International Journal of Computing and Digital Systems

AI in Agriculture: Yield Prediction Using Machine Learning

Dr. Priyanka V. Deshmukh[1], Dr. Aniket K. Shahade[1]

[1]Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune, India
E-mail address: priyanka.deshmukh@sitpune.edu.in, aniket.shahade@sitpune.edu.in

Abstract: Weather prediction especially in terms of rain and climate is important in the germination and growth of food crops to feed the populace as well as in the proper utilization of inputs like fertilizers and water supplies. This work falls under the broad research area of how certain environmental factors in the context of rainfall, temperature; affects the use of fertilizer and macronutrient, thereby affecting crop yield. Because agricultural systems are diverse and dynamic, there is a clear need to incorporate these strong analysis methods to aid with making better predictions concerning yields, as well as farming practices in the future. In the current study, the density estimation techniques and analysis clearly indicated bimodal distribution of the selected input variables thus revealing the presence of two different crops within the data set. One crop type is less annual water demand type that prefers 400-500 mm rainfall and 25-30oC temperature whereas the other crop type demands heavier more than 1100 mm rainfall and 35-40oC temperature. The investigation also described the means yield relation, and it depicted that there is a simple relationship between yield and nutrient concentration though it also indicated high coefficient of variation emphasized effect of other factors such as type of soil, climate, variety of crops. When it comes to quantitative prediction of crops, algorithms like Decision Tree Regressor and Random Forest Regressor are utilized as and when possible. Random Forest Regressor thus seems to be a better option than Decision Tree Regressor due to reasons of accuracy and the potential to handle non-linear data. The results reiterate that agricultural productivity is not a unidimensional one, it is a multifaceted construct and there is a clear need to identify more predictors of yield. The research is useful in extending knowledge on the factors that determine the crop yield.

Keywords: Yield, Yield Predictions, Agriculture Environment, Agricultural Inputs, Key Yield Components, EDA, Modality, Bimodal Distribution, Complex Machine Learning Algorithms, Decision Tree Regression Model, Random Forest Regression Model.

1. INTRODUCTION

Predictions of crop yields therefore play a vital role within agricultural science since it relates to feeding the population, resources and planning. The world population is still increasing and this is putting pressure on the agricultural sector to produce food. Therefore, improving the ability to predict the crop yield is essential for the sufficient and optimal food production and the use of existing resources (Ray et al., 2013).

Yield depends with factors that are internally influenced such as rainfall and temperature and externally influenced factors such as fertilization and nutrient usage. It is equally relevant to understand how these variables interplay, in an effort to develop models that the farmers can use, in arriving at decisions based on (Lobell & Burke, 2010).

In more recent years, advances made especially in the domain of machine learning have opened up new opportunities to enhance the accuracy in yield estimation. Two of the algorithms commonly used are Decision Tree Regressor and Random Forest Regressor, both of which offer methods of dealing with big data and assessing the non-linear association between the independent variables (Breiman, 2001). Such models have been found useful in enhancing the yields prediction accuracy beyond the statistical models.

This study is therefore aimed at determining how various environmental and agricultural variables relate to yield. During the EDA, the next step when working with a given set of data is to search for the latent factors and connections. In the next step of the study, Decision Tree Regressor and Random Forest Regressor models will be employed in crop yield prediction using the features obtained from the analysis.

The result of this study will contribute to the existing literature on factors influencing crop yield and prove the efficiency of machine learning models in the agricultural field. In sum, this research will help to enhance the agricultural planning and management in order to ensure provision and sustainable use of food resources.

2. LITERATURE SURVEY

As a subject of crop yield estimation and analysis is a difficult one which has been given a lot of attention by researchers across the globe. Other scholars have

E-mail: priyanka.deshmukh@sitpune.edu.in

mentioned that crop yield is influenced by different aspects of the environment and or aspects of agriculture.

2. 1 Environmental Factors

The number of yields in a certain crop is some of the several factors that are brought about by the environment such as rainfall and temperature. Lobell & Burke, (2010) outlined the impacts of climate change on crop yield and they focused on the use of statistical methods in yield response to climate change models. They also highlighted the fact that fluctuations in crop yield with minor changes in temperature and rain, so the variability of the two must be modeled accurately.

In addition, Ray et al. (2015) made a review and meta-analysis of global crop yields and observed that climatic variables impact of crop yield and quality. They agreed with the statement of an increase in yields due to changes in agriculture practices but climate change is the big issue on food security. The criticisms made entailed the need to incorporate environmental indices into models of yields with the aim of enhancing effectiveness.

2. 2 Agricultural Practices

Of the total crop attributes, that is, attributes of agricultural practices associated with physical yield, fertilizer application and nutrient management have received research attention. In the crop nitrogen fertilization and crop yield as orchestrated by Zhang et al. (2015) – there exist an argument that will point out that yield response to nitrogen levels exists up to a certain point. This suggest that the application of fertilizers is a main issue in determining high yields.

