CROP YIELD PREDICTION

Megha Veeregowda
Data Science
University of Colorado Boulder,
Boulder, Colorado, United States
meve3341@colorado.edu

Om Sai Madala
Data Science
University of Colorado Boulder,
Boulder, Colorado, United States
omma5510l@colorado.edu

Brahmendra Charan Attanti
Data Science
University of Colorado Boulder,
Boulder, Colorado, United States
Brahmendra.Attanti@colorado.ed

ABSTRACT

Coming from India, A country where Agriculture used to be the primary occupation for several decades, but people are slowly moving away from agriculture into different sectors. One of the main reasons being the risk factor involved in the production of yield. There are several factors which have an impact on the yield like Climatic conditions - temperature, rainfall, Humidity, Kind of soil - PH value, the amount of fertilizers being used and some other factors which have an impact on the crop.

There is a need to develop a machine learning algorithm which can predict which crop suits the best based on the environmental conditions and estimate what yield we can expect from that crop. To get higher production of crops, they should be cultivated at the right place and right time. By using this machine learning model, we can reduce the risk of crop failure and guarantee maximum profit for farmers to sustain their livelihood.

KEYWORDS

Machine Learning, Crop Yield Prediction, DBSCAN, Random Forest, Ridge Regression, Multilayer Perceptrons (Neural Networks)

INTRODUCTION

India's economy is primarily centered on agriculture because it is a wholly agrarian nation. Because of the overall development of the industry, modernization should be implemented in agriculture. Automatically, India's economy rises as well. Therefore, we must try to increase crop production and generate large profits. so that the government may act to manage food risk, implement policies, and solve the issues that farmers are now dealing with. Using machine learning algorithms, such as linear regression, Decision trees, and K closest neighbors for comparison analysis, one must construct an effective model for predicting agricultural yield. This model must be studied and performed well. Our system uses the Ridge regression approach to achieve machine learning and make predictions.

LITERATURE SURVEY

Lots of research has been done on predicting the best crop based on the environmental conditions at that phase of time and some amount of research is going on to estimate how much yield we can get.

"Crop Prediction using Machine Learning [2020]- M.Kalimuthu , P.Vaishnavi"

It was about predicting which crop would suit the best based on the conditions surrounding it given as input parameters.

To develop a Supervised Machine learning model, they used Naive Bayes Classifier along with Boosting Algorithm to predict the class label(classification). This work helps the farmers who have less knowledge in predicting the crops for developing a sustainable future.

Research by Ji et al

This was about predicting yields in the Fujian Rice field. Two machine learning models were developed using Multiple Linear regression algorithm and Artificial Neural Networks and their effectiveness was compared.

They found out that ANN's were slightly better performing than Multiple Linear Regression algorithms as the RMSE score for ANN was less than that of MLR algorithm.

PROPOSED WORK

We will implement two machine learning models; the first one is predicting the best crop(classification) based on the input parameters which are nothing but the environmental conditions.

Then this Crop along with the input parameters are sent to another machine learning model which predicts the yield which inturn calculates the profit (Regression) based on the market price of that particular crop.

We are going to analyze the data with respect to different crops by using the DBSCAN Algorithm.

For the Classification model we will be using Logistic Regression, Random Forest and the XGBoost algorithms and for the Regression model we will be using Ridge regression.

Process we followed -

1 - Based on Conditions

2 - Predicting Optimal Crop (Classification)

3 - Respective yields based on Market price (regression)

IMPORTANCE OF OUR PROJECT

A web tool called Crop Recommendation helps farmers decide what crops to produce in their fields.

Farmers can use the application to get advice on choosing the right crop and fertilizer. Utilizing this application will boost crop yield and suggest appropriate crops.

Predicting crop yields is crucial for maximizing agricultural productivity. An accurate crop yield prediction model can help farmers to decide on what to grow and when to grow.

STEPS UNDERTAKEN

Data collection is the 1st step we have in which we have searched for a dataset in various websites and found a dataset which has all the required attributes to work on and do some analysis.

