



Crop Selection and Yield Prediction using Machine Learning Approach

**PRITESH PATIL*, PRANAV ATHAVALE, MANAS BOTHARA,
SIDDHI TAMBOLKAR and ADITYA MORE**

Department of Information Technology, AISSMS Institute of Information Technology, Pune, India.

Abstract

In recent years, Agriculture sector has been researched a lot with the advancements in technologies like machine learning and smart computing. With the dynamic economics of Agri-produce, it is becoming challenging for farmers to utilize the land efficiently to get maximum profit in the specific landscape. Crop Yield Prediction (CYP) is crucial and is greatly dependent on environmental factors like soil contents, humidity, rainfall as well as area under cultivation and other required metrics. Due to insufficient incorporation of the multiple environmental circumstances, a number of existing tools and techniques used for CYP, such as historical averages, tend to produce inaccurate findings. In such situation, with multiple options of crop, it is essential for farmers to plan the crop strategy in advance. If the farmer can get estimate of the crop yield in advance, cultivation can be done accordingly. To solve this problem, machine learning approach is implemented as a base for accurate predictions. Crop prediction is done by classification model and yield prediction uses regression models to learn from the data. Multiple ML models are analyzed based on performance metrics. Best performer model is incorporated in backend. Among the used models for yield prediction, Random Forest Regression gives best results with MAE of 0.64 and R2 score of 0.96. For crop prediction, Naïve Bayes classifier gives most accurate results with accuracy of 99.39. The study emphasizes how machine learning could revolutionize crop management techniques by giving farmers insights about optimizing resource allocation and boost overall crop yield.



Article History

Received: 11 May 2023

Accepted: 31 October 2023

Keywords

Crop Yield Prediction;
Digital Agriculture;
Machine Learning;
Naïve Bayes;
Random Forest.

Introduction


The field of machine learning is advancing day by day. Learning is important when we need

sample data or experience rather than the ability to immediately design a computer program to solve a particular problem. When there is no human

CONTACT Pritesh Patil ✉ pritesh.patil@aissmsioit.org 📍 Department of Information Technology, AISSMS Institute of Information Technology, Pune, India.



© 2023 The Author(s). Published by Enviro Research Publishers.

This is an  Open Access article licensed under a Creative Commons license: Attribution 4.0 International (CC-BY).

Doi: <https://dx.doi.org/10.12944/CARJ.11.3.26>

knowledge or when people are unable to express their expertise, learning becomes important. Computers are programmed with machine in order to improve performance criteria based on actual or hypothetical facts. Computer program learns to optimize the parameters used for the model using training input or previous information. The model may be descriptive to draw conclusions based on model data or predictive which estimates trends in future.¹ A subset of artificial intelligence (AI), machine learning (ML) enables computers to learn for a specific dataset such as playing chess or making recommendations on social networks without having to be explicitly programmed. Precision farming and Agri-technology, now referred to as Digital Agriculture, are evolving into emerging fields in research that employ highly data-driven techniques to boost productivity in agriculture while shrinking the adverse effects on the environment. Machine learning (ML), alongside big data technology and robust computing infrastructure, has arisen to create potential solutions for unravelling, quantifying, and comprehending data-intensive processes in agricultural operational environments. Data analysis, as an evolved scientific discipline, is essential to the development of a wide range of crop management applications. Many times, it is possible to efficiently use ML without having integrating data from many sources. There tends to be less emphasis on data integration when large datasets are easily available, especially on a major scale. The main force behind this development is the complexity of data preprocessing and analytical processes, as opposed to the machine learning models' generally straightforward implementation.² Agriculture sector has a major contribution of almost 20% in India's GDP in year 2019-20.³ Also, it is the principal source of employment in India. In addition to being a significant part of the global economy, it is crucial for the continued existence of humanity. Weather, pests, and the readiness of harvesting operations are the main factors that influence agricultural production. For managing agricultural risk, it's essential to have accurate crop history information.⁴ Unethical practices are being used to produce higher yields of less-nutritious hybrid cultivars as the population grows. These techniques tend to harm soil quality. It results in environmental loss. Given the changing patterns of weather conditions and also economics, it is getting difficult to choose right

crop for farmer. The use of various fertilizers is also unclear because of seasonal climate variations and changes in the availability of fundamental resources like soil, water, and air. The agricultural yield rate is continuously decreasing in this situation.⁵ Farmers today cultivate crops based on knowledge gained from earlier generations. Since the traditional method of cultivation has been refined, there are either excessive or insufficient yields without really meeting the need.⁶ If the producer knows yield estimates in advance, it would help to form the crop strategy. Machine learning is a rapidly expanding methodology that supports and provides a guide in decision process in various applications of multiple different industries. The majority of modern gadgets benefit from models being examined before deployment. The primary idea is to increase the efficiency and profits of the agriculture industry by using data as a tool with models. Precision farming, which prioritizes quality above unfavorable environmental variables, would be the main focus.⁷ ML has advanced its applications in agriculture in areas like predicting soil properties, rainfall analysis, yield prediction, disease and weed detection, ML based computer-vision and many more.⁸