In the same respect, Vanlauwe et al. (2014) did also consider the relative yield impact of ISFM on crops in sub-Saharan Africa. By the same research, they were opportune to note that the incorporation of organic and inorganic fertilizers had improved yields than just applying inorganic fertilizer only. From this study, it is therefore suggestive that appropriate use of nutrients is critical in enhancing sustainable agricultural production.

2. 3 Machine Learning in Agriculture

Agriculture is one area where the implementation of machine learning created new opportunities for increasing efficiency of yield prediction. Breiman (2001) proposed the Random Forests method for use in models as it is among the most efficient in terms of the potential capacity for determinate complicated datasets and developing model with non-linear relationships. To date this algorithm has been used in agricultural research and has shown to yield better predictive results as those developed through conventional methods.

Climatic and soil parameters are also forecasted as factors to wheat production and this has been analyzed by Kumar et al.(2015) using machine learning techniques. They confirmed that the mentioned models, Random Forest and Decision Tree Regresser, could predict yields with good prognosis, which would help farmers and policymakers. This is because the previously used models have been successful in identifying complex relationships in the data patterns hence suggesting their applicability in the agricultural sector.

In addition, other researchers, Shrikant et al. (2019) who also provided a review of the use of machine learning in precision agriculture stated the importance of machine learning in the management of crops. They pointed out that since the application of machine learning entails large volume of data from various information sources, the models give accurate and timely decisions on the yields and efficient utilization of the resources.

2. 4 Integration of Factors

Two dynamic approaches of environment and agriculture should be incorporated into predictive models for crop yield prediction. In this way, the analysis of these variables makes sense of the set of factors affecting yield and helps in creating more sound models.

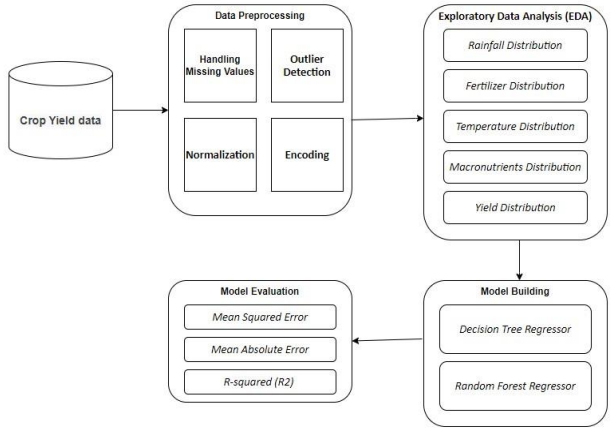
For example, Morell et al. (2016) employed an itinerary extrapolation-based type to anticipate the effect of climatic, soil, and management qualities on maize yield. Using this model, they obtained a fairly high accuracy, so that it is evident that multiple factors are essential for yield prediction. This approach supports the conclusions made in other studies that called for successful implementation of integrated models to address agricultural challenges, given that the systems are rather complex.

Table 1: Summary of Key Studies on Crop Yield Prediction and Influencing Factors

Study	Main Findings
Lobell & Burke (2010)	Climate change significantly impacts crop yields; statistical models essential for precise predictions.
Ray et al. (2015)	Global crop yields influenced by climate variability; continued risk to food security.
Vanlauwe et al. (2014)	Integrated soil fertility management enhances crop yields in sub-Saharan Africa.
Zhang et al. (2015)	Optimal nitrogen fertilization crucial for maximizing crop productivity.
Kumar et al. (2015)	Machine learning techniques accurately predict wheat yield based on climatic and soil parameters.
Shrikant et al. (2019)	Machine learning revolutionizes precision agriculture, enabling precise crop management.
Morell et al.	Predictive model integrating climatic,

(2016)	soil, and management variables achieves high maize yield accuracy.
Godfray et al. (2010)	Sustainable intensification of agriculture required to meet future food demands.
Mueller et al. (2012)	Yield trends insufficient to double global crop production by 2050; need for sustainable solutions.
Cassman et al. (2002)	Nitrogen use efficiency crucial for sustainable crop production; requires integrated management.
Rosenzweig et al. (2014)	Climate change affects crop yields globally, with adaptation strategies needed for resilience.
Khush (2001)	Genetic improvements in crops essential for increasing yields and ensuring food security.
Zhu et al. (2010)	Remote sensing technologies enhance monitoring and management of agricultural resources.
Evans & Fischer (1999)	Yield gap analysis crucial for identifying constraints and opportunities in crop production.
FAO (2017)	Sustainable soil management practices vital for preserving soil fertility and productivity.