Exploratory Data Analysis is the crucial process of doing a preliminary examination of data to uncover patterns, spot anomalies, test hypotheses, and validate assumptions using summary statistics and graphical representations.

Exploratory data analysis is a procedure used in data analytics to fully comprehend the data and discover its many aspects, frequently using visual aids. This enables you to better understand your data and identify insightful patterns in it.

Once the EDA part is done, we have performed **data pre-processing** in which we have made various transformations on the attributes of the data set like standardizing them, changing the categorical variables into numericals so that our machine learning model can give us better outputs.

Now, after preprocessing we have the data ready in hand and we proceed to **feature selection** in which we will get an idea of which attributes are important i.e., which are having the most effects or impact on the target label. This step is mainly useful when we have more attributes, we will only pick some important columns from them as ML models don't perform well when the dimensionality of data is more.

Here comes the **Modelling** part in which we have used four models. For the Classification model we will be using Logistic Regression, Random Forest and the XGBoost algorithms and for the Regression model we have used Ridge regression.

Once the modeling is done, we will use some evaluation metrics to see which models have performed and based on the models also we can say which attributes are the most important and then using the best performing model, we will use that in the final implementation of our project.

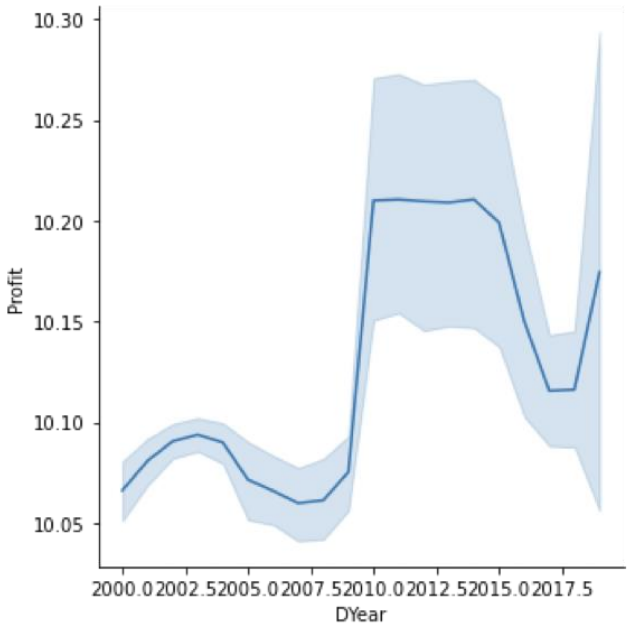
METHODOLOGY

In the proposed framework, we have used different machine Learning algorithms to estimate the crop yield based on different factors like rainfall, humidity, temperature, pH value of the soil, and fertilizer compositions. Here the fertilizer composition is the Nitrogen, Phosphorus and Potassium contents in soil which affects the crop yield. In the proposed architecture, we are going to use Logistic Regression with Multinomial Probability Distribution to predict the crop based on different factors. Here we are training the Regression model with temperature, humidity, ph, rainfall, and Nitrogen, Phosphorous, and Potassium features. Based on these conditions, we are going to predict the best crop.

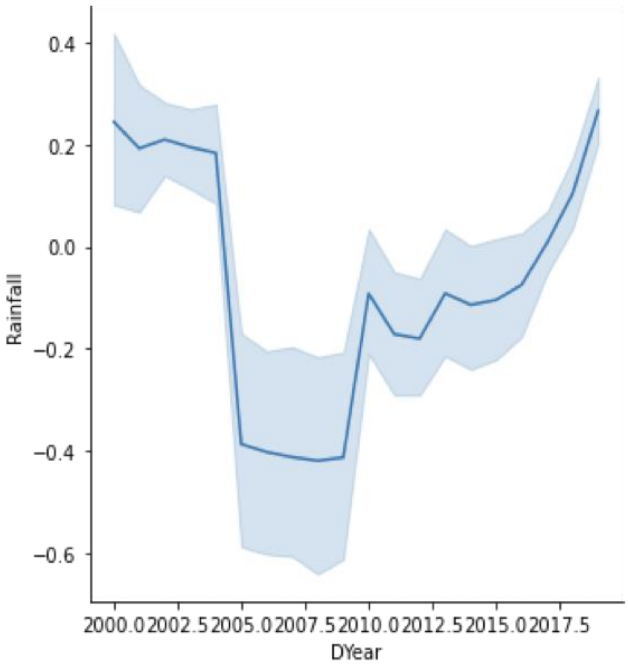
After predicting the best crop, we are going to predict the yield of the crop using different features like Investment, Market Price, Seeds used etc. This is the second model in the Machine Learning Pipeline. Here we are predicting the yield of the crop using Ridge Regression, XGBoost algorithms. The output of the first model will be the input to the second model. Yield of the crop is highly dependent on different factors like Investment, Market Price, and amount of seeds used per hectare. We train the model with these features and predict the crop yield. Here we have chosen Ridge Regression because of high dimensionality and collinearity of the data. Ridge Regression will penalize different features that have correlation between them, but it won't make the features zero.