The use of computer vision, machine learning, and IoT applications will assist boost productivity, enhance quality, and ultimately increase the profitability of farmers and related industries. To increase the overall harvesting output, precision farming is crucial in the world of agriculture.⁹ For example, smart irrigation systems, crop disease prediction, crop selection, weather forecasting, and determining the minimal support price are all examples of techniques employed in agriculture. These methods will increase field productivity while requiring less work from farmers.¹⁰ Crop yield estimation may be used for a variety of purposes, including helping farmers enhance production, optimizing the supply-demand cycle for fertilizers, insecticides, and other agricultural products, predicting prices, and calculating the risk levels for agricultural insurance.¹¹

Literature Review

Prior research¹² used data that included nutrients and other environmental elements to anticipate crops. For CYP, several feature selection techniques and ML models are employed. In this study, the

following factors were looked at: To assess the effectiveness of feature selection and classification algorithms, F1 Score, Mean Absolute Error (MAE), Logarithmic Loss (LL), Accuracy (ACC), Specificity (S), Recall (R), Precision (P), and Recall (R) were utilized. (AUC). Using Modified Removal of recursive Features, six variables - average soil and air temperatures, min and max air temperatures, precipitation, and humidity are selected. A variety of data splitting validation techniques, including (25- 75), (30-70), (35-65), (40-60), (45-55), (50-50), (55-45), (60-40), (65-35), (70-30), and (75- 25), are used and evaluated against the previously stated accuracy criteria. Additionally, versions of the feature selection techniques such as MRFE, RFE, and Boruta have been applied. According to the results, the Random Forests Classifier is the most accurate in comparison with kNN and other classifiers discussed above. As characteristic ranges broadened, the measurement values decreased.

Another study by Anakha Venugopal, Jinsu Mani, Aparna S Rima Mathew, Prof. Vinu Williams¹³ uses several machine learning approaches to forecast the agricultural production. By taking into account variables like temperature, rainfall, area, and other characteristics, Farmers will be able to select the crop that will provide the highest produce by using the forecasts made by ML models. The study is focused on Kerala's Agri-produce. Among the classifier models utilized here, Random Forest has the highest accuracy, followed by Logistic Regression and Naive Bayes.

A Research¹⁴ A smartphone app which is used in the proposed method connects farmers to the internet. GPS helps user in locating his location. The user enters the location and soil type. The most profitable crop list can be picked using machine learning algorithms, and they can also forecast crop yields for user-selected crops. Machine learning models, including random forest (RF), artificial neural network (ANN), support vector machine (SVM), multivariate linear regression (MLR), and k-nearest neighbor (KNN), are used to estimate crop productivity. Random forest demonstrated the best outcomes with 95% accuracy. The algorithm also makes recommendations on when to apply fertilizers to increase yield. This research focused

on the limitations of present approaches and their applicability for yield prediction. The suggested approach then connects the farmers with an effective yield forecasting system via an app for smartphones. To assist them in selecting a crop, people may select from a number of attributes. The integrated prediction system assists farmers in estimating the crop produce. A user may research possible crops and their yield using the integrated recommendation system in order to make better educated judgements. Based on data from states of Maharashtra and Karnataka, several ML models like RF, KNN, MLR, SVM, and ANN were built and compared for accuracy. Results confirm RF Regressor, which has a 95% accuracy rate, is the best standard algorithm when applied to the presented datasets.

In,¹⁵ the Random Forest Algorithm is used. In spite of extensive research into challenges and topics like weather, temperature, humidity, and rainfall, there are still no acceptable remedies or ideas to deal with the difficulty we face. In nations like India, there are numerous different sorts of rising economic growth, including in the agriculture sector. Additionally, crop yield predictions can be made using the processing. The current study proved the value of data mining techniques for predicting agricultural output based on input features related to the climate. All new grains and regions chosen for the investigation should have accuracy of prediction above 75%, demonstrating improved predictive performance. The produced website is user-friendly. The website was developed utilizing data from that area to predict crop yield.

According to a study,¹⁶ selecting the best crop before sowing will increase agricultural yield. It depends on a variety of factors, such as the soil type and its composition, climate, local terrain, crop yield, market prices, etc. Techniques like Decision Trees, K-nearest Neighbors, and Artificial Neural Networks have a position in the crop selection framework, which depends on a variety of different factors. Machine learning has been used to choose crops based on how natural disasters like hunger could affect them. Researchers have employed artificial neural networks to choose crops depending on soil and climate with success.

When attempting to create a high-performance predictive model, ML studies face a variety of difficulties. To tackle the issue at hand, it is essential to choose the appropriate algorithms, and both the algorithms and the supporting platforms must be able to handle the sheer amount of data.¹⁷

A study¹⁸ suggested a method for unsupervised fuzzy categorization that identifies crop kinds with springtime harvests. The categorization outcomes likely to get better with time. Strategy used in¹⁹ made use of the Bayesian network categorization supervised learning model. Crop information is analyzed with environmental parameters like temperature and rainfall to categorize crops.