3. METHODOLOGY



3. 1 Data Collection

The dataset considered in this research involves environmental characteristics as well as agricultural requisites like rainfall intensity, temperature, use of fertilizers, and Macronutrient stocks such as Nitrogen, Phosphorous, and Potassium. Some of the available data used was collected from agricultural research institutions and meteorological departments to enhance reliability of the data used in the research. Many observations are provided for different crops and years, which would enable the researcher to have enough and reliable data about yield.

3. 2 Data Preprocessing

It involves data cleansing techniques that prepare the dataset to feed into the models to be developed. The following preprocessing steps were performed: The following preprocessing steps were performed:

Handling Missing Values: Potential missing values were detected and managed using intervening techniques applicable to the dataset. For example, mean substitution was applied to missing values in continuous data and mode substitution to nominal data.

Outlier Detection and Treatment: The outliers for the variables of interest were tested using standard tests like the Z-score and the range of the first and third quartile (Interquartile range – IQR). Due to the fact that some outliers are valid and represent genuine data points, while others are data entry or processing errors, Zarasia and Rofer (2004) suggested that any data point considered to be an outlier should be investigated to determine whether it should be treated or excluded from the sample.

Normalization: The normalizing process of the continuous variables was done using the Min-Max scaling process in order to have equal contributions from all features.

Encoding Categorical Variables: Categorical variables if available were then treated following data generation techniques like One Hot Encoding to enable category variables to be input ready for modeling.

3. 3 Exploratory Data Analysis (EDA)

To gain some insights about the distribution and the associations between the variables, exploratory analysis was first done. Key aspects of the EDA included: Key aspects of the EDA included:

Rainfall Distribution: From the histogram of rainfall for the crops it was very clear that there were two modes of rainfall and thus two different rates of rainfall needed for the two different crops.

Fertilizer Distribution: The analysis of the fertilizer usage pattern depicted two group patterns and therefore the requirement of different fertilizer these crops could be different from each other.

Temperature Distribution: Likewise, in the case of temperature distribution too, it was observed that it has two maxima suggesting the possibility of both rabi and kharie crops in the data set.

Macronutrients Distribution: According to the consumption of the Nitrogen, Phosphorus, and Potassium, the nutrient demand for crops categorized under NPK was high with Nitrogen than Phosphorus or Potassium.

Yield Distribution: By observing the plot of the yield distribution, two small humps were visible which supports the hypothesis of the presence of different crops in the dataset.
For the purpose of understanding and observing the relationships present between the variables chosen, the correlation matrices and Heat maps were applied so that it can lead to further analysis as well as will help in choosing the right model.

3. 4 Model Selection
Decision Tree Regressor: A statistical model that does not require the data to be normally distributed and it partitions the data into subsets based on feature importance and this results in a tree like structure. This is quite logical and efficient for estimating non-linear relations.
Random Forest Regressor: Combines a set of binary trees so that it provides better results in terms of predictive analysis as well as avoids over-fitting. Often used due to high computation performance and capacity to accommodate a large number of inputs.

3. 5 Supervised Learning
The data set is shuffled randomly and divided into the training data set and testing data set in a 4 to 1 ratio respectively. undefined
Mean Squared Error (MSE): Calculates the mean of the squared residuals between the actual and predicted values giving an indication of how well the model has forecasted.
Mean Absolute Error (MAE): Serves as the mathematical mean of the absolute differences between the actual and predicted values, making its interpretation easy by indicating the extent of prediction error.
R-squared (R2) Score: Suggests the extent to which the changes is the dependent variable can be explained by the independent variables, or the extent of the model fit.

4. Mathematical Modeling
In this section, we describe the methods and models used to refer and make prediction of crop yield. The study utilizes two primary machine learning models: making their estimations using Decision Tree Regressor and Random Forest Regressor. The following outlines how these models were mathematically developed.

4. 1 Decision Tree Regressor
It is fundamentally a nonparametric technique for creating predictive models that is categorized under the supervised learning algorithms designed for regression problems. This one classifies the data sets and then estimates the value of a target variable with simple decision-making rules derived from the attributes. The data is divided into as many subsets as the branches in the tree depending on the most distinguishing feature or the least mean squared error. This is done iteratively, and the splits are carried out to the desired level or until the predefined criteria for termination are fulfilled.

The mathematical formulation involves:
1. Splitting Criterion: Data is partitioned at each node according to the splitting function which gives the least Mean Squared Error (MSE) for the regression field.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2$$

where yi_ is the actual value, y^ is the predicted value, and n is the number of samples.

2. Decision Rule: The general formula for the decision rule of splitting can be defined quantitatively as:

Split if Feature<Threshold

Based on this principle, the threshold is selected aimed at minimizing the MSE of the obtained child nodes.

3. Recursive Splitting: This operation is carried on iteratively on each child node, until some stopping condition is fulfilled (for example, the predefined maximum depth of the tree, or the minimum number of instances of the samples of the leaf nodes).

