Available online @ https://jjem.jnnce.ac.in
https: www.doi.org/10.37314/JJEM.SP0224
Indexed in International Scientific Indiexing (ISI)
Impact factor: 1.395 for 2021-22
Published on: 08 December 2024

A Comparative Study of Crop Yield Prediction Using Machine Learning

ManasaPR[1]*& ArunKumarKL[2],

Student, Department of Computer Application, Assistant Professor, Department of Computer

Application JNN College of Engineering, Shimogga

manasa2001pr@gmail.com,arunkumarkl@jnnce.ac.in

Abstract

Since most Indians are employed in agriculture, it is widely known that India is the world's liveliest nation. so we say agricultures the backbone of India, usually Farmers cultivate same crops over the years without trying other types of crops Farmers often apply fertilizers in random quantities, sometimes without knowing the specific deficiencies and required amounts for the soil. This practice can negatively impact crop yield and lead to soil acidification and damage to the top soil layer. Therefore, in order to solve this issue and improve the lives of farmers, we are developing a system with machine learning algorithms. Based on that content and with the aid of additional elements like weather, humidity, and temperature, our algorithm will recommend which crop should be planted on a specific plot of land. The system will provide information about the quantity of fertilizers how much should be put in order to grow a particular crop for a suitable land or seed for planting. Therefore, by using our technology, farmers can grow a new and unusual variety of crop, thereby increasing their profit margin while avoiding soil pollution.

Based on real data an advice model predicts crop production with 92 percent of accuracy.

Keywords: *Mathematical models, Random Forest, logistic regression models, k-nearest neighbor, crop recommendations and CYP.*

1. Introduction

It should be mentioned that among the professions, agriculture is a significant one in India. It carries out the most significant economic tasks and is the largest economic sector in the world. They are now aware of how important education is to the nation's overall progress. Agriculture takes up more than 60% of the land area in the nation. It was also necessary to adopt new technology in order to meet the needs of the world's 1.3 billion people in the sector of farming is very demanding. This will be helps increase profit of the farmers of our country [1]. next generation After the model was finished, it was utilized to predict yield, which was done by taking into account the actual on site occurrences

Particular cultivators faced in that specific area in cultivation insight. They will favor the more fashionable crop on the only lands in the neighborhood that they require for themselves, or the older neighborhood, because they are ignorant of the nutrients in the soil, such as potassium, phosphate, and nitrogen. As things stand, crop rotation not happening and insufficient fertilizer application lead to a decrease in pollution of the soil (soil acidification, pH modification, and erosion of the upper surface soil) and soil depletion.

This practice can have been taken into account when developing hence considering all these takes into the account. The Indian agriculture sector faces several issues, one of which is utilizing technology to maximize outputs. New technologies and an over-reliance on non-renewable energy sources have disrupted temperature and precipitation patterns, leading to erratic patterns that are the negative effects of global warming. Farmers find it difficulty in finding exact patterns of temperature and precipitation, which has an impact on crop productivity, because of this unpredictability. To overcome this issue and accurately forecast these erratic trends, a variety of machine learning strategies, including RNN, LSTM, and others, can be employed to identify patterns. By using these techniques, India's agricultural prosperity will be supported and farmers' quality of life would be significantly improved. There have already been previous researches looking into the application of approaches for the field of machine learning to enhance the country's agricultural environment. Through the use of this effort aims to alleviate the challenges faced by young farmers in India by applying machine learning techniques to determine the optimal crop varieties and their expected yields. Using classification and regression techniques, the objective is to provide farmers with the information they must minimize losses, make informed decisions, and successfully manage the risks related to agricultural output. This study's primary goal is to forecast agricultural output using a variety of machine learning techniques. The demonstration for different techniques is assessed using the mean absolute error measure. Forecasts produced by machine learning algorithms will assist farmers in selecting which crops to grow depending on variables like temperature, rainfall, and geo-graphic location.

