International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 05 | May - 2022 Impact Factor: 7.185 ISSN: 2582-3930

AGRICULTURAL PRODUCT PRICE AND CROP CULTIVATION PREDICTION BASED ON DATA SCIENCE TECHNIQUE

Yogeshwaran.R[1], Saai Sathish.S.A[2] , Irfan Ahmad.J[3],Mrs.Karthikayani[4]

[1]Department of Computer Science and Engineering, SRM Institute of Science & Technology, Vadapalani, Chennai, Tamilnadu, India
[2] Department of Computer Science and Engineering, SRM Institute of Science & Technology, Vadapalani, Chennai, Tamilnadu, India
[3] Department of Computer Science and Engineering, SRM Institute of Science & Technology, Vadapalani, Chennai, Tamilnadu, India
[4] Department of Computer Science and Engineering, SRM Institute of Science & Technology, Vadapalani, Chennai, Tamilnadu, India

---

ABSTRACT - Agriculture is primarily responsible for increasing the nation's economic contribution across the world. However, due to a lack of implementation of ecosystem control technology, the majority of agricultural lands remain underdeveloped. Crop output is not improving as a result of these issues, which has an impact on the agricultural sector. As a result, agricultural production is increasing as a result of plant yield prediction. To avoid this issue, agricultural industries must use machine learning algorithms to forecast crop yield from a given dataset. The use of supervised machine learning techniques (SMLT) to analyse datasets in order to capture various information such as variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments, and so on. A review among machine learning algorithms was conducted to see which one was the most competent in predicting the best crop. The findings reveal that the suggested machine learning algorithm approach has the greatest accuracy when comparing entropy calculation, precision, recall, F1 Score, sensitivity, specificity, and entropy.

***Key Words*:** Algorithms, Decision making, Agriculture, Data Science Techniques.

1.INTRODUCTION

Machine learning models and advancements are now available to agricultural farmers. The use of AI and machine learning in the food tech sector is beneficial. Farmers Business Network may employ machine learning and analytic techniques to produce price data outcomes. Crops are being managed and monitored by robots. Sensors aid in the collection of agricultural data. SMLT may also be used to estimate agricultural production in a specific location based on historical data and price predictions. According to studies, if AI and machine learning are employed in agriculture, the industry would increase in the next years, improving the economy of the country.

Crop cultivation forecast is an important aspect of agriculture, and it is dependent on a variety of elements including soil, weather conditions such as rainfall and temperature, and the amount of fertiliser used, notably nitrogen and phosphorus. These elements, on the other hand, differ from one location to the next, making it impossible for farmers to grow comparable crops in all of them. Agriculture relies on the ability to predict a viable crop for production.

Of great concern to many farmers is the uncertainty of crop prices. Farmers are unable to design a fixed output schedule due to price fluctuations. This issue is particularly prevalent in plants with short shelves, such as tomatoes. Companies are surveying the earth and monitoring plant life in real time using satellite imagery and weather data. Companies can detect pests

and diseases, predict tomato production and yield, and anticipate prices using technologies such as big data, AI, and machine learning. They may advise farmers and governments on future pricing strategies, demand levels, the best crop to plant for good yields, the use of pesticides, and more.

2. RELATED WORKS

The RFE and non-RFE (NRFE) approaches were compared by Gregorutti et al. . The approach was validated using the available data from the Uni of California tech repository, and the importance metric was utilized as a criteria ranking for FS. The RFE was shown to be more efficient than the NRFE based on the findings. Hall and Holmes examined many FS approaches and evaluated them using benchmark data sets. The wrapper strategy is the best for FS, according to the results. Liue investigated the pre recorded classification and clustering algorithms. They utilised real-world applications to showcase the FS approaches in their research. Granito professor, Recursice feature elimination and MRFE were compared. The performance of agro-industrial goods was assessed using proton transfer reaction-mass spectrometry (PTR-MS) data. According to their findings, the RF-RFE outperforms the SVM-RFE. To assess various FS approaches, Azofra and Bentez employed different set of data the University of California, Orange (Org), and Silicon Graphics (SGI). The wrapper technique is the best for picking characteristics, according to an experimental investigation. Altman prof. Has suggested a better Recursive feature model for FS using the existing pipeline hyper tuning measure. When the