RESULT ANALYSIS

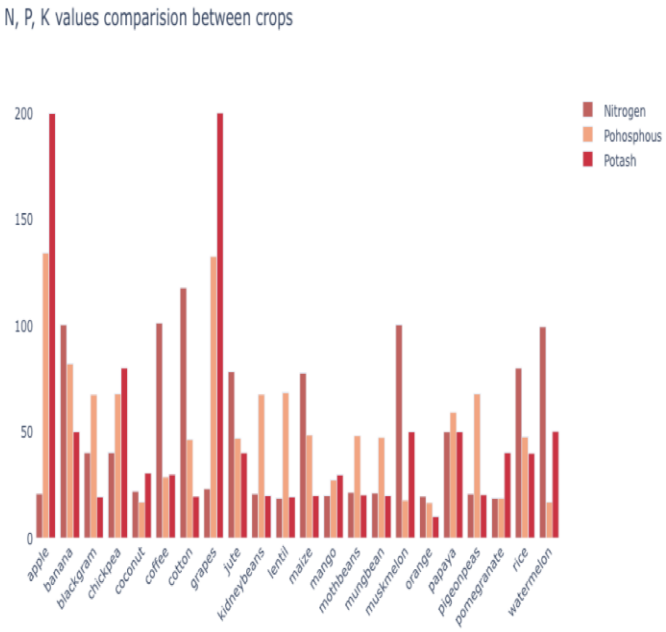
After the code has been executed, a graphic representation of the response is shown for each of the original parameter values. These variables include the use of pesticides, fertilizers, water, area, and the sun. Based on the data for these parameters, the yield is predicted:



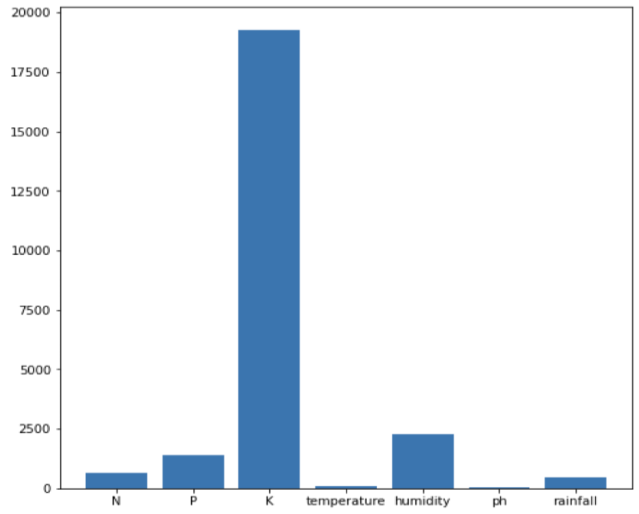
The above plot is the variation of profit for 2o years from 2000 to 2020. From the plot the profit was very low for about 5 years from 2005 to 2009. This can be due to many factors like temperature, humidity, and rainfall variations during that period. We know that 2005 to 2008 was the period of Economic and Industrial crisis. During this period agriculture is one of the fields that was highly affected due to unavailability of fertilizers and seeds for the crops.