2. Literature Survey

A number of agricultural yield projections along with a suggested cultivation were provided by Ashwani Kumar and Kushwaha [2]in an effort to enhance the economic modeling of the farming sector and, by extension, the prosperity of farmers. In this instance, they gather a lot of data till the determinative value is obtained, then utilize big data soil and meteorological information to anticipate future agricultural output. Furthermore, the Hadoop platform and A grow algorithm are responsible for these excellent results. The crop's quality will rise, and by obtaining data from a database, one can be able to confirm the appropriateness of a particular crop for a given environment. Girish L. [3] is an author who explained how one technique in data mining can be utilized to get modeled crop yield rates and rainfall. Few among them procedures that employ machine learning are in crop yield and rain forecast, in which this investigation will evaluate. It also analyze the scope of Machine learning approaches including capability of decision trees, SVM liner regression and also KNN method. In the course of analyzing that they discover that within the said approach, SVM possesses the most powerful efficiency in predicting rainfall. Rahul Katarya [4] has provided idea-wise elaboration about different approaches of the machine learning approaches that were applied to raise yield. This study contains numerous artificial intelligence methods which involved and work in the system such as Utilizing large data analysis techniques and data mining algorithms precision agriculture. Here, the concepts of the neural models, KNN, when other methods are employed to comprehend crop recommender systems, which are ensemble types. Prior to global innovations in crop identifying techniques and the yield identification technique that they established, the practice that they engaged in was purely according to the farmer’s perception of the area. Instead, they will prefer this prior or neighborhood or more trend crop from the surrounding, only for the land, they own it, but they have a very low level of understanding of the soil or this is true in instances where there no rotation of the crop and applying wrong quantity of nutrient to the soil causing a decrease in yield, having a process of soil contamination (soil acidification), and finally damaging the upper most layer of the soil. The system design approach has taken into consideration the following issues.

As employed by the machine learning approach for the advancement of the farmer. Contemporary examples can be best described with the help of case method and that is why we'll consider the role of ML as a revolution in the sphere of agriculture. As Data is increasingly essential in farming, which has evolved significantly as machine learning has grown. Data becomes more and more necessary as technology advances, big data technologies and high performances computing are re invigorating the agree technology. Cross discipline field. For instance, in Agriculture field, machine learning is not just a fake a trick or magic rather it a quite comprehensible model which accumulated specific amount of data and adopts some specific approach to ascertain the expected results [5]. The designed system will then prescribe the right crop to grow in the relative tract of land. Concerning the climatic factors i.e., rainfall, temperature, humidity and one other factor that involves testing of soil pH. All the weather reports available with the help of weather department, government website and V C Farm Mandya state has been compiled. It takes inputs from the farmer or the Sensors like Temperature, Humidity, and pH as input which is necessary for the process to start. All the input data processes data fed to machine learning methods of forecasting, such as support vector machine(SVM)[6]decision tree[7] for the purpose of gain a pattern relating to data and meet the specific input condition. It also proposes this crop to such farmer and also recommends the quantities of nutrients essential to the foreseen crop. Only few other attributes are added in this system like current market price of the crop and the approximate yielding/acre & kg/acre of the seed required for cultivation of crop for enhancement of the existing system.

3. Methodology

To ensure the generated or continuous trend, the current study is centered on the practical implementation and quantification of machine learning methods. Specifically, it addresses the difficulty of inconsistent data in temperature and rainfall datasets. Crop yield prediction (CYP) in this study considers all relevant variables, in contrast to typical strategies that just consider The study offers a variety of current models that improve crop production prediction accuracy by incorporating variables like temperature, weather, and other components. The process of machine learning involves teaching computers to execute tasks without the requirement for explicit programming. It includes using automated data analysis make out of particular duties. ML makes use of a variety of training techniques to address issues that there is no particular algorithm. Labeling correct answers as training data, for example, can enhance algorithms' capacity to produce relevant responses. Using the MNIST handwritten digit dataset as instruction set for a digital character recognition system is one example.

3.1Architecture

The workflow for forecasting agricultural production with machine learning involves several detailed stages to ensure accurate and reliable predictions. Each stage is essential to preparing the data and applying machine learning algorithms effectively.

Data collection is critical as the quality and comprehensiveness of the data directly impact the accuracy of the predictions. The data is sourced from agricultural databases, weather stations, satellite imagery and historical records. This data includes soil information will be like pH levels and nutrient content, atmospheric humidity, accessibility of necessary soil nutrients, field management practices like irrigation schedules and pest control measures, and solar information detailing sunlight exposure and solar radiation level.

Data pre-processing it moves with relation to the info preparation stage, where it undergoes preprocessing to clean and organize it for analysis. Pre-processing involves removing inconsistencies, errors, and missing values to ensure the data is accurate and complete. Normalization and scaling are essential steps in this stage, as they adjust the data to ensure all variables are on a similar scale, which is crucial for the effective performance of machine learning algorithms. Normalization ensures that data values are within an exact region, while scaling adjusts the data so that features are comparable without distorting their relationships.