A study by D. A. Reddy, B. Dadore, and A. Watekar²⁰ highlights how despite being one of the nations with the highest agricultural output, India's agriculture productivity is still fairly low. Productivity needs to be increased so that farmers may get better profit from decreased costs. In order to reliably and successfully propose a suitable crop based on soil data, it offers solutions such as offering a recommender utilizing an ensemble approach with a large proportion of voting methods employing random tree, CHAID, kNN, and naive bayes classifier. Soil types, soil characteristics, and crop yield data collection are taken into consideration when advising the farmer on the best crop to grow. The majority voting process, which is the most popular assembly technique, is used in this system. Any number of primary learners may be used in the voting process. A minimum of two base learners are required. The chosen learners complement one another and impart knowledge to the others. With more competition, a better forecast may be made. The specified training data set is used to train the model. When a new record has to be categorized, each model chooses the class independently. Class predicted by consensus of learners is chosen as class label for current record.

A study²¹ says Building a random forest, a group of decision trees that considers two-thirds of the records in the datasets, takes into account data sets on temperature, production, perception, and rainfall. These decision trees are then applied to the remaining data to ensure accurate categorization. For accurate crop production prediction based on

the input qualities, the test data may be applied to the generated training sets. The RF method and the dataset were used to evaluate the efficacy of this technique. The advantage of the random forest approach is that overfitting is less of an issue with random forests than it is with decision tree-based model. The random forest does not need to be trimmed. The loaded data sets are divided into train, test data of 67 or 33 percentage points, or 0.67 or 0.33 respectively. In order to enable the mapping of attribute values to appropriate values and list placement, the training data must be categorized. By contrasting the initial data with model predictions, the probability is determined. Based on the result, the highest likelihood is utilized to make a forecast. The accuracy may be calculated by comparing the generated class value with the test data set.

According to a different study,²² agriculture has positive economic effects on the country. It falls short, nevertheless, in terms of using modern machine learning techniques. As a result, our farmers ought to be knowledgeable with all of the most recent machine learning technology and fresh approaches. The productivity of agriculture is increased by using these methods. To increase agricultural productivity rates, a number of machine learning approaches are used. These techniques can help with agricultural problems. We may also assess the accuracy of the yield by looking at several ways. Thus, we may perform better by contrasting the accuracy of several crops. In agriculture, sensor technology is widely used. The study helps increase agricultural yield rates. helps choose the right crop for the chosen site and season.

Materials and Methods

Data Pre-processing

A technique called data pre-processing transforms unprocessed, uncleaned data ready for further analysis. Data may be gathered from multiple sources, but as they are collected in raw form, analysis is not possible. We convert data into a comprehensible format by using several strategies, such as substituting missing values and null values. Fields in the dataset which are insignificant for label prediction are eliminated. If required, One-Hot Encoding is performed on the dataset to have dataset ready for regression model fitting. The division of train and test data is the final stage in the

data preparation process. As training the machine learning algorithm usually requires as much data points as possible, the data typically has uneven distribution. Training dataset, which in this case makes up 70% of total data, is used to train machine learning models and make accurate predictions.

Factors Affecting Crop Yield

The yield of every crop is impacted by a wide range of variables. These are essentially the characteristics that aid in estimating a crop yield. For crop yield prediction, this study includes parameters such as temperature, rainfall, area, humidity, soil nutrients, pH, and AUC (Area under Cultivation).

Comparison and Selection of ML Algorithm

We first must assess and compare different algorithms before selecting the one that best matches this particular dataset. Machine Learning is an effective way to solve crop prediction problem as it learns from past data and gives predictions on current parameters. In order to make precise predictions and stand by erratic patterns in weather conditions like temperature and rainfall, various machine learning classifiers like Logistic Regression, Naïve Bayes, Random Forest, KNN are used and compared for the performance metrics and the model with best accuracy is selected for crop prediction. For Yield prediction, regressors like Linear Regression, Random Forests and Decision Trees Regression are compared for metrics like MAE, Median Absolute Error and R2 Score. The model with best values is selected for predicting yield.

Naïve Bayes

Based on Bayes' theorem, Naïve Bayes model is frequently employed in many classification tasks. The multinomial, Bernoulli, and Gaussian algorithms make up the three Naive Bayes algorithms. Naive Bayes Algorithm is mostly employed for classification problems. It operates under the presumption that each feature has an equal chance of occurring and that the likelihood of each feature occurring is independent of the probabilities of the occurrence of all other features. The Bayes theorem determines likelihood of an event happening when another event is occurred. Multi-class classification makes use of Bayes theorem. Also, in comparison to other ML techniques, it is quicker and simpler to construct. Additionally, it doesn't need a lot of training data.

Both discrete and continuous data may be used with it. It is extremely scalable and unaffected by insignificant features.

Decision Trees

A decision tree is a type of tree structure that resembles a flowchart and is frequently employed in supervised machine learning for classification and prediction. A DT may be transformed into a set of rules, with each path serving as a different rule, with each path travelling from the root node to each leaf node. In a decision tree, each leaf node has a class that may be reached if an attribute matches the prerequisite for the branch that leads to it. In a decision tree, each internal node corresponds to a test, condition, or attribute.⁶

KNN

The machine learning approach known as kNN, which is supervised and nonparametric, is used to solve classification and regression issues. Labeled data is used with supervised algorithms. The technique relies on the distances between the points, which may be calculated in a few different ways. The fact that the distance must always be either zero or positive should be taken into account. The distance is squared, raised to a given power, or the absolute values are used to do this. Pre-processing of all the labelled data is necessary before we apply the kNN algorithm. All of the data must first be normalized. As kNN struggles to function when there are too many features present, feature selection must then be used to eliminate the insignificant features. Missing data must be filled in. Else, that particular record must be eliminated. The performance can be enhanced by including more train samples. The fundamental drawback of KNN is that as the size of dataset grows, cost of computing rises, the algorithm's speed decreases.