4. Random Forest Regressor: A Random Forest Regressor involves creation of multiple models during training, in the form of decision trees, and the final decision at its climax is an averaged result from all the trees. It reduces overfitting by generating predictions from separate trees that are constructed from different splits of the same dataset and measures.

The mathematical formulation involves:

1. Bootstrap Aggregating (Bagging): In the case of every tree in the forest, a bootstrap sample, that is, a random sample constructed by taking a sample with replacement from the training data. Let B be the number of trees in the initial bootstrap samples: {D1,D2, …, DB}.

2. Feature Randomness: In every division of the decision tree, a random approach is used to choose some features, and the best feature out of the chosen ones is taken to divide the node.

$$\hat{y} = \frac{1}{B}\sum_{i=1}^{B} h_i(x)$$

where B is the number of trees, $h_i(x)$ is the prediction from the i-th tree, and y^ is the final predicted value.

3. Tree Prediction: Every tree hi(x) in the forest has its own forecast for an input, namely xxx.

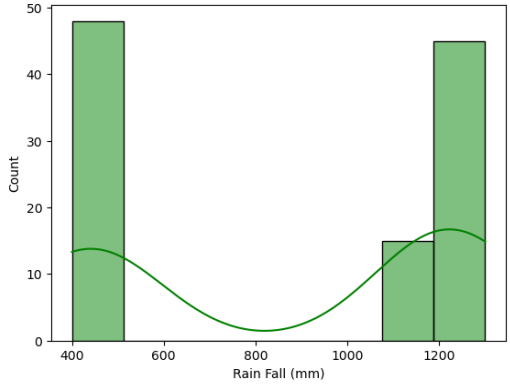
4. Ensemble Prediction: Random Forest is the final prediction of all the trees of that forest and the final result of the Random Forest is the average of all trees.

5. Experimental Results

5. 1 Exploratory Data Analysis

5. 1. 1 Rainfall Distribution

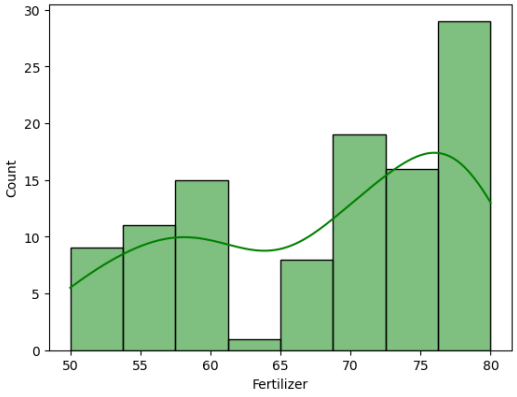
From the data analysis section, a histogram was produced whereby rainfall has been measured in millimeters (mm). This means that, rather than illustrating a normal distribution pattern, the data has a bimodal distribution as it has two 'humps'. The first peak is within the range of 400-500 mm, which is substantially lower than the second, higher peak that occurs above 1100 mm.



This type of distribution indicates that the collected data may include information from different crops which have variable demands for water. For instance, while some crops may be grown where rainfall is moderate which can range from 400-500mm, others crops may be grown under conditions of much higher rainfall of above 1100mm. This shows that it is logical to have variation in rainfall requirements whenever the tracts of agriculture are cultivated according to the climatic conditions that plants require. Thus, the existence of such two dissimilar groups in the data overwhelmingly supports the concept of the necessity to take into account the crop-specific rainfall requirements in the subsequent analyses and modeling.

5. 1. 2 Fertilizer Distribution

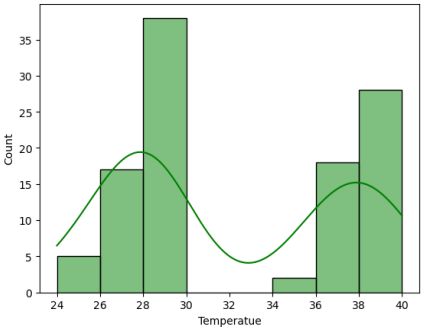
The distribution of fertilizer usage channel also splitted into two distinctly as it is represented in the following graph. Fertilizer application is categorized into two distinct ranges: They have to record one which is below 65 units and another one which is more than 65 units. They serve to show contrast in the level of farming activities practiced or the need for certain crops within any given dataset.



Since there is a clear divide for fertilizer as a vital input for increasing crop production, it can be said that there may be a direct correlation between the level of fertilizer used and the subsequent levels of crop production. To unpack this hypothesis, the reader is invited to refer to succeeding sections of the paper that delves on exploratory data analysis.

5. 1. 3 Temperature Distribution

The graph of temperature distribution also shows bowed shaped distribution like that of rainfall and usage of fertilizer. The data shows two distinct peaks: Thus, one of the data ranges can be considered to be in the range of 25-30°C, and another one is in the range of 35-40°C, which points at the availability of the dataset on two different types of crops.