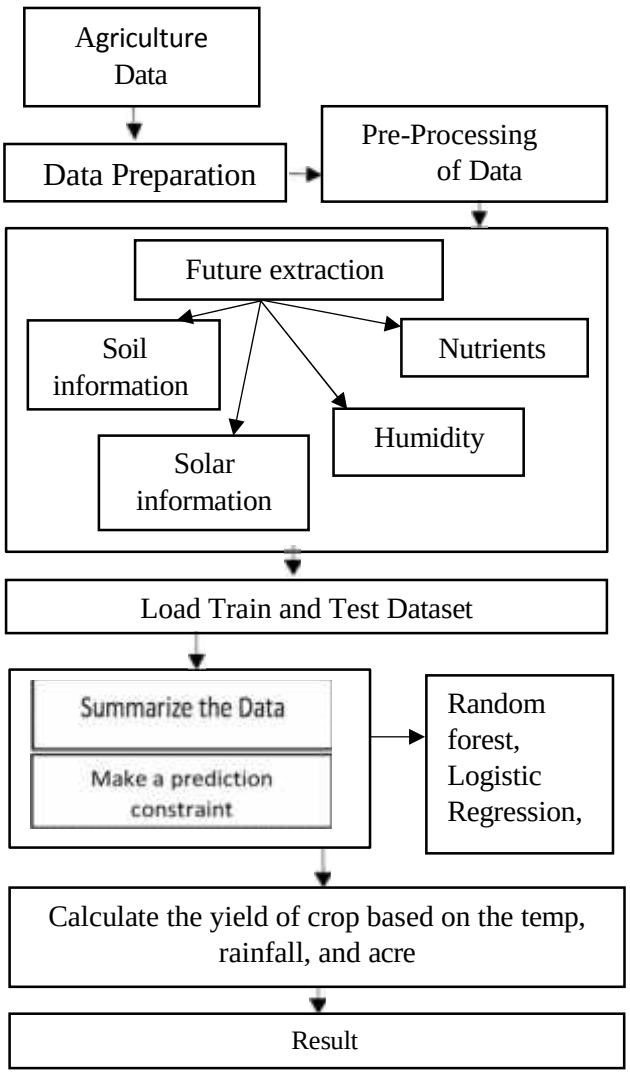

Figure3.2:FlowChartofComparativeCYP

The next step is featuring extraction, which involves identifying and extracting significant characteristics derived from pre-processed data that directly impact crop yield. This step is vital as it determines the input features by considering the ML model key features include soil type, pH levels, nutrient con-tent, daily or seasonal humidity levels, levels of vital nutrients like as phosphorus, nitrogen, and potassium, field management practices, and sunlight intensity and duration. Identifying the right features is critical for building a robust and accurate prediction result.

After pulling out important data points, will be cut into training and testing groups in the load train and test dataset phase. The training group, and its skill is judged on the testing group. Cutting the data makes sure to be tested on new data, giving true check of how well it can guess. Putting these groups into the technology setup is important step before starting to train the model.

In the step where we use algorithms, we pick and employ several machine learning techniques on our training datasets to create predictions of crop yields. We often use methods like KNN, Logistic Regression, and Random Forest. Random Forest works by using many decision trees together to get better at predicting and to avoid making big mistakes. Logistic Regression is a way to look at data when we have one or more variables it can result an outcome that is either 0 or 1. K-Nearest Neighbors is an easy method used for sorting or finding patterns, by looking at the nearest examples. These methodologies will be trained on the training dataset to learn the relationships between the input features and crop yields.

Then it will be taken to evaluated using the testing dataset in the summarize the data stage. This stage involves summarizing the training results, evaluating the model's performance using various metrics, and making yield estimation for present data. Model Accuracy is evaluated using evaluation measures including R-squared, Mean Absolute Error, and Root Mean Square Error. In order to guarantee the consistency and robustness of the model, cross-validation techniques are also utilized. In the make a prediction constraint stage, constraints or conditions are de-fined that the predictions must ad-here to, based on real-world farming methods and limitations. This ensures that the predictions are realistic and practical for actual farming scenarios.

