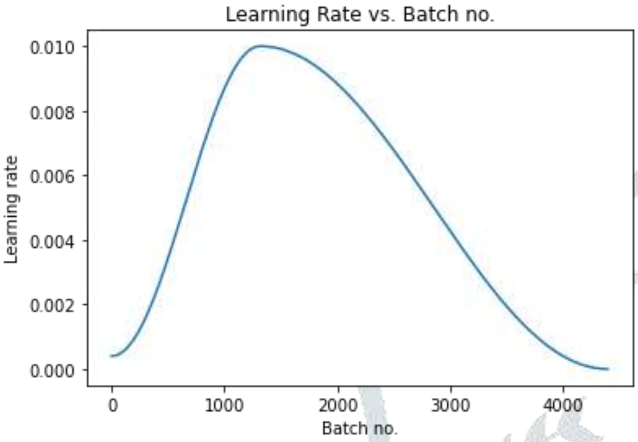
Crop recommendation: The crop recommendation model was trained by using the following six algorithms namely Decision tree, Naive Bayes, Support Vector Machine, Logistic Regression, Random forest, XGBoost. The accuracy of each model is given below:

MODEL	**ACCURACY**
Decision Tree	0.900
Naive Bayes	0.990
SVM	0.106
Logistic Regression	0.952
Random Forest	0.990
XGBoost	0.993









Disease Detection: The deep learning model was created by implementing ResNet and adding the required parameters. The images above show the accuracy and loss vs the number of epochs. The accuracy reaches its maximum at the second epoch and the loss is drastically reduced at the second epoch. The learning rate increases and reaches a maximum when the number of batches is around 1200.

VI. CONCLUSION

In this paper, we have created an easy to use web - application system based on machine learning and deep learning. By using this system, we are able to successfully provide features such as - crop recommendation using Random Forest algorithm, fertilizer suggestion using rule-based classifier and plant disease detection using ResNet and Convolutional Neural Networks. The user can provide their custom inputs using forms and get the results based on their inputs. The system suggests which crop is better to yield and provide suggestions for fertilizers based on the crops they desire to grow. The plant disease detection not only detects the disease but also provides a cure to rectify that disease.

The future enhancements may include the possibility of showing the crop's profitability based on the market trends. Analyzing market demands and recommending the suitable crops to cultivate. Integrating e-commerce sites to purchase the required fertilizers and farming tools. Predicting the life span of the yielded crops. Adding Farmer's native language as a translation option will create a better user experience for the Farmers.

REFERENCES

[1] S. Veenadhari, B. Misra, and C. D. Singh, "Machine learning approach for forecasting crop yield based on climatic parameters," Oct. 2014, doi: 10.1109/ICCCI.2014.6921718.

[2] N. Balakrishnan and G. Muthukumarasamy, "Crop Production - Ensemble Machine Learning Model for Prediction," Int. J. Comput. Sci. Softw. Eng., vol. 5, no. 7, pp. 148–153, Jul. 2016

[3] J. H. Jeong et al., "Random forests for global and regional crop yield predictions," PLoS One, vol. 11, no. 6, Jun. 2016, doi: 10.1371/journal.pone.0156571.

[4] V. Singh, A. Sarwar, and Sharma Vinod, "Analysis of soil and prediction of crop yield (Rice) using Machine Learning approach.," Int. J. Adv. Res. Comput. Sci., vol. 8, no. 5, pp. 1254–1259, Jun. 2017.

[5] R. B. Guruprasad, K. Saurav, and S. Randhawa, "Machine Learning Methodologies for Paddy Yield Estimation in India: a Case Study," Nov. 2019

[6] S. Agarwal and S. Tarar, "A hybrid approach for crop yield prediction using machine learning and deep learning algorithms," in Journal of Physics: Conference Series, 2021

[7] D. Paudel, H. Boogaard, A. de Wit, … S. J.-A., and undefined 2021, "Machine learning for large-scale crop yield forecasting," Elsevier, Accessed: Jun. 06, 2021.

[8] S. A. Shetty, T. Padmashree, B. M. Sagar, and N. K. Cauvery, "Performance Analysis on Machine Learning Algorithms with Deep Learning Model for Crop Yield Prediction," in Springer, 2021, pp. 739–750.

[9] D. Elavarasan and P. M. D. R. Vincent, "A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters," J. Ambient Intell. Humaniz. Comput., 2021, doi: 10.1007/s12652-020-02752-y.

[10] N. R. Prasad, N. R. Patel, and A. Danodia, "Crop yield prediction in cotton for regional level using random forest approach,"