measurement was used to ranking measure and Gin importance were evaluated, it was discovered that the model- beat the Gin Recursive model model significantly. The Boruta FS approach was introduced by Kurs and Rudniki , and the Boruta programme offered a user-friendly interface for their algorithm, which was tested using the Madalon data set. Ruß and Kruse suggested an unique FS approach for wheat yield prediction that included a comparison of two regression models: Supporting vectors regressions and non-binary tree model i.e is regression. In terms of variable selection, Darset evaluated the Recursive feature and elimination using recursive feature, concluding that completed work would never scale to the given prominent data. The RFE method was utilised by Hsieh et al. to identify critical factors that influence rice blast disease (RBD).

3. METHODOLOGY

Upload data, ensure cleanliness, and cut and clean your data for the analysis in this part of the report. Make sure users clearly record their procedures and protect your cleaning choices.

For example, using this program, you can import any algorithm and class to separate train tests from sklearn and numpy modules. Then we wrap the method of uploading data () to a dynamic database. Next we use the train test split method to split the database into training and test data. The A prefix in the variable represents the number of elements, while the B-value indicates the target values. This method randomly separates the database into training and testing data at 64:36. Then we combine any algorithm. In the next line, we combine our training data in this way so that the computer is trained to use this data and lastly the part of the training is over.

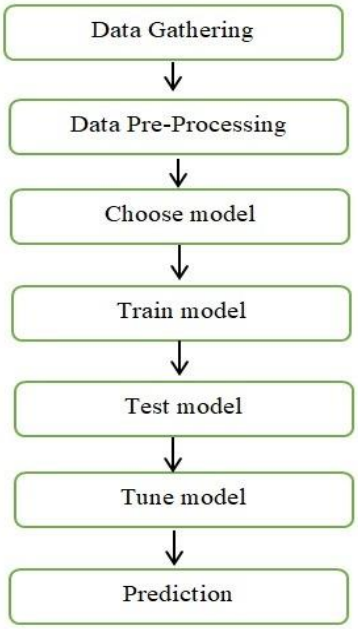


Fig -1: Methodology

Data Pre-processing:

Machine learning verification methods are used to determine the error rate of the Machine Learning model, which can be described as near to the actual data error. If the dataset is large enough to represent the whole thing, you will not need such strategies to verify.Eventhough, in real situations, working with data samples will not be a real representation of a given database population. To find the null values, multiply the value and describe the data type regardless of the float variable or the total number. Sample data used to provide an unbiased assessment of model equity in the training database while adjusting the hyper parameters of the model.

A verification set is defined here to test a particular model, but this is a standard test. As machine learning engineers they use this data to fine-tune the hyper parameters of a model. Data collection, data analysis, and process for data content, quality, and composition can add you to your to-do list. During the data identification process, it helps to understand your data and its properties; this information will help you decide which algorithm you can use to build your model.

Variable	Description
Crop	Crop name
State Name	Indian state name
Cost of Cultivation (`/Hectare) A2+FL	Cultivation amount for A2+FL Scheme
Cost of Cultivation (`/Hectare) C2	Cultivation amount for C2 Scheme
Cost of Production (`/Quintal) C2	Production amount for A2+FL Scheme
Yield (Quintal/ Hectare)	Yield of crop
Crop year	Crop year list
District Name	District name for each state
Area	Total area of each place
Rainfall	Water availability of each crop
Average humidity	directly influences the water relations of plant and indirectly affects leaf growth
Mean Temperature	Climate of each crop