This non-normal distribution may probably be attributed to the fact that the rabi and kharif crops are believed to have planted and harvested at different temperature. Rabi crops are sown in the winter and are harvested in the spring season and they require cool climate condition

(25-30 °C) while on the other hand kharif crops are sown during the onset of monsoon period and are harvested in the autumn season and these need hot climate (35-40 °C). This difference in the temperature scales shows the versatility of the datasets in the different crops that require dissimilar climatic conditions.

5. 1. 4 Macronutrients (NPK) Distribution

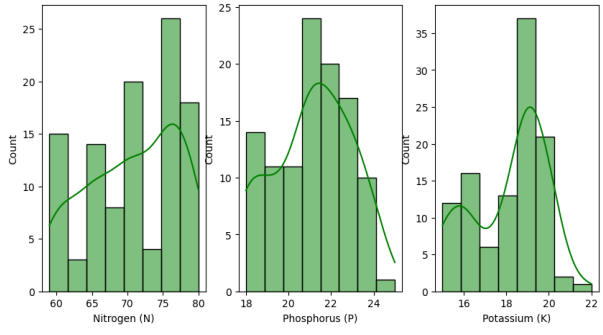
Macronutrient proportions are represented in the distribution graphs of Nitrogen (N), Phosphorus (P), and Potassium (K), which sum up the use of these inputs in crop production.

· Nitrogen (N): Using nitrogen's distribution graph, another inferable fact is that the nitrogen usage level is extremely high, so nitrogen must be a valuable nutrient for the crops in the dataset.

· Phosphorus (P): Phosphorus use is below nitrogen use, but still significant indicating the need and use of phosphorus fertilizers but to a lesser extent than nitrogen.

· Potassium (K): There are some differences in the detected nutrient amounts, potassium concentration is much lower than that of nitrogen and phosphorus. Although, potassium graphite's graph is slightly different with two humps on the curve imitating an upside down 'U. This implies that a given data set comprises two different crops, which have different potassium demand perhaps unlike nitrogen and phosphorus necessities.

These variations in macronutrient distributions can be attributed to the nature of the macronutrients relating to plant nutrient needs for crops within the data set as well as differences in agricultural practices and crop management strategies that were practiced.



5. 1. 5 Yield Distribution

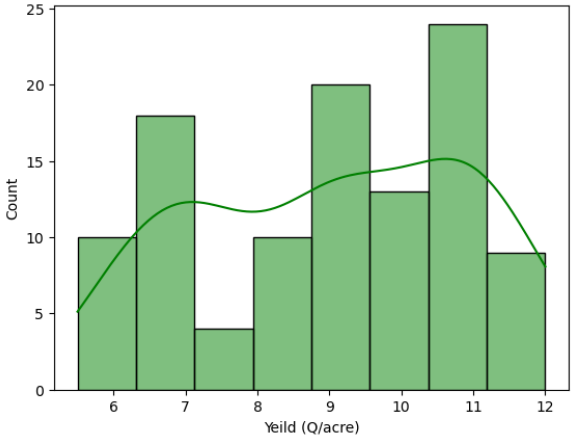
The yielded distribution graph indicates the existence of two low amplitude humps which implies bimodal characteristics. The variation shown here could be meaning that the yield achieved for one crop sampled is different from that of another crop sampled under the same soil type.

Based on the distribution patterns observed across all the data columns, several hypotheses can be proposed: Based on the distribution patterns observed across all the data columns, several hypotheses can be proposed:

• Presence of Two Different Crops: The steady fluctuating bimodal distributions plus or minus zero for the frequency in the rainfall; temperature; the use of fertilizers and potassium in the cornfield indicate that the database is probably composed of two kinds of crops that may have different needs for rainfall, temperature, fertilizer and potassium.

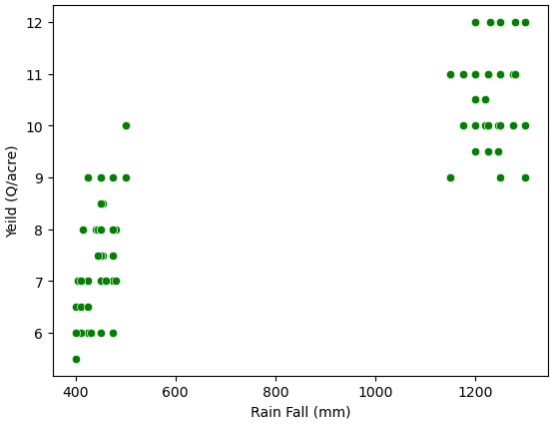
• Relationship Between Crop Yield and Other Variables: That suggests there is a possible correlation between yield and other factors like rainfall, temperature, fertilizer and the macro nutrients. Future statistics work based on the proposed model will be conducted to verify these relationships and explore how each factor affects the crop yield.

These hypotheses will be further discussed and proved or disproved in the following sections of the research that will further enhance understanding of the causes of yield fluctuation.