Random Forests (RF)

The RF technique is a perfect example of ensemble learning in action since it connects several classifiers to tackle the challenging problem and improve a model's efficiency. The "forest" created with this approach is actually a collection of decision trees. In each decision split, RF characteristics are chosen at random. Picking traits that encourage prediction and lead to increased efficiency reduces the correlation across trees. The Random Forest ML

classification approach generates the final output by combining the results of all the decision trees after segmenting the dataset into smaller subsets or trees. The Bagging subcategory of ensemble learning methods includes Random Forest. A Sample of rows and features from the primary dataset are selected at

random and fed into the Random Forest Technique's decision trees. It can also carry out jobs requiring both regression and classification. It also works well with huge, highly dimensional data sets, and most significantly, it greatly improves the model's accuracy and fixes the overfitting problem.

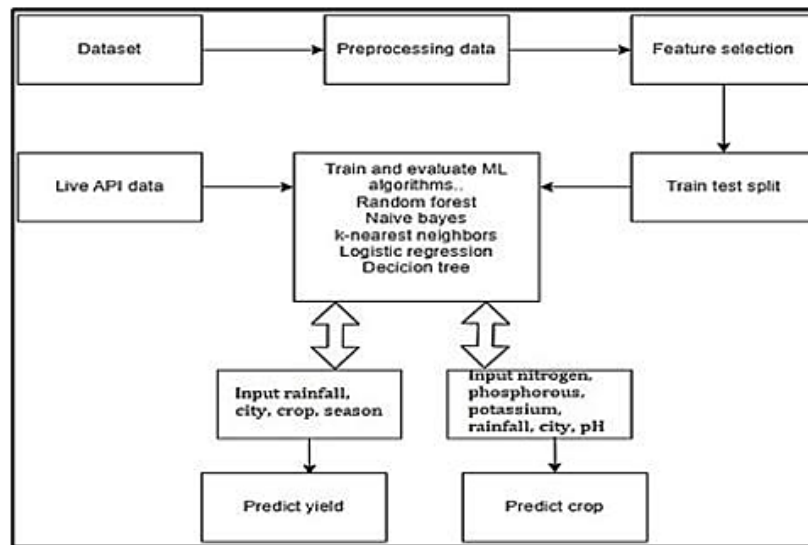


Fig. 1: System Architecture

System Architecture

System architecture is represented in Figure 1. End user interacts with web user interface (UI) which is hosted on a server. Open Weather Map API is connected to server to deliver weather data. The machine learning models are trained and tested by admin and are loaded in the server for predicting crop and yield in tone per hectare of land area.

Datasets

The public datasets have been chosen because they are readily available and easily accessible. Kaggle is a popular platform for finding and sharing datasets, so we were able to find datasets that met our criteria. We selected 3 datasets namely

India Agriculture Crop Production²³

The dataset has following features: State, District, Crop, Year, Season, Area, Area Units, Production, Production Units, Yield. This dataset is used to build regression model for yield prediction. Yield is the required label. It has total 12176 records containing

27 unique crops and 4 unique seasons. Crops are as follows: Arhar/tur, bajra, castor seed, gram, groundnut, jowar, linseed, maize, moong, niger seed, other cereals, kharif pulses, rabi pulses, summer pulses, ragi, rapeseed and mustard, rice, safflower, sesamum, millets, soyabean, sugarcane, sunflower, tobacco, urad, wheat, oilseeds.

District Wise Rainfall Normal²³

This dataset is used for collecting district wise rainfall data to predict yield. It is used for extracting district wise average annual rainfall for each district of Maharashtra, India. This feature is combined with Yield dataset mentioned above to get estimates of production for particular crop in the given season.

Crop Recommendation²³

The dataset is used for crop prediction. It has features like N, P, K, rainfall, humidity, pH and crop. N, P, K stands for Nitrogen, Phosphorous and Potassium nutrients in soil. It has 2200 total records containing 22 unique crops. Data consists 100

records for each of the following crops: rice, maize, chickpea, kidney beans, pigeon peas, moth beans, mung beans, black gram, lentil, pomegranate, banana, mango, grapes, watermelon, muskmelon, apple, orange, papaya, coconut, cotton, jute, coffee.

Data Pre-Processing

Crop Prediction

Prior to start modelling the data, we need to carry out data-pre-processing. It is done in following steps as shown in Figure.

Handling Missing Values

The line `df = pd.read_csv('Crop_recommendation2.csv', na_values='')` reads the CSV file into a DataFrame (df), replacing any occurrences of '=' with NaN values, which are commonly used to represent missing data in pandas.

Separating the Target Variable

The line `b = df['label']` extracts the target variable column ('label') from the DataFrame df and assigns it to the variable b.