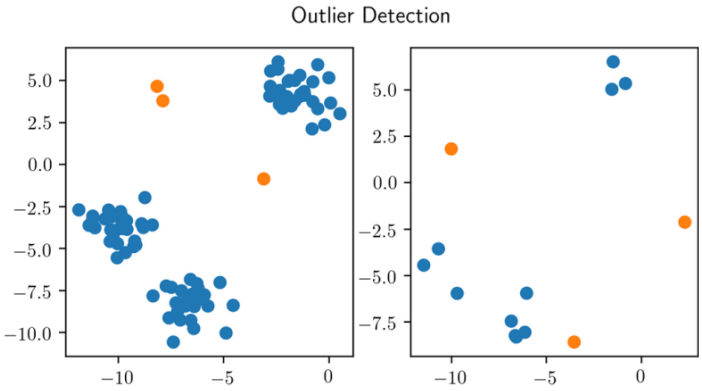
4. ALGORITHMS AND PROPOSED WORK

ML and mathematics, segregation is a supervised machine learning method by how a Ai is used in the computer to plan from a given data input and uses the trained data to find the value of existing target column. This data set may be bi-class or it may be multi-category as well. Some examples of classification problems are: Recognizing speech, handwritten notes attention, bio metric detection, document fragment etc. In Supervised Reading, algorithms read data with labeled data. After the data is trained, we need to specify a algorithm for the model that determines which label should be assigned to the new observation on basis on the probability and associates the similar observations with the new untrained data.

The goal here is to make a learning model of a predictive harvesting crop price and crop prediction using the machine learning algorithms that can take the models of the monitored controlled machine by predicting the results in the most accurate way by comparing the controlled algorithms. The purpose of this study was to use machine learning to analyze plant database records in the agricultural industry. It is very difficult for farmers to expect agricultural yields. We strive to reduce the risk factor in crop selection. The demo database is now integrated into the machine learning model, and the model is trained using this data set.

Analysis of demonstration test data using data visualisation:

Viewing data is an important skill in machine learning and machine learning. The statistics really focus on quantitative definitions and data measurement. Data viewing provides an important strategic tool for gaining quality understanding. This can be useful when exploring and knowing a set of data and can be helpful in identifying patterns, corrupted data, external objects, and much more. With a little background information, data perception can be used to identify and demonstrate important relationships in strategies and charts that are more visible and participatory than related or value measures.



Hyperparameter Tuning:

Once the data is preprocessed,we can pass the clean data for the feature selection also known as hyper paramenter tuning from the list of available parameters we define in the pipeline function of python.Once the pipeline model is triggered the gridsearch selects the best paramters to use for the model.

Hyperparameters: Variations of linear regression are common as the deviation of data in the dataset. The deciduous tree has a mutli level in-depth also a small amount of leaf spot recognition as hyperparameters.

Correct Hyperparameters: Hyperparameters control the upper and lower extremities of the model. Accurate hyperparameters often differ from different databases. To get the best hyperparameters, the following steps are followed:

1. For each hyperparameter setting the model is tested

2. Hyperparameters provide the best model selected.

Hyperparameter Search: Grid search selects the hyperparameter values grid and checks them all. Guess function is required to initiate min and max values for every parameter. The randomized searches randomly measure a example of points on a gridCV. It works much better than grid search. Smart hyperparameter tuning selects a few hyperparameter settings, checks verification matrices, corrects hyperparameters, and re-verifies verification matrices.