The above plots are the variation of rainfall and Average Temperature for about 20 years. If you consider the profit variation, we came to know that profit from the crops is very less during the period of 2005 to 2009. If the results of profits are compared with the rainfall and temperature variations, we can see that the rainfall is very less during this period. and the temperature is very high during this time. From the above comparison, we can say that the rainfall and temperature are one of the important factors that will influence the crop production and thereby the profits from the crops.



These are the optimal proportions of fertilizers - Nitrogen, phosphorus, and potassium values for each crop. We can't say that one chemical is the best among the three and it will be better to use that one in more proportions because every crop has different requirements or suitable conditions in which it will be grown. Based on the crop requirements we are supposed to add fertilizer in those optimal proportions to get better yields.



The above graph represents feature selection. It is in most cases important, especially when we have more attributes, we will only pick some important columns from them as ML models don't perform well when the dimensionality of data is more.

We can see from the graph that potassium along with humidity are two key features which have a strong impact or effect on the species, mainly on the crops in Mysore regions as our data is majorly concentrating on that part.

MODEL RESULTS

Classification Part

The algorithms used for classification problem, which is nothing about predicting the best crop based on the given set of conditions are Logistic regression, Random Forest and XGBoost techniques.

The method of modeling the likelihood of a discrete result given an input variable is known as **logistic regression**. The most popular type of logistic regression models a binary result, such as true or false, yes, or no, and so on. Using multinomial logistic regression, events with more than two distinct possible outcomes can be modeled. When attempting to establish which category a new sample most closely resembles, classification problems are a good place to employ logistic regression as an analysis technique.

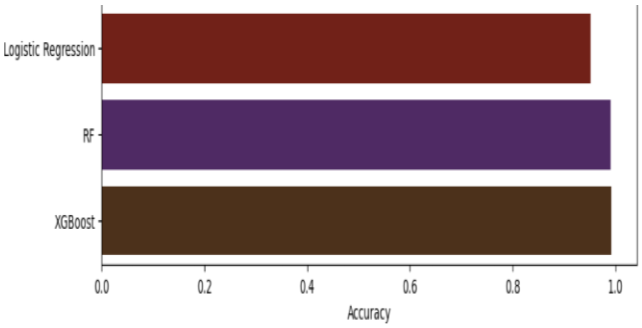
The other two techniques, namely Random Forest and XGBoost are ensemble techniques.

In **Random Forest**, n records at random are selected from a data set with k records. After that, distinct decision trees are built for every sample. Each decision tree will produce an output, and the final output for classification and regression, respectively, is based on majority voting or averaging.

XGBoost is one of the best machine learning models going around whether it be classification or regression problem. Data scientists are using it so often because of the better results it is giving compared to many other machine learning algorithms in most cases.

Extreme Gradient Boosting (XGBoost) is a distributed, scalable gradient-boosted decision tree (GBDT) machine learning library. In contrast to gradient boosting, which operates as gradient descent in

function space, XGBoost offers parallel tree boosting. A second order Taylor approximation is employed in the loss function to provide the connection to the Newton Raphson technique.




```
Logistic Regression --> 0.9522727272727273
RF --> 0.990909090909091
XGBoost --> 0.9931818181818182
```

We can see that as expected Ensemble techniques were giving better outputs than the logistic regression, we have used with multinomial distribution instead of the binomial logistic which can predict only two class labels. So, as our problem was a multi-class classification as our target label crop involves around 15+ categories we had to use logistics with multinomial distribution. But all the three algorithms we used have performed well enough.

Regression Part

For predicting or estimating the yield based on a given set of conditions with some specific crop we have used Ridge regression. The main idea of using this type of regression is that it gives the best results when there are conditional dependencies among the data attributes. Our data set has attributes like Rainfall, temperature, humidity and we can say that more rainfall leads to increase in humidity and lower temperature values. So, when these kinds of conditional dependencies are present, ridge regression performs the best.

