Crop Yield Prediction Using Machine Learning

Abstract:

Machine learning plays a vital role in various fields of agriculture. Crop yield prediction is one of the most important domains where it plays a major role. Not only yield prediction but also can guide us in farming in a much better way so that the yield is maximum and the cost is minimum. In this case study, we have closely examined the various models how it behaves through the given data sets. A detailed comparison is made. The main objective of the case study is the proper use of Machine Learning prediction power to estimate which crop should be grown, by keeping account of various attributes like temperature, rainfall, precipitation, and others. Farmers can benefit from yield estimation because it allows them to minimize crop loss and obtain the best prices for their crops.

Introduction:

India is a developing country, so it faces lots of issues, that could be resolved if handled with much accuracy and precision. For accuracy and precision, the power of computers and their learning capabilities can bring a massive breakthrough. Machine learning has long been used in agriculture. The crop yield prediction is one of the most difficult issues, and various models have been already implemented over this topic. As we know, the yield of any crop depends upon various factors like soil nutrients, rainfall, temperature, precipitation, soil texture, climate, environment and many more. The presence of so many deciding factors make it more complex to accurately predict the yield. Basically, regression approaches are being used for future prediction while descriptive models are being used to gain insight from collected data and clarify what has happened. To develop a high-performance predictive model, ML studies have to face lots of challenges like availability of accurate data, selection of proper algorithms, check the capability of the model to handle huge amounts of data, and accuracy checks.

Agriculture is the primary occupation for most Indians and contributes around 17% of the country's GDP according to 2018 statistics. Farmers solely depend on the knowledge which passed down through the centuries. Due to which farmers fail to estimate the exact amount of crops they need. For Indian farmers this became the main reason for their adversity. The number of suicide cases of farmers are increasing day by day due to low crop prices they get and high cost of development. Soil resources are depreciating, water is polluted, and farmer incomes are slowly decreasing.

Here comes the role of machine learning, where it can resolve lots of problems by accurately predicting the yield of crops. They can earn more if they know which crop will yield the maximum and when. Yield estimation impacted our economy in a positive way and helped lots of farmers earn their living in better way. Resources for machine learning are Satellite images, vast records of soil parameters, climate history which helps us to give useful information regarding the crop. There, we passed these data through various models and chosen the best out of it.

Literature Review:

In the research paper the author took the data of past 18 years, of various districts of Maharashtra. Data contains previous years’ climate, soil, and yield of various crops. They tried to minimize the error between actual and predicted values and have used Python language to build the regression models. They had used different machine learning models that are good at regression and giving the best results. Their performance are evaluated using metrics like R squared and Mean Absolute Error.

Authors handled various challenges like selection of proper dataset, data preprocessing to get the best result, training regression model using less computationally power and handling more error rate due to continuously changing the environment in the districts.

Dataset Description:

It includes district-wise different parameters like temperature, precipitation, humidity, soil type, crop type, area of the field, seasons. Since some models can not take strings as input so they had performed encoding technique. In place of label encoding, they have used one hot encoding technique which minimizes the problem of ranking. Label encoding can affect the interpretation of the model for values.

Proposed Framework:

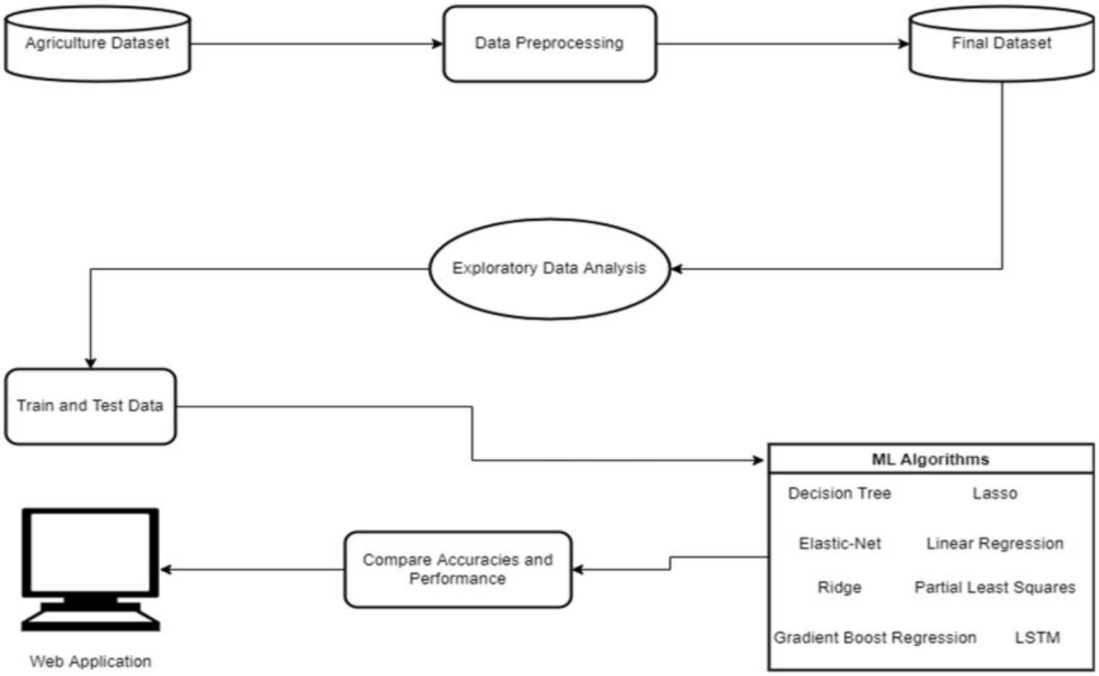


Fig. 1. Work Flow Diagram.

In the data preprocessing step they have removed the null, extreme outliers, and repeated values using various python libraries. After all these steps, they have performed Exploratory Data Analysis i.e, exploring the data set by taking the help of several visualization tools. During this, they had explored various relation between the parameters, found various patterns between them and they came to know how different factors are affecting the yield of crops.

After this they split the dataset into train and split and used them over various machine learning models to find out which ML model is working better. To analyze the model performance, they have used R-squared method (Coefficient of Determination). Then percentage is evaluated. To compare the efficiency of different model they have used Mean Absolute Error, Root Mean Square Error.

$$R^2 = 1 - \frac{Ss_{res}}{SS_{tot}}$$

R^2 = coefficient of determination
SS res = sum of squares for residuals
SS tot = total sum of the squares

$$RMSE = \sqrt{\frac{1}{n-p}\sum_{1=1}^{n}(y_i - \widehat{y}_i)^2}$$

p: Polynomial number

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x|$$

MAE = Mean Absolute Error

n = Number of errors
Σ = Symbol for summation
|xi – x| = Absolute error

Exploratory Data Analysis:

EDA involves use of various plots like scatter plot, histograms and many more to properly interpret the data and extract more out of it. It also involves the data preprocessing step. Since they have preferred one hot encoding over label encoding, they have to face another problem of multicollinearity. Multicollinearity is determined by Variable Inflation factor. According to their calculation they have Variable Inflation factor less than 5, which resolves the problem of Multicollinearity.

Feature Engineering:

Since they had also used a neural network for their training they needed more data. Here comes the concept of feature engineering where they have created new features depending upon the already present features. As they have used the concept of Deep learning, it requires more data to give better result. In order to increase the data they created new features like precipitation ration, precipitation humidity ratio, precipitation to mean temperature ratio. Then they normalized the data for the proper training of the LSTM model.

Web Application:

Since everything is now over internet, it would be better if this model is also deployed over internet. They had built a web application where any end user can give the required details and get the predicted crop yield. The output will be show over the webpage. To built this web application they had used Streamlit.