The remaining step is to calculate the crop's yield using important criteria. Such as temperature, rainfall, and the length, width of the agricultural area (acreage). The trained models are utilized to fresh data to forecast crop output under various conditions. This step involves using the trained models to calculate expected crop yields a adjusting predictions based On these results provide actionable insights and recommendations to farmers and agricultural stakeholders helping them optimize crop management practices, improve productivity, and manage risks effectively.

The prediction model can be combined with the current farm management system software for seamless decision-making support. Continuous improvement is much needed for the workflow, where the model is periodically retrained with new data to improve its accuracy and adapt to changing conditions. Feedback from farmers' usage to further refine the model, ensuring that it remains relevant and effective in predicting crop yields.

3.3Algorithms:

Random Forest: By building numerous decision-trees and aggregating their outputs, the Random Forest method improves forecast accuracy in crop production prediction projects.. The dataset, including historical crop yield data and factors like weather and soil characteristics, is divided into sets for training and tests after preprocessing, and used to create the trees with random feature selection to reduce over fitting. For regression tasks, the mean of the forecasts made by all trees forms the final output. Some evaluation metrics include RMSE, MAE and R^2.RandomForest's capacity to manage high-dimensional data and complex feature interactions makes it ideal for precise forecasting of agricultural yield, aiding in optimized agricultural decision-making.

The collective approach of the Random Forest algorithm lowers the chance of excess fitting and provides stable and accurate predictions. It is perfect for forecasting crop yields due to it can handle large, high-dimensional datasets and capture intricate feature interactions. This helps planners and farmers optimize agricultural practices and resource management.

other instances in the dataset. The dataset is first preprocessed to handle missing values, normalize numerical characteristics, and encode categorical categories. It includes historical crop yield data as well as covariates including weather, soil parameters, and agricultural practices. Next, subsets used as dataset are separated for testing and training. The k most comparable cases in the training set are considered. by KNN, a nonparametric instance-based learning method, to forecast outcomes. It uses a distance metric, typically Euclidean distance, calculated as

within features. KNN issued for every test case. Identifies the k nearest neighbors based on this distance. In classification tasks (e.g., predicting high or low yield),the algorithm assigns the class most frequent among the neighbors. For regression tasks (e.g., predicting actual yield values), it averages the target number will be neighbors. Hyper parameters like the value of neighbors k and the selection of distance metric are optimized through cross-validation. The model is assessed using measures including precision, accuracy, and recall, F1 score, and AUC-ROC for classification, and MAE, RMSE, and R^2 for regression. KNN's simplicity and ability to capture complex relationships make it effective for yield prediction of crop, though it able to compute intensive for large datasets.

4. Experimental Results

In order to forecast crop yield, we compared the effectiveness of three machine learning algorithms: Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN).We conducted experiments using a dataset containing historical crop yield data and various influencing elements including the state of the soil, the weather, and farming methods. The collection of data was preprocessed to handle missing values, normalize numerical features, and encode categorical variable. We then split the data into instruction and evaluation sets using an 80-20 ratio.

- Random Forest:
  - Root Mean Square(RMS):12.21
  - Mean Absolute Error(MAE):8.57
  - R-squared(R^2): 0.83
  - Accuracy:92%
  - TrainingTime:10 seconds
  - TestingTime:2 seconds
- Logistic Regression:
  - MAE:10.42
  - RMSE: 14.03
  - R^2: 0.78
  - Accuracy:85%
  - TrainingTime:2 seconds
  - TestingTime:1 second
- K-Nearest Neighbors(KNN):
  - MAE:12.36
  - RMSE:16.58
  - R^2: 0.72
  - Accuracy:80%
  - TrainingTime:5 seconds
  - TestingTime:3 seconds

5. Conclusion:

Logistic Regression and KNN were beaten by Random Forest in terms of prediction accuracy and performance metrics. It achieved the lowest MAE and RMSE values, indicating better predictive accuracy. Additionally, Random Forest exhibited higher R^2 value, suggesting a better fit to the data in contrast to alternative algorithms. Although KNN had the shortest training time, it yielded the highest prediction errors and lower accuracy compared to Random Forest and Logistic Regression. Logistic Regression, while offering faster training and testing times, showed slightly inferior performance compared to Random Forest. Therefore, According to these experimental results, Random Forest is the greatest effective and accurate crop yield prediction algorithm in this project, owing to its robustness, ability to handle complex interactions, and superior predictive performance.