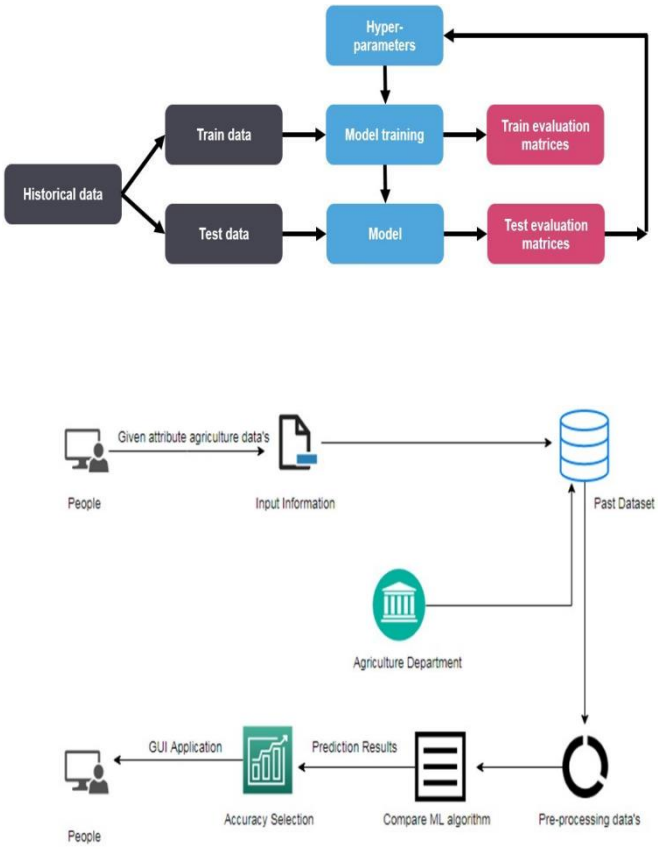


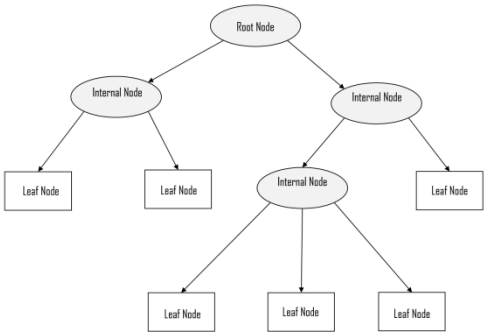
Fig -2: System Architecture

Decision Tree Algorithm:

This algo is one of the most industry used sub-tree based algorithm. It falls under the class of machine learning algorithms(SML). It handles both regression output variables and phase. Decision tree ideas:

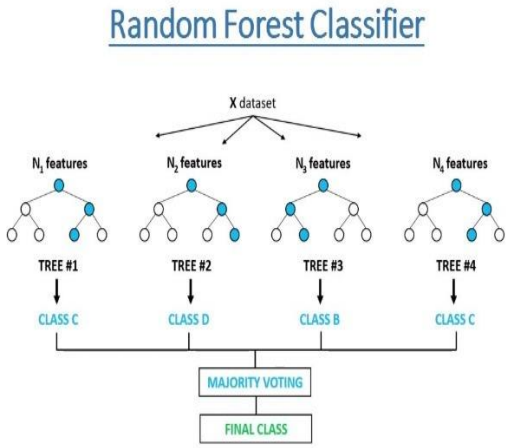
- In the beginning, we look at all the training set as root.
- Preferences are considered as part of gaining knowledge, attributes are considered continuous.
- On the basis of qualitative values and records are still distributed repeatedly.
- We use mathematical methods to order attributes such as root or internal node.

The decision tree constructs models that divide or recede in the form of a structural algorithm. It splits the data set into several small sets while at the same time the associated decision tree is further developed. The decision area has more than one branches and the leaf area depicts a split or probability. The highest dominant value node in a tree with the accuracy prediction called the base-root node. This type of algorithm can handle both regressional and classification data values.



Random Forest Classifier:

Random forests or random decision-making are an integrated learning method for classification, retreating and other activities, which works by creating a number of deciduous trees during training and class extraction which is a classroom mode (planning) or predictable meaning (retreat) of individual trees. Random decision-making forests correct the practice of deciduous trees to be most suitable for their training set. Random Forest is a type of machine-readable algorithm based on group readings. Shared learning is a form of learning where the various type of algos or the pre-exisiting algo are used several time for creation of the most probable guessing model. The RF algorithm includes algos of the same type multiple deciding tree, which leads to a vast variety of sub trees, so the word. It can be used for both retransmission or partition operations.