Creating a Preprocessing Pipeline

The code creates a pipeline (my_pipeline) using scikit-learn's Pipeline class. The pipeline consists of two steps

Imputation

The missing values in the DataFrame are imputed using the SimpleImputer transformer with a strategy

set to "mean". This strategy replaces the missing values with the mean of the corresponding column.

Standardization

The features are standardized using the Standard Scaler transformer. This step scales the features to have zero mean and unit variance.

Applying the Preprocessing Pipeline

The line `X = my_pipeline.fit_transform(df)` applies the preprocessing pipeline (my_pipeline) to the entire DataFrame (df). It fits the pipeline on the data to learn the mean values (for imputation) and the standardization parameters. Then, it transforms the data by applying the learned transformations.

Train-Test Split

The `train_test_split()` function from scikit-learn is used to split the processed features (X) and the target variable (b) into training and testing sets. The `stratify=b` parameter ensures that the class distribution is maintained in both the training and testing sets. The split is performed with a test size of 30% (`test_size=0.3`) and a random state of 42 (`random_state=42`).

These pre-processing steps help handle missing values, standardize the features, and split the data into training and testing sets for further analysis and model training.

```
import pandas as pd
df = pd.read_csv('Crop_recommendation2.csv', na_values='')
df.head()

b = df['label']

from sklearn.impute import SimpleImputer
imputer = SimpleImputer(strategy="median")
imputer.fit(df)

from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
my_pipeline = Pipeline([
    ('imputer', SimpleImputer(strategy="mean")),
    ('std_scaler', StandardScaler())
])

from sklearn.model_selection import train_test_split
a = my_pipeline.fit_transform(df)
a = df
X_train, X_test, y_train, y_test = train_test_split(a, b, stratify=b, test_size = 0.3, random_state=42)
```

Fig. 2: Preprocessing for Crop prediction

Yield Prediction

In data preprocessing, we did clean the data containing missing values, outliers, or errors that need to be addressed before the data can be used for machine learning. Also, we did data integration of District wise rainfall normal²³ and India Agriculture Crop Production²³ as we required it to be merged for passing it to the machine learning model. We

did data transformation for India Agriculture Crop Production²³ dataset as it had categorical variables which need to be encoded as numerical values to pass to the machine learning model. We used One Hot Encoding for data transformation. We had to do data reduction to limit the dataset to the state of Maharashtra otherwise the dataset would have been too large in terms of the rows and columns.

```
import pandas as pd
df_rainfall = pd.read_csv('district wise rainfall normal.csv', na_values='')
df_rainfall=df_rainfall[df_rainfall['STATE_UT_NAME']=='MAHARASHTRA']

df = pd.read_csv('India Agriculture Crop Production1.csv', na_values='')
df=df[df['State']=='Maharashtra']
df=df[df['Year']>="2004-05"]
df=df[df['Production Units']=='Tonnes']
df = df.drop('Area', axis = 1)
df = df.drop('Production', axis = 1)
df = df.drop('State', axis = 1)
df=df.drop('Production Units', axis=1)
df=df.drop('Area Units', axis=1)

merged_df=pd.merge(rainfall_data, df, on='District')
from sklearn.preprocessing import OneHotEncoder
df_onehot = pd.get_dummies(merged_df, columns=None, prefix=None)
b = df_onehot['Yield']
df_onehot = df_onehot.drop('Yield', axis = 1)
a=df_onehot
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(a, b, test_size = 0.3, random_state=42)
```

Fig. 3: Preprocessing for Yield Prediction

Feature Selection

A machine learning model's performance can be improved through feature selection, which is the process of choosing a subset of the relevant features from available data. For Crop Prediction, following features were selected: Nitrogen, Phosphorous, Potassium, Temperature, Humidity, pH and Rainfall. For Yield Prediction, features selected are as follows: City, Crop, Annual Rainfall (in mm), Season.

Train Test Splitting of Data

We have split the data in the ratio 70:30 using random sampling and stratification. Choosing an appropriate train-test split is important in ML, because it can affect the accuracy and generalization of the resulting model.

API Integration

The city of user taken as an input is given to API call as a parameter. The temperature and humidity fields from API response are given to crop prediction model

as input along with other data. This helps the system to give real-time predictions. "OpenWeatherMap" API is used for the same. For creating an API URL, base URL and API key is used which is unique with each subscription. User's city name is passed in complete URL as a parameter and response is collected. From the collected response, required fields i.e., temperature and humidity are passed to ML model for predicting crop.

Training and Evaluation of Models

The crop prediction uses the multi-class classification machine learning model to predict the crop for a set of given input features. Whereas the yield prediction incorporates the regression model to predict the yield for a given set of input features. For training and evaluation of models, Google Colab Platform is used. While the User Interface for the project is built using ReactJs, the backend is built using Python Flask framework.

Application and Advantages over existing versions

The model can be used to create an impact on right crop selection as the user would get fair prediction on yield as well as crop. Also yield prediction would be important in financial assessment of crop strategy. Model is useful if the user wants to compare yield for multiple crop options and then select the best one. It could also be used in a wide geography to estimate the yield for a particular crop. This project can be used directly by end users as farmers for taking predictions for their conditions. Instead, it can also be used by government agencies for planning and policy making if modified with wider access to reliable closed source government data. It can also be used by NGOs which work for educating farmers in adopting new technologies and precision agriculture. Also, it can be used in fields where monetary calculations come in picture as it is dependent on how much yield could be produced like in insurance claims or loan policies.