Methodology:

In this case study we have explored the Dataset and understood the various steps taken by the author to get better result. Different Algorithms like Linear Regression, Lasso Regressipon, Gradient boosting algorithm, Elastic Net Regression, LSTM model are used for training of the model. Out of these LSTM outperformed all these models and gave the best result. Below their is the description of various algorithm:

1. Multiple linear regression

MLR, also called multiple linear regression, is a mathematical method for predicting the result of an answer parameter by integrating several logical factors. It depicts the relationship between a continuous dependent parameter and several independent parameters. We chose to use MLR to predict yield because there was non-linear behaviour between our parameters.

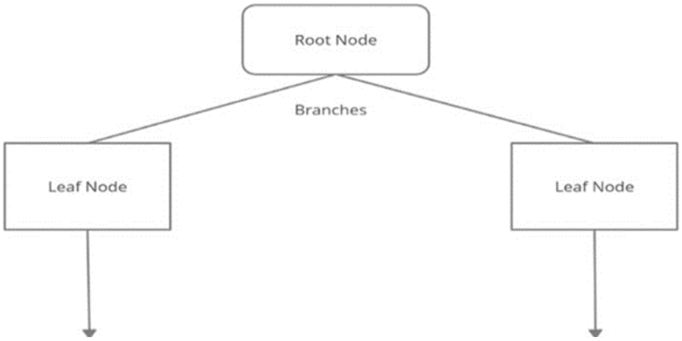
$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_i X_i$$

Y: Dependent Variable

X: Independent Variable

2. Decision tree regressor

The Decision Tree is a Supervised Machine Learning technique that learns decision tools from associated features to estimate a target value. It disintegrates a large dataset into smaller manageable subsets. A tree with leaf and decision nodes is the final product. A leaf node has more than two branches that represent the attribute values that have been evaluated. A minor alteration in the dataset can cause huge instability. They also take more time to train and are expensive.



Set of Possible Answers Set of Possible Answers

3. Elastic-Net regression

Elastic-Net is a normal regression approach that incorporates the Ridge and the Lasso regression's L1 and L2 penalties. Elastic net can deal with a large number of parameters and even when there exists a correlation between the parameters. It can decrease the effect of various parameters without completely removing all of the parameters.

$$\frac{1}{2m}\sum_{i=1}^{m}(y-Xw)^2 + alpha^{*} ratio^{*} \sum_{j=1}^{p}\left|w_j\right| + 0.5^{*} alpha^{*}(1-ratio)^{*} \sum_{j=1}^{p} w_j^{2}$$

4. Lasso regression

LASSO is known as Least Absolute Shrinkage and Selection Operator. It's a modified form of linear regression that implements shrinkage as a factor. The Data Values converge at a central position known as shrinkage, which is close to the mean. Lasso is very reliable when using variable selection or feature selection/elimination. Unlike MLR it can avoid overfitting even on smaller datasets. It is fast during training.

Cost(W) = RSS(W) + l * (sum of absolute value of weights):

5. Ridge regression

The Ridge Regression is a method for assessing multicollinear values in these models. Where there is multicollinearity, the least squares es-timates are without bias, but the values of variance are noted to be high, so they may be distant from the actual value.

$$\beta^{ridge} = argmin_{\beta \varepsilon Rp} \sum\nolimits_{i=1}^{n} \left(y_i - x_i^T \beta\right)^2 + \lambda \sum\nolimits_{j=1}^{p} \beta_j^2$$

$$argmin_{\beta \varepsilon Rp} ||y - X\beta||_2^2 + \lambda||\beta||_2^2$$

6. Partial least square regression

Partial least squares (PLS) is a common method for soft modeling. The responses are component quantities that the researcher hopes to predict in future samples. Where the parameters are multiple and strongly collinear.

$$max_{\alpha} Corr^2(y, X\alpha) Var(X\alpha)$$

$$subject\ to\ ||\alpha|| = 1,\ \widehat{\varphi}_t^T S\alpha = 0,\ l = 1, \ldots, m-1.$$

7. Gradient boosting regression

Gradient boosting is a machine learning technique that can be used for a variety of applications, including regression and classification. It returns a prediction model in the form of an ensemble of weak prediction models, most commonly decision trees. The resulting approach is called gradient-boosted trees.