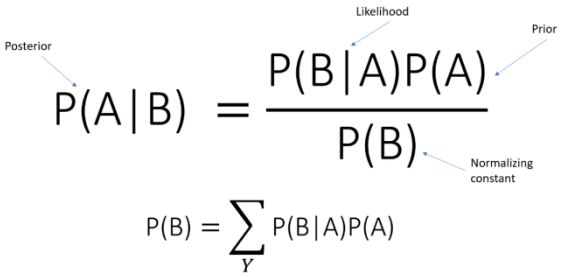


Naive Bayes algorithm:

The Naive Bayes algorithm is an accurate method that uses the opportunities of each attribute belonging to each category to make a prediction. It is a supervised learning method that you can come up with if you want to model a modeling problem that is predictable.

It facilitates the calculation of opportunities by assuming that the probability of each attribute belonging to a particular category is independent of all other attributes. This is a powerful concept but it results in a faster and more efficient way.

The probability value of a category given an adjective value is called a conditional probability. By multiplying the conditional opportunities together for each attribute of a particular category, we have the potential for example data for that category. To have a prediction and then use the probable of the model for each class and then select the class value with the high value found.



$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(B) = \sum_{Y} P(B|A)P(A)$$

Metrics used for determining the model efficiency:

True Positive Rate (TPR) = TP / (TP + FN)

Good false rating (FPR) = FP / (FP + TN)

Accuracy: Part of the total number of correct predictions if not at all how often the model predicts automatic and non-error people.

Accuracy rate:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Accuracy rate is the accurate performance measure of the model and is a prediction of accurately predicted observations in all comments. One might think that, the high accuracy then the model will be best and vice versa. The point is valid in some cases where accuracy is a good but only if you have a database that is symmetry in nature where the false value and negative negative values are similar.

Accuracy: Part of a good prediction that is really true.

Accuracy = TP / (TP + FP)

The accuracy of the predicted positive predictive accuracy is higher than the predictable positive predictive value. The question is if this metric answer is for all passengers listed as survivors, how many actually survived? High accuracy is related to low level of inaccuracy. We found an excellent 0.788 accuracy.

Standard formula:

F 1-Score :

F 1 value = Two * (Remember * Accuracy) / (Remember + Accuracy)

After Comparing Algorithm with prediction, we conclude the best algorithm with maximum accuracy. The Algorithms used are Random forest Classifier, Decision Tree, Suppirt Vector Machine, and Naive Bayes. It is critical to reliably evaluate the performance of various distinct machine learning algorithms, and it will be discovered how develop a test harness in Python using scikit-learn to compare multiple different machine learning algorithms.This test harness may be used as a framework for your own machine learning tasks, and you can add additional and alternative algorithms to compare. Each model will have an own set of performance

characteristics. You may evaluate how accurate each model is on unseen data using resampling approaches such as cross validation.

4. RESULTS AND DISCUSSION

After doing this study, we found that Random Forest Classifier gives maximum accuracy and it interfaced with GUI. Thus, we can predict the crop using such machine learning Algorithms.

