IME673A: Applied Machine Learning

Crop Yield Prediction

Submitted to
Prof. Veena Bansal

Submitted by:
Mohd Juned Khan (22114016)
Aditya Kumar Yadav (22114004)
Achyut Raj (22114003)
Sachin Khanchi (22114024)

Objective

The objective of crop yield prediction using machine learning is to develop models that can accurately predict the yield of a crop given various environmental and farming factors. Agriculture plays a critical role in the global economy. With the continuing expansion of the human population, understanding worldwide crop yield is central to addressing food security challenges and reducing the impacts of climate change. Crop yield prediction is one of the challenging tasks in agriculture. It plays an essential role in decision making at global, regional, and field levels. Agricultural yield primarily depends on weather conditions (rain, temperature, etc), pesticides and accurate information about the history of crop yield is an important thing for making decisions related to agricultural risk management and future predictions. Machine learning is an important decision support tool for crop yield prediction, including supporting decisions on what crops to grow and what to do during the growing season of the crops. Several machine learning algorithms can be applied to support crop yield prediction.

Input Feature:
Area,
Crop item,
Year, pesticides,
avg_temp,
average_rain_fall

Target variable:
Crop Yield

Result

1) The crop being potatoes has the highest importance in the decision making for the model, where it's the highest crops in the dataset. Cassava too, then as expected we see the effect of pesticides, where its the third most important feature, and then if the crop is sweet potatoes, we see some of the highest crops in features importance in dataset. If the crop is grown in India, makes sense since India has the largest crops sum in the dataset. Then comes rainfall and tempratures. The first assumption about these features were correct, where they all significantly impact the expected crops yield in the model.
2) India is the largest producer of Cassava, Potatoes, Sweet potatoes, Rice, Paddy and Wheat.
3) Japan is the largest producer of Yams.
4) Canada is the largest producer of Maize.
5) Guatemala is the largest producer of Plantains and others.
6) Mexico is the largest producer of Sorghum.
7) Brazil is the largest producer of Soybeans.

Models

1) Linear Regression

For Test Dataset
MAE 29560.91045089605
MSE 1783894627.0277524
RMSE 42236.176756753826
R2 score 0.7517263042968824
Adjusted R2 score 0.7484001495219124
For Train Dataset
MAE 29834.66684990498
MSE 1833665770.3418365
RMSE 42821.32378082019
R2 score 0.7463873302940077
Adjusted R2 score 0.7449422438569365

We can observe that the R2 score for both train and test dataset is reasonable. So, it the case of neither underfitting nor overfitting.

2) Polynomial Regression

For Test Dataset

MAE 435034744369.9456
MSE 1.6035594743231268e+27
RMSE 40044468710711.18
R2 score -2.231755345512107e+17
Adjusted R2 score -9.309419639181964e+17

For Train Dataset

MAE 10354.943187566392
MSE 327346024.075237
RMSE 18092.70637785395
R2 score 0.9547250647167361
Adjusted R2 score 0.9328434816027943

We can observe that the R2 score for train data is high and for test data is low. So, it is the case of overfitting.

3) Support Vector Regressor

For Test Dataset

MAE 44851.2464284077
MSE 6398464539.068653
RMSE 79990.40279351425
R2 score 0.10949312034941983
Adjusted R2 score 0.09756288464118235

For Train Dataset

MAE 45667.099279686816
MSE 6512786502.876216
RMSE 80701.83704771668
R2 score 0.0992223343343146
Adjusted R2 score 0.09408969806271528

4) Random Forest Regressor

For Test Dataset
MAE 5399.3866124493325
MSE 177210889.51508757
RMSE 13312.058049568728
R2 score 0.9753366584594443
Adjusted R2 score 0.9750062404866522

For Train Dataset

MAE 2102.6875245557508
MSE 26664374.587207817
RMSE 5163.755860534831
R2 score 0.9963120742424931
Adjusted R2 score 0.9962910604205131