$$F_m(x) = \ F_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} 1 R_{jm}(x)$$

$$\gamma_{jm} = arg_\gamma min \sum_{x_i \ \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma)$$

8. Long short-term memory (LSTM)

A Deep Learning approach for modeling sequential data is Recurrent Neural Networks (RNN). The main principle of RNN is that it, saves the output of a layer then feeds it back to the input to guess the output. Multiple RNNs that are each trained for a particular task are controlled by logic gates in the long short-term memory.

$$\widetilde{c}_t = \tanh(w_c[\widetilde{h}_{t-1}, x_t] + b_c)$$

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c}_t$$

$$h_t = o_t * \tanh(c^t)$$

$$c_t - \ cellstate(memory)attimestamp(t).$$
$$\widetilde{c}_t - represents\ candidate\ for\ cell\ state\ at\ timemstamp(t).$$

Result:

As in the above scenario LSTM performed well, we can say that models involving neural networks can perform better as compared to the normal regression models. The main idea behind the use of LSTM rather than RNN is to avoid the problem of vanishing gradient descent. RNN can face vanishing gradient descent where earlier layers fail to learn anything through back propagation. LSTM solved this issue by using different kind of activation function i.e, LSTM cell.

Having the capability of long term dependency LSTM worked really well.

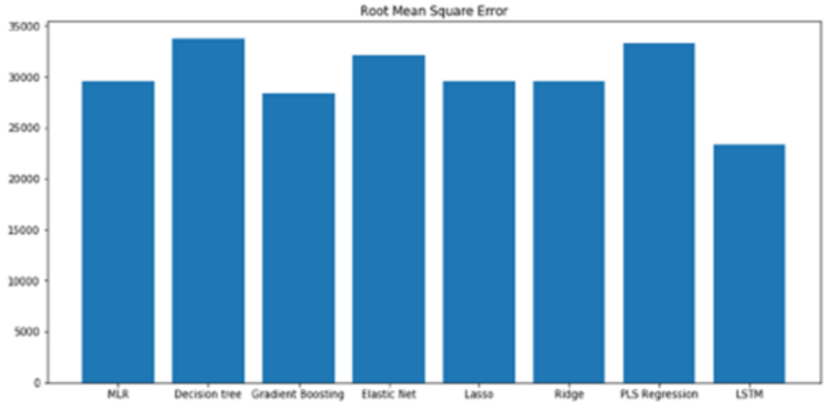


Fig. 7. Root Mean Square Error.

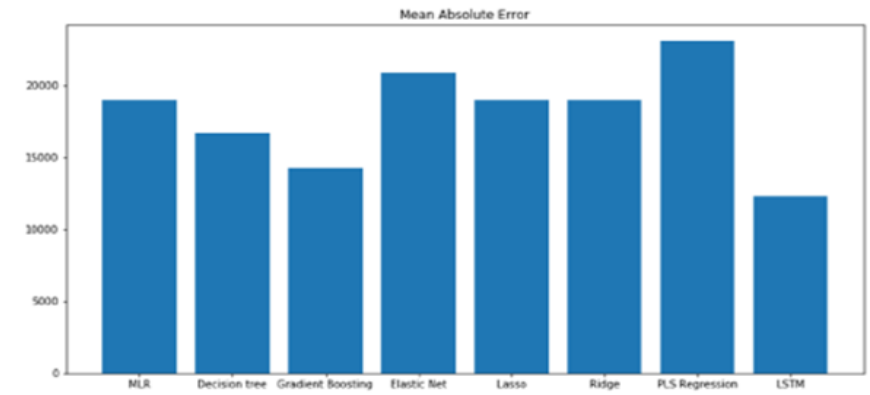


Fig. 6. Mean Absolute Error.