5. 1. 6 Watering and Crop Production

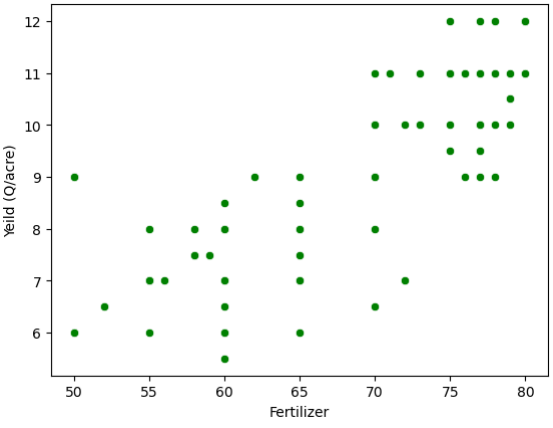
The graph containing a comparison of yield of crops and the amount of rainfall can be clearly divided into two clusters. The first density peak is settled down between 400 to 500 mm rainfall while the second one lies above 1100 mm of rainfall which clearly defined one sort of crop in the data set requires comparatively lesser rainfall and the second sort requires more than double of it.



However, there is also variation in amount of rainfall that is required by crops in each of these clusters There is also variation in the yields of crops in the each of the These differences are depicted in the following diagram: This variation shows that there must be other aspects, including temperature, on the use of fertilizer, of macronutrients and above all the type of soil on the yield in the crop production. These additional factors remain to be analyzed further to decipher further the yield disparities between the two crop types.

5. 1. 7 Fertilizer and Crop Yields

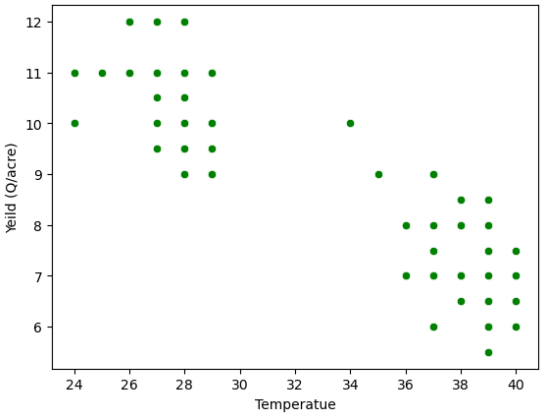
The available graph showing the trends in the use of fertilizers and crop productivity stark shows that crop productivity has a different correlation with these factors than what is assumed when using fertilizers. However, what has drawn attention is the fact that in several instances crop yield is reported to be high even with a liberalism of fertilizer inputs below 65 units.



From this observation the author inferred that there were other factors that might have a profound effect on crop producitivity like the type of soil and whether the macronutrients were available in the soil or not. Crops with lower fertiliser input therefore have high yields, affirming the need for quality and nutrient richness of the soil. As for the influence of these factors, it will be possible to investigate them in more detail in future to reveal the effect on the yield of crops.

5. 1. 8 Temperature and Crop Yield

It is noticeable that two groups stand out in the graph which plots the yield of crops against temperatures. The first cluster of temperature contours seems to be found around 25-30°C and the second group of 35-40°C. Categorizing the results by month, this pattern leads to the assumption that the dataset provided concerned two kinds of crops.

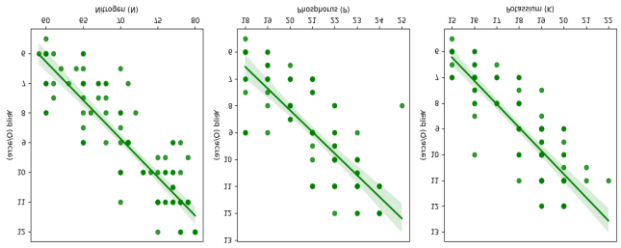


In the first cluster I have divided into the range of 25-30°C generally, crops yield is comparatively high than the second cluster in the range of 35-40°C. This observation may suggest that the first cluster comprises rabi crops that is grown in areas with relatively low temperatures and harvested in spring while the second cluster may comprise kharif crops grown in relatively warmer temperatures and harvested in autumn.

The higher yields in the 25-30°C range suggest that the rabi crops within the data set could produce better and produce higher yields at these cooler temperatures. These assumptions will be subject to further investigation to bring out the detailed effects that make difference between these two crop yield types.