The project improves the prediction accuracy by suitable data gathering cleaning and selecting best accurate model. Also, the project incorporates both crop as well as yield prediction. So, the project is using classification as well as regression models for necessary functionality. It adds value to the modern agriculture setup by providing a way to add to the reliability of crop selection which in turn improves the yield and financial stability.

Results and Discussions

Crop Prediction

First, datasets are loaded and cleaned from insignificant features. After Data Preparation, data is split into training and testing data and various models are fitted and tested for accuracy. Feature Importance is calculated to determine the relative significance or contribution of individual features in ML model. For crop prediction model, Drop Column Importance, also called as "permutation importance" or "feature importance by feature shuffling," is calculated.

$$\text{Drop Column Importance} = \text{Baseline Metric} - \text{Shuffled Metric}$$

Drop column importance is based on the idea that removing a feature that is crucial to the performance of the model would cause it to perform less significant than before. It is calculated in following steps

1. Train a model with all features
2. Measure baseline performance with a validation data
3. One feature is determined of which importance is to be calculated
4. Train a model with all other features except the selected one
5. Calculate performance with a validation data
6. The feature importance is the drop in performance from baseline
7. Follow same steps 3 through 6 for every feature

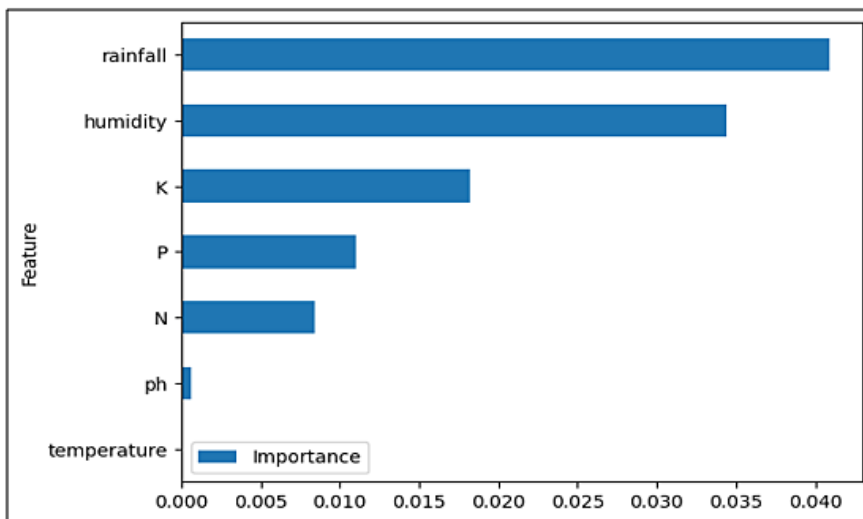


Fig. 4: Feature Importance for Crop Prediction

As shown in above Figure 4, rainfall is most important feature for crop prediction followed by humidity, Potassium(K), Phosphorous(P), Nitrogen(N) and pH.

Classification Models' results for Crop prediction as depicted in Figure 5.

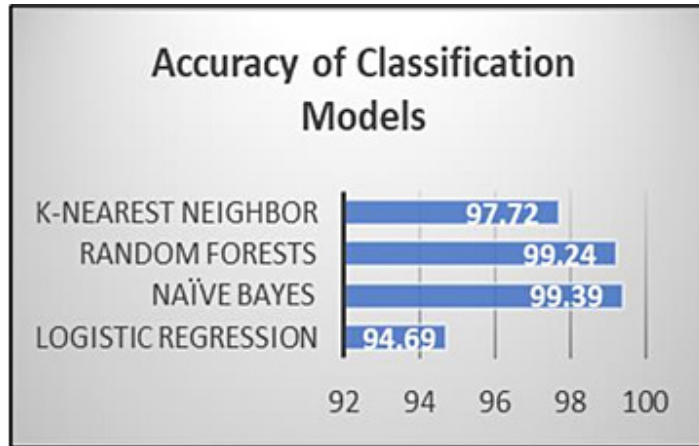


Fig. 5: Accuracy Metric for Classification Models

When trained on the dataset, KNN gives accuracy of 97.72%, RF gives accuracy of 99.24%, Naïve Bayes Classifier has 99.39% accuracy score. Logistic

Regression has accuracy of 94.69%. Based on these results, Naïve Bayes classifier is incorporated in the backend for Crop Prediction.

```

feature_importances = rfr.feature_importances_
categorical_features=('District','Crop','Season')
# Initialize a dictionary to store aggregated importances for original features
aggregated_importances = {feat: 0 for feat in categorical_features}

# Map and aggregate feature importances
for i, feature in enumerate(categorical_features):
    for col_name in X_train.columns:
        if col_name.startswith(feature):
            aggregated_importances[feature] += feature_importances[i]

# Sort the aggregated importances in descending order
sorted_importances = dict(sorted(aggregated_importances.items(), key=lambda item: item[1], reverse=True))
print(sorted_importances)

feature_index = X_train.columns.get_loc('Rainfall (mm)')
importance_of_specific_feature = feature_importances[feature_index]
print(f"Importance of 'Rainfall (mm)': {importance_of_specific_feature}")

{'Crop': 0.49430545245645613, 'District': 0.42547144722421487, 'Season': 0.0008708377708168945}
Importance of 'Rainfall (mm)': 0.012513866094829844
    
```

Fig. 6: Code Snippet for Feature Importance in Yield Prediction

Yield Prediction

Calculating feature importance for a Random Forest Regressor with one-hot encoded features involves determining the contribution of each feature to the model's predictive performance. It is done through following steps.