5) Gradient Boosting Regressor

For Test Dataset
MAE 24897.241377953585
MSE 1652064330.9950194
RMSE 40645.59423842908
R2 score 0.7700737976441862
Adjusted R2 score 0.7669934466078403

For Train Dataset
MAE 24834.617735953936
MSE 1664300156.552293
RMSE 40795.8350392818
R2 score 0.7698121365832982
Adjusted R2 score 0.768500524825938

6) ANN

R2 score 0.9215656492720136

Learning from project

- We have learned multivariate outlier detection using isolation forest.
- Feature transformation(yoe-johnson) to remove skewness
- The 'Cumulative_explained_variance' shows linear relationship with the 'n_components' in the PCA that is why, its not reasonable to apply PCA. If we apply PCA with n components around 50 then it will explain less than 70% of the variance of the whole data.
- Machine learning is an important decision support tool for crop yield prediction
- Improvements in crop yields have been essential to feed a growing population, while reducing the environment impact of food production at the same time.
- To improve crop yields, It is very important to identify the factor affecting the crop yield.
- The four most important factors that influence crop yield are soil fertility, availability of water, climate, and diseases or pests.
- Machine learning models can help identify the key factors that affect crop yield, such as weather conditions, soil properties, and farming practices. By analyzing these factors, farmers can make more informed decisions about crop selection, planting strategies, and resource allocation.
- Predicting crop yield can also help farmers optimize their resource management
- By using machine learning models to predict crop yield, farmers can optimize their operations and increase their efficiency. For example, models can be used to predict optimal planting times, irrigation schedules, and pest management strategies.
- Machine learning models can also help farmers adopt more sustainable farming practices. By predicting crop yield, farmers can minimize waste, reduce chemical use, and improve soil health, leading to a more sustainable and resilient agricultural system.
- Finally, machine learning models can provide farmers with a tool to support their decision-making processes. By analyzing historical data and predicting future outcomes, models can help farmers make more informed decisions that improve their overall productivity and profitability.

In summary, learning from the prediction of crop yield using machine learning can help farmers make more informed decisions, optimize resource management, increase efficiency, enhance sustainability, and improve profitability.

Approach and Methodology

Predicting crop yield using machine learning typically involves the following approach and methodology:

Data Collection: The first step is to gather data on various factors that affect crop yield, such as Area, Crop item, Year, pesticides, avg_temp, average_rain_fall. This data can be collected through various sources, such as government agencies, weather stations, and farmers. We have got this data from Kaggle.
Data Preparation: Once the data has been collected, it needs to be cleaned and pre-processed to remove any errors or inconsistencies. We have preprocessed the data by checking for missing values, outliers, and other anomalies that could affect the accuracy of the predictions.

Feature Engineering: The next step is to identify the most important features that affect crop yield and create new features that could improve the accuracy of the predictions. For example, weather data can be combined with soil data to create new features that capture the interactions between these factors.

Model Selection: There are several machine learning models that can be used for crop yield prediction. We have used linear regression, decision trees, random forests, and neural networks. The choice of model depends on the complexity of the problem, the size of the dataset, and the performance metrics used to evaluate the models.

Model Training: The selected model is trained using the pre-processed data and the identified features. The training process involves adjusting the model parameters to minimize the difference between the predicted and actual crop yields.

Model Evaluation: Once the model is trained, it needs to be evaluated using a separate dataset to assess its performance. This involves comparing the predicted crop yields with the actual crop yields and calculating performance metrics such as accuracy, precision, recall, and F1-score.

Model Deployment: Once the model has been evaluated and found to be satisfactory, it can be deployed in the field to predict crop yields. This involves integrating the model into a user-friendly interface that can be accessed by farmers and other stakeholders.

In summary, predicting crop yield using machine learning involves data collection, data preparation, feature engineering, model selection, model training, model evaluation, and model deployment. The success of the approach depends on the quality of the data, the choice of features and model, and the accuracy of the predictions.