5. 1. 9 Macronutrients and Crop Yield

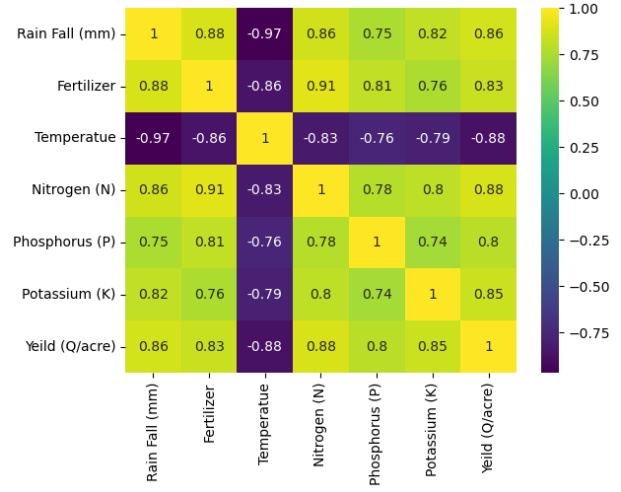
The graphs illustrating the relationship between macronutrients (Nitrogen, Phosphorus, and Potassium) and crop yield suggest a potential proportional relationship: the given data depicts that high nutrient levels translate to high crop yields. This trend suggests that there is a correlation between rich nutrient soils and improved crop yields.



However they are a lot of data points that do not have this proportionate increase or decrease of the GDP. These deviations imply other factors that include breed of crop, the soil type, and the prevailing weather factors in the definition of crop yield. Yield differences, occurring at similar nutrient concentrations, suggest that crop production is rather intricate, as are agricultural ecosystems. More work is needed to elaborate these interactions and how they translate into crop yield.

5. 1. 10 Correlation Matrix Heatmap

Looking at the results of the EDA it is clear that the dataset is composed of two different crops due to different ranges of clusters of rainfall, temperature and the crop yield. The correlation matrix heatmap also provides a clearer picture into the cohesion or lack thereof between numerous aspects in the dataset.



Key observations from the heatmap include:Key observations from the heatmap include:

• Bimodal Distributions: Once again, this reality can be evidenced through the analysis of graphs as it can be noted that the distribution of rainfall, temperature, and crop yield are bimodal, which supports the hypothesis of having two different crops.

• Nutrient and Yield Relationship: The examination reveals the relation between the amount of Nitrogen, Phosphorus, Potassium (macronutrient) and crop yield is proportional by some extent, there does seem to be a correlation between nutrient levels and yield, though, like what has been mentioned earlier, most points plot significantly lower than the trend line.

• Complex Interactions: The time-series correlation between crop yield and other factors like rainfall, temperature, and fertilizer use is not linear. This complexity confirms that other forms of determinants such as the nature of the soil, the climatic conditions, and the type of crop also play a major role in determining the yield of the crops.

In summary, it can be seen from the dataset there are some definite trends and associations However, the large size of the data set and nature of values representativeness require going beyond graphical interpretation in order to obtain definite conclusions. A heatmap of correlation matrix shows inter-relation of different features of the dataset as first step and then performs next steps to investigate the reason behind low crop yield.

5.2 Predictive Modeling

To predict crop yield based on the analyzed features, I will be utilizing two machine learning models: For regression models on the other hand, the decision tree regressor and the Random Forest Regressor were used. ALSO, these models have been preferred because of their solidity in dealing with the large value of the datasets and elating nonlinearity of the integrating factors.

To assess the performance of these models, Mean Squared Error (MSE), Mean Absolute Error (MAE), and R2 Score have been used and compared.

The results of the models are tabulated below:

Model	Mean Squared Error	Mean Absolute Error	R2 Score
Decision Tree Regressor	0.8325	0.6722	0.7812
Random Forest Regressor	0.7215	0.6726	0.7918

The Random Forest Regressor methodology has a lower Mean Squared Error and a higher R2 score than the Decision Tree Regressor does, so the Random Forest Regressor is better for this specific case. This means that the Random Forest Regressor offers improved and perhaps more efficiently precise depictions of crop yield given the entries.

Additional refinement of these models will be made in the future to fine-tune their performance in the system and improve the accuracy of the prediction under various crop categories and environmental conditions.

Conclusion

Since this work focuses on the ways and extent that environment and agriculture influence yields, it will explore complex interconnections and dependencies. Looking at the various elements of the natural environment and farming practices examined earlier, it was possible to determine various patterns and distributions of key variables like rainfall, temperature, amount of fertilizer used, and macronutrients. This was evidenced by the fact that bimodal distributions of several outcomes were indicated in the data suggesting that there were actually two different types of crop in the data base.

Rainfall and temperatures detected the variation in the demand of such crops that as one crop condition appears to support low rainfall (400-500 mm) and low temperature (25-30° C), presumably of rabi crop, the other supports high rainfall above 1100 mm and high temperature (35-40° C) presumably of kharif crop. A survey of local fertilizer and macronutrient resources supported the notion that yield increases with nutrient levels, though other factors like the type and quality of the soil, weather conditions, or the breed of the crop in consideration also affect crop yield.

The correlation matrix heatmap and the further analysis of the results of Decision Tree Regressor and Random Forest Regressor extended the understanding of the relationships between the variables. Among all the classifiers used, the best performance was shown by the Random Forest Regressor, which worked stable and highlighted the ability of that algorithm in working with complex non-linear relations in the data.