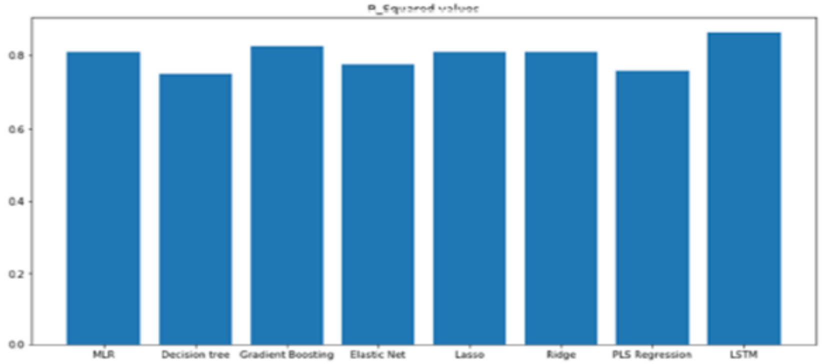


Fig. 5. R_Squared Values.

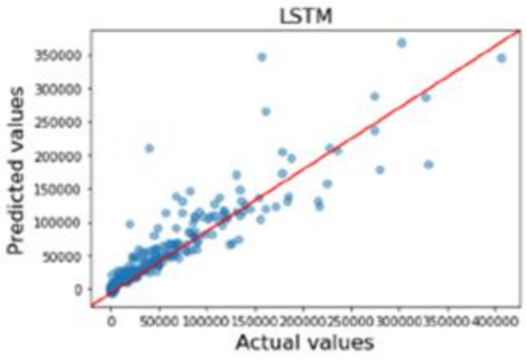


Fig. 15. LSTM Scatterplot.

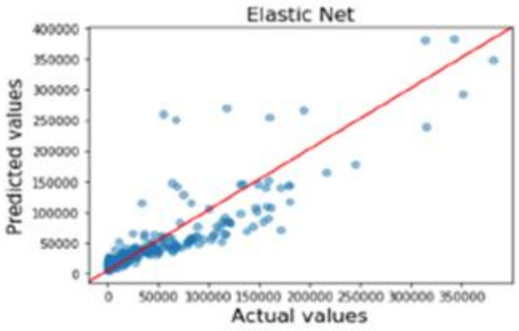


Fig. 13. Elastic-Net Regression Scatterplot.

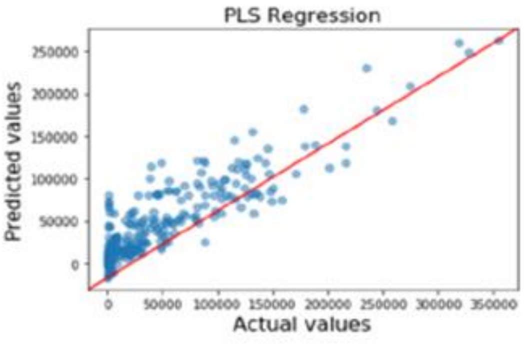


Fig. 11. Partial Least squares Regression.

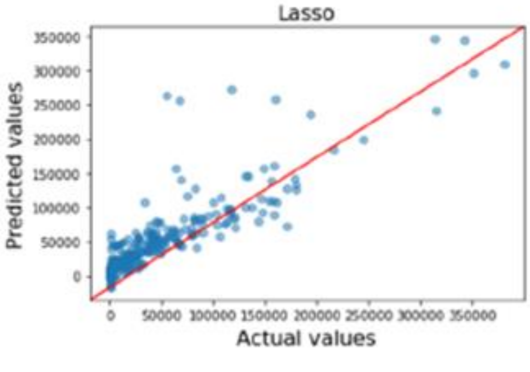


Fig. 12. Lasso Regression Scatterplot.

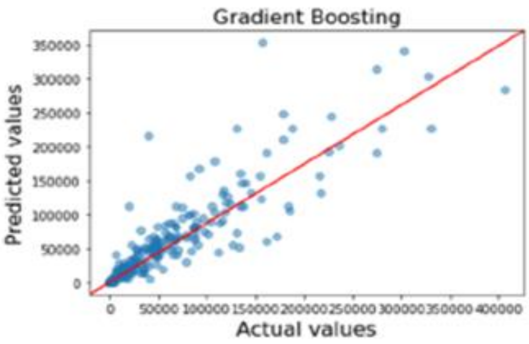


Fig. 14. Gradient Boosting Scatterplot.