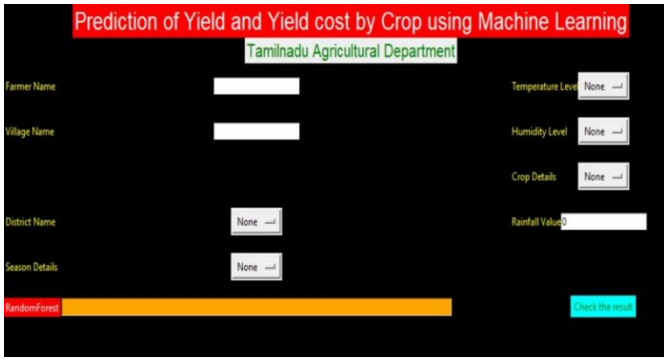


Fig -4: GUI for price prediction

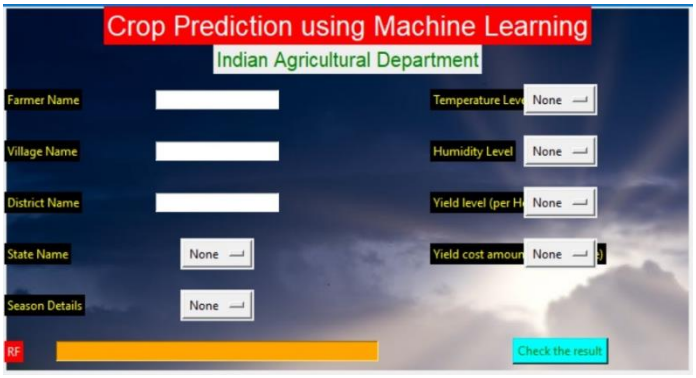


Fig -5: GUI for Crop prediction

5. CONCLUSION AND FUTURE SCOPE

The testing method began with data processing, followed by null values found or zeros instead of a required feature analysis, data visualisation, and lastly model creation and assessment. Finally, we use a machine learning method to estimate the harvest, with varying outcomes. This leads to some of the following crop forecast findings. Because this system covers the most sorts of crops, farmers may learn about crops that have never been farmed before. It also includes all conceivable crops, which aids farmers in deciding which crop to plant. Furthermore, this system considers previous data production, which will assist the farmer in gaining insight into the demand and pricing of various crops in the market. The remaining SMLT algorithms will be involved in determining the optimum accuracy with which to anticipate agricultural output and cost. The agricultural department wishes to automate the detection of yield crops during the eligibility procedures that can be done in real time To simplify this procedure by displaying the forecast result in online framework Django or Flask in Python or create a desktop Tkinter application.Though the model has high accuracy score now because of a limited use of dataset maybe in the future due to the drift in incoming new features the model can be outperformed or new feature engineering techniques can be incorporated.

REFERENCES

[1] G. Mariamal , A. Surliandi , S. P. Raj , and E. Poonkothai Determing the features of Land Suitability for Crop Cultivation using the features of soil and Environmental Characteristics Using MRFE and comparing with Various Classifiers .

[2] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," IEEE Trans. Comput. Social Syst., vol. 8, no. 1, pp. 214–226, Feb. 2021.

[3] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," Statist. Comput., vol. 27, no. 3, pp. 659–678, May 2017.

[4] D. H. Zala and M. B. Chaudhri, "Review on use of BAGGING technique in agriculture crop yield prediction," Int. J. Sci. Res. Develop., vol. 6, no. 8, pp. 675–677, 2018.

[5] M. Gopal P S and B. R, "Selection of important features for optimizing crop yield prediction," Int. J. Agricult. Environ. Inf. Syst., vol. 10, no. 3, pp. 54–71, Jul. 2019.

[6] S. Nembrini, I. R. König, and M. N. Wright, "The revival of the gini importance?" Bioinformatics, vol. 34, no. 21, pp. 3711–3718, Nov. 2018.

[7] M. A. Al Maruf and S. Shatabda, "IRSpot-SF: Prediction of recombination hotspots by incorporating sequence based features into Chou's pseudo components," Genomics, vol. 111, no. 4, pp. 966–972, Jul. 2019

[8] M. Lango and J. Stefanowski, "Multi-class and feature selection extensions of roughly balanced bagging for imbalanced data," J. Intell. Inf. Syst., vol. 50, no. 1, pp. 97–127, 2018.

[9] R. Rajasheker Pullanagari, G. Kereszturi, and I. Yule, "Integrating airborne hyperspectral, topographic, and soil data for estimating pasture quality using recursive feature elimination with random forest regression," Remote Sens., vol. 10, no. 7, pp. 1117–1130, 2018.

[10] P. S. Maya Gopal and R. Bhargavi, "Feature selection for yield prediction in boruta algorithm," Int. J. Pure Appl. Math., vol. 118, no. 22, pp. 139–144, 2018.