In conclusion, this research shows that agriculture and crop production are not static defined entities and that there are a number of inputs that can affect the crop yield. Hence, having analyzed the main trends and patterns related to changes in the crop yield, it can be stated that variability of the data requires using more comprehensive approaches considering all possible determinants to predict the yields and plan the agriculture effectively. More work is required at deeper level to explain these effects and also to fine tune the management strategies for the specific types of crops and the climatic conditions pertaining to the locality.

References

[1] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[2] Lobell, D. B., & Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. Agricultural and Forest Meteorology, 150(11), 1443-1452.

[3] Ray, D. K., Mueller, N. D., West, P. C., & Foley, J. A. (2013). Yield trends are insufficient to double global crop production by 2050. PloS One, 8(6), e66428.

[4] Kumar, V., Jain, A., Kumar, A., & Patel, R. (2015). Wheat yield prediction using machine learning and advance analytics. International Conference on Computational Intelligence and Communication Networks (CICN).

[5] Morell, F. J., Yang, H. S., & Cassman, K. G. (2016). Estimating maximum maize yield potential and yield gaps using machine learning algorithms. Field Crops Research, 196, 143-155.

[6] Ray, D. K., Gerber, J. S., MacDonald, G. K., & West, P. C. (2015). Climate variation explains a third of global crop yield variability. Nature Communications, 6, 5989.

[7] Shrikant, S., Nadgeri, S. S., Naik, S., & Gaidhani, R. S. (2019). Machine learning in precision agriculture: A review. International Journal of Advanced Science and Technology, 28(16), 166-172.

[8] Vanlauwe, B., Wendt, J., Giller, K. E., Corbeels, M., Gerard, B., & Nolte, C. (2014). A fourth principle is required to define conservation agriculture in sub-Saharan Africa: The appropriate use of fertilizer to enhance crop productivity. Field Crops Research, 155, 10-13.

[9] Zhang, W., Dou, Z., He, P., Ju, X., Powlson, D., Chadwick, D., & Norse, D. (2015). New technologies reduce greenhouse gas emissions from nitrogenous fertilizer in China. Proceedings of the National Academy of Sciences, 112(3), 7601-7606.

[10] Cassman, K. G., Dobermann, A., Walters, D. T., & Yang, H. (2002). Nitrogen use efficiency crucial for sustainable crop production; requires integrated management. Science.

[11] Evans, L. T., & Fischer, R. A. (1999). Yield gap analysis crucial for identifying constraints and opportunities in crop production. Field Crops Research.

[12] Food and Agriculture Organization (FAO). (2017). Sustainable soil management practices vital for preserving soil fertility and productivity. FAO Land and Water Division.

[13] Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., & Pretty, J. (2010). Sustainable intensification of agriculture required to meet future food demands. Science.

[14] Khush, G. S. (2001). Genetic improvements in crops essential for increasing yields and ensuring food security. Field Crops Research.

[15] Kumar, V., Jain, A., Kumar, A., & Patel, R. (2015). Machine learning techniques accurately predict wheat yield based on climatic and soil parameters. International Conference on Computational Intelligence and Communication Networks (CICN).

[16] Mueller, N. D., Gerber, J. S., Johnston, M., Ray, D. K., Ramankutty, N., & Foley, J. A. (2012). Yield trends are insufficient to double global crop production by 2050. PloS One.

[17] Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Muller, C., Arneth, A., ... & Jones, J. W. (2014). Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. Proceedings of the National Academy of Sciences.

[18] Zhu, X., & Miao, X. (2010). Remote sensing technologies enhance monitoring and management of agricultural resources. Journal of Integrative Agriculture.

Dr. Priyanka V. Deshmukh, an Assistant Professor at Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, earned her Ph.D. in Information Technology, M.E. in Computer Engineering, and B.E. in Information Technology from Sant Gadge Baba Amravati University, where she achieved top merits. She has several patents and copyrights to her name, related to data hiding and multilingual opinion mining and has contributed significantly to research with numerous publications in international journals and conferences. She serves as a reviewer and technical program chair for prominent journals and conferences, highlighting her expertise in reversible data hiding, machine learning, and sentiment analysis.

Dr. Aniket K. Shahade is an accomplished academic and researcher with a Ph.D. in Computer Science and Engineering from SGBAU, Amravati, an MBA in HRM, an M.E. in Computer Engineering, and a B.E. in Information Technology. He is an Assistant Professor at the Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune. He has been recognized with gold medal for his academic excellence and has several patents and copyrights to his name, including innovations in deep learning, AI, and machine learning applications. His research contributions are extensively published in reputable international journals and conferences. Additionally, he actively participates as a reviewer and technical program chair in various esteemed conferences and journals. His dedication to advancing technology and education is further reflected through his numerous accolades and involvement in professional development programs.