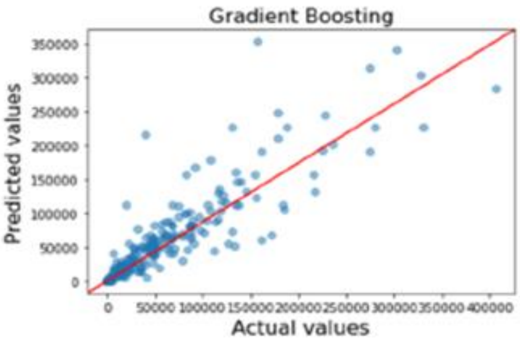


Fig. 14. Gradient Boosting Scatterplot.

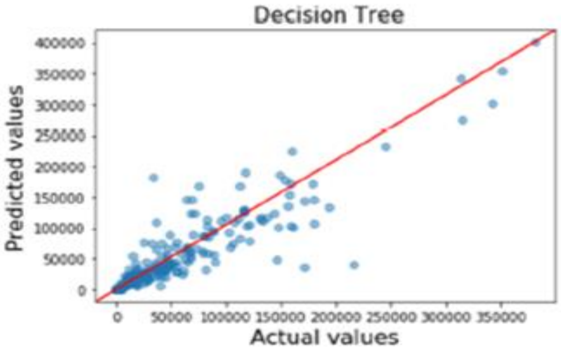


Fig. 9. Decision Tree Scatterplot.

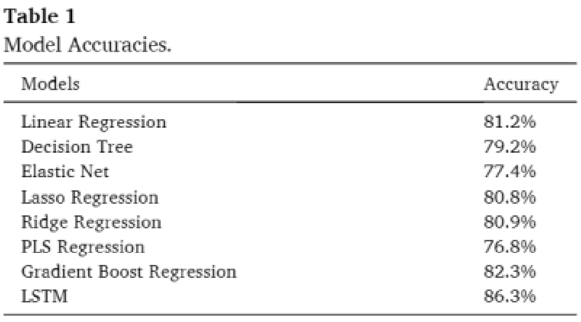


Table 1
Model Accuracies.

Models	Accuracy
Linear Regression	81.2%
Decision Tree	79.2%
Elastic Net	77.4%
Lasso Regression	80.8%
Ridge Regression	80.9%
PLS Regression	76.8%
Gradient Boost Regression	82.3%
LSTM	86.3%

Conclusion:

In the above case study, we understood how the author performed data preprocessing, found many relation between the parameters, performed plotting to visualize the data set properly, performed feature engineering to handle the lack of data for LSTM model training. Trained several models and compared their result with each other. This project was completed end-to-end by creating web application for the farmers. The only area where the LSTM model can limit is the presence of long term dependencies. One of the biggest limitations of LSTMs is their inability to handle temporal dependencies that are longer than a few steps. This was demonstrated in a paper published by Google Brain researchers in 2016. Another limitation of LSTMs is their limited context window size. This means that an LSTM can only consider a limited number of inputs when making predictions; anything outside of the context window is ignored completely. This can be problematic for tasks like machine translation, where it's important to consider the entire input sentence (not just the last few words) in order to produce an accurate translation.

We can overcome such problem by trying other models like Bidirectional LSTM and Neural Stack Machine and can find something more better than LSTM. In future we can make more stronger models which will give accurate results, not only that we can also merge image input to get more info. In that case we can also merge the power of CNN (Convolutional Neural Network) or can perform some kind of Ensemble Learning. As the web application was present only in local host, deployment of that web app along with some add features can also be kept as a future scope.

References:

S Iniyan, V Akhil Varma, Ch Teja Naidu (2022), Crop yield prediction using machine learning technique.