1. Train the Random Forest Regressor
2. Access Feature Importances: Random Forest Regressor has built in attribute named feature_importances_.
3. Map Feature Importances to Original Features: Every one-hot encoded feature is

- mapped with its original feature.
- 4. **Aggregate Feature Importances:** By aggregating we get every categorical feature's importance
- 5. **Rank Feature Importances** in descending order of importance.

As shown in Figure 7, Crop is most important feature in order to predict yield followed by District, Rainfall and Season.

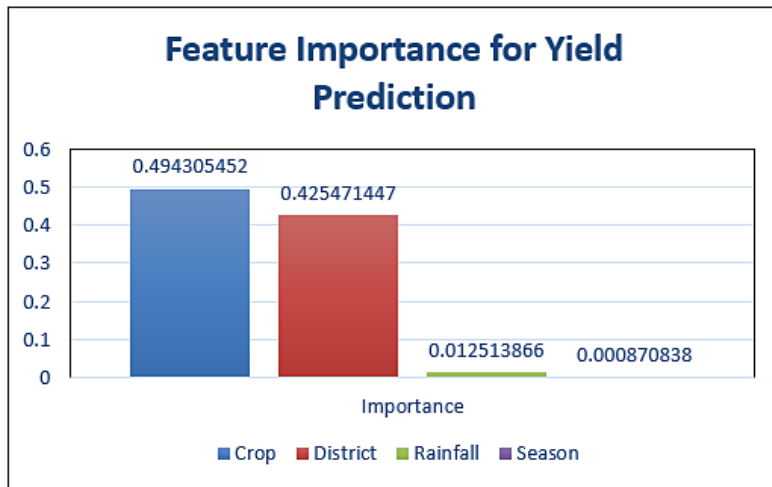


Fig. 7: Feature Importance for Yield Prediction

Yield prediction is done by regression. For comparison between different regression models, performance metrics like Mean Absolute Error,

Median Absolute Error and R2 Score are used. The results are depicted in figure 8.

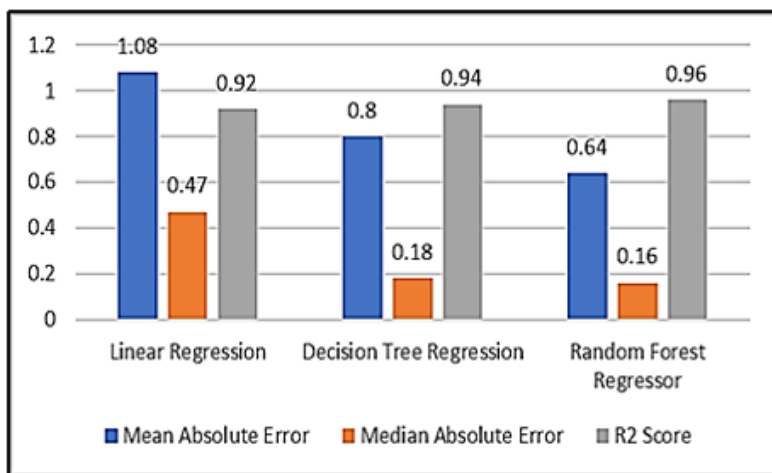


Fig. 8: Regression Results for Yield Prediction

Random Forest Regressor gives most reliable results when given required inputs with Mean Absolute Error of 0.64, Median Absolute Error is 0.16

and R2 score of 0.96. Decision Trees Regressor has Mean Absolute Error of 0.80, Median Absolute Error of 0.18, R2 Score of 0.94. Linear Regression

has Mean Absolute Error of 1.08, Median Absolute Error of 0.47 and R2 Score of 0.92.

Conclusion

Crop yield prediction is a complex process which relies on several different factors including weather, soil, fertilizers, pest infestations, etc. In this paper, we predict the crop yield using weather and soil parameters. The research is based on the datasets limited to districts in Maharashtra. The system incorporates regression techniques to estimate yield and multi-class classification to predict type of the crop. Among the used models for yield prediction, Random Forest Regression gives best results with MAE of 0.64 and R2 score of 0.96. For crop prediction, Naïve Bayes classifier gives most accurate results with accuracy of 99.39. The suggested method aids farmers in choosing which crop to plant in the field and how much yield any crop would give in that specific environment. Dataset used in the research can be improved by taking real time data through IoT devices. Also, various factors like irrigation and fertilizers use can be included for better prediction. Mobile App can be developed for mobile devices with added services like price

estimates in accordance with current market prices. Paid datasets may bring more reliable and accurate data which in turn might help in model accuracy. They may contain more features which may help correlate more with label.