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(h_\theta(x^{(i)}), y^{(i)}) + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

Any data that exhibits multicollinearity can be analyzed using the model tuning technique known as ridge regression. This technique carries out L2 regularization. Predicted values differ much from real values when the problem of multicollinearity arises, least-squares are unbiased, and variances are significant.

The penalty term is lambda. The ridge function's alpha parameter serves as a placeholder for the provided here. So, we may regulate the penalty term by varying the values of alpha. The penalty is greater with larger alpha values, which reduces the magnitude of coefficients.

We have received an **RMSE** value of 0.635 which is very good in case of estimating the yields, this much amount of error should be pretty much fine.

EVALUATION METRICS

To evaluate the **classification** algorithm implemented using Random Forest, we will use Classification Report and confusion matrix.

The **classification report** provides a summary of how well the built-in ML model classified data. It mostly consists of (N+3) rows and 5 columns. The name of the class label appears in the first column, and is followed by Precision, Recall, F1-score, and Support. Three rows are for accuracy, a macro average, and a weighted average, and N rows are for N class labels.

Precision: It is determined in relation to the anticipated values. The precision for class-A is determined by how many of the total predicted values correspond to class-A in the actual dataset. The confusion matrix [i][i] cell to [i]column ratio is what determines this.

Recall: It is computed considering the dataset's real values. The recall for class-A is the proportion of entries in the dataset that the ML model correctly identified as belonging to that class. It is the ratio between the sum of the I row and the [i][i] cells of the confusion matrix.

The F1 score is the harmonic mean of recall and precision.

A classification problem's prediction outcomes are compiled in a confusion matrix. Count values are used to describe the number of accurate and inaccurate predictions for each class. This is the **confusion matrix's** secret.

To evaluate the **Regression** algorithms implemented using Ridge Regression we will use RMSE score.

Root mean square error or root mean square deviation is one of the most used measures for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}},$$

Conclusion

To conclude that we can say that in classification algorithms ensemble techniques were giving better results than logistic regression and attributes like Rainfall, temperature, humidity, pH, and different fertilizers which involves chemicals like Nitrogen, phosphorus and potassium are the most important attributes which are to be given utmost consideration to yield better results.

FUTURE IMPACT

We all know that many people throughout the world make their living through agriculture, and it plays a crucial role in our daily lives by giving us access to resources, employment possibilities, and technology in addition to food. It has served as the basis for human existence and survival. It has been a source of food, nourishment, and medicine for at least several thousand years. Agriculture can help reduce poverty, raise incomes, and improve food security for 80% of the world's poor, who live in rural areas and work mainly in farming. The farmer will learn which crops can be grown in the field at that specific moment through crop prediction and yield advice. The difficulty for the farmer in selecting the best crop for their field is reduced by this technique. By analyzing the data, it offers recommendations for the farmer to grow crops that are appropriate for the soil. The proposed model helps in increasing agricultural production and reduces the money and time of the farmer.

One scenario in which our project can help is that as crop production takes few months of time, we can implement sensors in the crop field and from time to time on a weekly basis or so figuring out the conditions at that instant in the crop field and then comparing them with the optimal conditions for that particular crop and if in case any fertilizers composition is less than optimal proportions then we should add it. Doing this repeatedly will help us to make better profits.

REFERENCES

[1] D. J. Reddy and M. R. Kumar, "Crop Yield Prediction using Machine Learning Algorithm," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 1466-1470, doi: 10.1109/ICICCS51141.2021.9432236

[2] R. J, V. K. G. Kalaiselvi, A. Sheela, D. S. D and J. G, "Crop Yield Prediction Using Machine Learning Algorithm," 2021 4th International Conference on Computing and Communications Technologies (ICCCT), 2021, pp. 611-616, doi: 10.1109/ICCCT53315.2021.9711853

[3]https://www.researchgate.net/publication/348020414_Crop_Yield_Prediction_Using_Machine_Learning_Techniques

[4] F. F. Haque, A. Abdelgawad, V. P. Yanambaka and K. Yelamarthi, "Crop Yield Prediction Using Deep Neural Network," 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), 2020, pp. 1-4, doi: 10.1109/WF-IoT48130.2020.9221298