Acknowledgements

We are grateful to Prof. Pritesh A Patil for providing valuable input and feedback throughout the research process.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Conflict of Interest

The authors declare no conflict of interest regarding this research. However, it should be noted that the first author of this paper is an employee of a company that develops and markets machine learning software for crop yield prediction. The results and conclusions presented here are solely based on the authors' research and do not reflect any external influence.

References

1. Alpayđın, Ethem. "Introduction to machine learning, second edition." MIT Press, 2010. ISBN: 978-0-262-01243-0.
2. Liakos, K.G.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D., "machine learning in agriculture: a review", *Sensors*, vol. 18, no. 8, pp. 2674, August 2018. <https://doi.org/10.3390/s18082674>
3. Sabitha, "A study on sectorial contribution of gdp in india from 2010 to 2019", *AJEBA*, vol. 19, no. 1, pp. 18-31, January 2020. Article no. AJEBA. 62227.
4. Jain A., "Analysis of growth and instability in the area, production, yield, and price of rice in India", *Journal of Social Change and Development*, vol. 2, pp. 46-66, N/A, 2018.
5. Wolfert S, Ge L, Verdouw C, Bogaardt MJ, "Big data in smart farming– a review. *Agricultural Systems*", vol. 153, pp. 69-80, May 2017.
6. Sangeeta, Shruthi G. "Design and implementation of crop yield prediction model in agriculture." *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, no. 4, pp. 305-310, Apr. 2020.
7. Johnson LK, Bloom JD, Dunning RD, Gunter CC, Boyette MD, Creamer NG, "Farmer harvest decisions and vegetable loss in primary production. *Agricultural Systems*", vol. 176, pp. 102672, November 2019.
8. Sharma A, Jain A, Gupta P, Chowdary V. Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access*. 2020 Dec 31;9:4843-73.
9. Meshram V, Patil K, Meshram V, Hanchate D, Ramkteke SD. Machine learning in agriculture domain: A state-of-art survey. *Artificial Intelligence in the Life Sciences*. 2021 Dec 1;1:100010.
10. Reddy, D. J., & Kumar, M. R. (2021). Crop Yield Prediction using Machine Learning Algorithm. 2021 5th International

- Conference on Intelligent Computing and Control Systems (ICICCS). doi:10.1109/iciccs51141.2021.9432236
11. Ranjini B Guruprasad, Kumar Saurav, Sukanya Randhawa, "Machine learning methodologies for paddy yield Estimation in India: a case study", 2019.
 12. S. P. Raja, B. Sawicka, Z. Stamenkovic and G. Mariammal, "Crop prediction based on characteristics of the agricultural environment using various feature selection techniques and classifiers," *IEEE Access*, vol. 10, pp. 23625-23641, 2022, doi: 10.1109/ACCESS.2022.3154350.
 13. Venugopal, Anakha, S, Aparna, Mani, Jinsu, Mathew, Rima, Williams, Vinu. "Crop yield prediction using machine learning algorithms." *International Journal of Engineering Research & Technology (IJERT) NCREIS – 2021*, vol. 09, no. 13, pp. 1-6, Dec. 2021.
 14. S. M. PANDE, P. K. RAMESH, A. ANMOL, B. R. AISHWARYA, K. ROHILLA and K. SHAURYA, "Crop recommender system using machine learning approach," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1066-1071, doi: 10.1109/ICCMC51019.2021.9418351.
 15. Suresh, N., *et al.* "Crop yield prediction using random forest algorithm." 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 279-282, 2021, doi: 10.1109/ICACCS51430.2021.9441871.
 16. E. Manjula and S. Djodiltachoumy, "A model for prediction of crop yield," *Int. J. Comput. Intell. Inform.*, vol. 6, no. 4, pp. 298–305, 2017.
 17. van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review in computers and electronics in Agriculture, 177, pp. 105709. doi: 10.1016/j.compag.2020.105709.
 18. M. Liu, T. Wang, A. K. Skidmore, and X. Liu, "Heavy metal-induced stress in rice crops detected using multi-temporal Sentinel-2 satellite images," *Sci. Total Environ.*, vol. 637-638, pp. 18-29, Oct. 2018.
 19. K. E. Eswari and L. Vinitha, "Crop yield prediction in tamil nadu using bayesian network," *Int. J. Intell. Adv. Res. Eng. Comput.*, vol. 6, no. 2, pp. 1571-1576, 2018.
 20. D. A. Reddy, B. Dadore, and A. Watekar, "Crop recommendation system to maximize crop yield in ramtek region using machine learning," *Int. J. Sci. Res. Sci. Technol.*, vol. 6, no. 1, pp. 485-489, Feb. 2019.
 21. Priya, P., Muthaiah, U., Balamurugan, M. "Predicting yield of the crop using machine learning algorithm." *International Journal of Computer Science and Mobile Computing*, vol. 4, no. 5, pp. 1-7, May 2015.
 22. Medar, Ramesh, S, Vijay, Shweta. "Crop yield prediction using machine learning techniques." *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 9, no. 5, pp. 1-6, May 2019.
 23. <https://www.kaggle.com/>