Intellomod – Crop Yield Prediction

Maharaj Mahaadev, Koragatla Sai Rithvik

***Abstract*— The project is about the prediction of yield of crops and thus in turn predicting the price of crops with the help of various models that are available. Doing so would help farmers greatly increase their profits and plan on the future on what crop they should cultivate next. Several systems are in place today, and we are improving the prediction while adding more features to the application. The main problem to solve is the uncertainty regarding the yield of crops grown. The solution would be to successfully predict the output that is the yield of crops to be acquired in the future. The benefit in doing so, would be saving time, resources etc. and mainly the loss or to greatly increase the profit to be gained from selling the crops. The models we used were tested and replaced to get the prediction with the best accuracy as possible as of now.**

***Index Terms*—Artificial Intelligence, Decision Trees, Linear Regression, Machine Learning**

I. Introduction

Farming is an activity which is important for a society. But in several cases, it has been observed that many farmers go through hard times because there is no way of knowing the price of the crops, they produced for selling. In fact, there is no certain way that is mathematically used to predict the quality of the crops. Since that's the case the price of crops varies widely across various buyers, and that is not good because farmers could suffer from losses.

Farming is a skill [1], it is better to know beforehand what crops to grow in a region depending on several factors like the humidity, soil type, weather etc. If such calculations are made and farming is done, then the yield could be greatly increased. Otherwise, all these factors would hinder productions and as a consequence either profit would be lost or there is also a chance of going in loss. Many farmers had to go through difficult scenarios and unfortunate events and many others went bankrupt and has been in huge debts.

If there was a way to solve all that then it would be really great. Fortunately, now such solutions do exist which are the prediction systems which we can use to predict which crop to grow as well as the yield which we would receive from that. But all these systems already do exist in place. In this project, what's different is the prediction of the crop price estimations. And working on the previous projects we have improved the efficiency of the predictions and made it much effective.

To solve such problems, it would be good to have a system which predicts the price of the crops from various inputs and tells us the estimated price that the crop would get. Combining various factors like region, availability, past year trends, quality of soil, fertilizers used, weather etc. By feeding all these data into a model we can get the output which would be the estimated price for the crop. This will provide a way for farmers to understand how much their crop is actually worth. Even though things are always changing (like weather and climate), that set of data is predicted by the weather prediction systems.

There are several factors in nature which can't be predicted, but we can try to reduce the randomness of the situation by predicting it much better. And that's the exact idea behind this project. Even though it can never be 100%, it should be easier if the system can predict at least 85% of the cases/ days or more than that. So, if we combine all these data then it would be possible to get the price of the crop.

Food is important for any country, by increasing profits in the food industry the nation can greatly prosper. By predicting what price, a crop yields can be a great way to increase the national income. The price of the crops which are exported can be predicted beforehand, saving time and money. It would be easier to export good quality crops for great money. All these can be done with the help of predicting the prices of crops from simple inputs.

There are lots of crop prediction systems available in the market but what makes this one different is because we have improved the efficiency of algorithms by trying out several ones and finding the ones which provide with the most accurate price. We have also removed inputs which provide minimum value for prediction and added more effective inputs which actually matter in prediction.

In doing so, we added weather as well as temperature of the day. Parameters such as the day in which the crop was planted etc. The removed features are the quality index of the soil, the parent branch to which the crops belong etc.

Compared to other learning models, by adding and removing several things we have made the prediction to be the best. In the future we can improve on this again. The project has also been expanded to accommodate prediction of prices for the crops

Submitted on 17/04/2022 as project for the course of Artificial Intelligence and Machine Learning, Bennett University.

Maharaj Mahaadev is with Bennett University, Surabhi Road, Kochi, India

Koragatla Sai Rithvik is with Bennett University. Andhra Pradesh, India

which are to be exported.

The training model used mainly is decision tree regression. With the help of the model, the output came down as the most accurate from the other ones we tested. Parts of this model can be improved, and the dataset and features used could also be changed in the future.

The dataset contains several attributes. A glimpse of those attributes is crop names/ type, weather, temperature, date, location, date of crop sprouting, quality of seeds, fertilizers used, type of soil. All of these will be explained in detail in the future reports. The output is the price we are trying to predict for the crop. For this we mainly compare it with the price of the crop in the past year. Past year insights and how much that crop did in the market will provide us with useful knowledge for predictions.

.

II. Related Work

Niketa et al 2016 [2] through research showed us that the quality of crops depends on the seasonal climate. In India, the climate varies widely across the regions, several places in India have a behavior of not having consistent weather. Farmers face problems due to this. To solve that problem, they used some machine learning models to predict crops to grow depending upon the climate. They used past year data to predict the future one. SMO classifier was used for this. Main features that were considered are minimum temperature, maximum temperature, average temperature, and previous year's crop information and yield information. With the help of SMO, they classified the data into two groups, the one with high yield and the other with low yield. However, the obtained result from the project had low accuracy compared to some other models.

Eswari et al 2018 [3] found out that the yield of the crop depends on the perception, average, minimum and maximum temperature. They also decided to add one more feature for better prediction that is crop evapotranspiration. The crop evapotranspiration will contain the weather and growth of plant. The feature helped in better prediction for the yield of crops. The algorithm used was Bayesian networks and they classified it into two groups as false and true and compared it with the actual output of previous years and checked the accuracy. Their final project gave better results compared to the previous model. So, we can conclude that the Bayesian model is better that SMO in predicting the yield of crops.

Shruti Mishra et al 2018 [4] studied and found out that the previous year data of crops is great for predicting the future of that crop. With the help of these data, farmers can decide which crops to grow. They used dataset with attributes like area of land, and production, crop used and the season, they used these features in the prediction. The accuracy of the data was compared to other methods. Based on the data they used, they finally got the IBK model as the one with the most accurate prediction.

Chlingaryana et al 2017 [5] studied the relationship between the crop and the nitrogen levels present in the soil. Remote sensing was used for finding out these data. Such equipment can directly help the farmers in producing great yield. Remote sensing data across various places is combined for learning purposes. Nitrogen present in the soil is what makes it fertile and thus improves the crop yield. The algorithms were fed with various data including the nitrogen present in soil, type of soil, previous year data etc. for the accurate prediction of results. The soil, crop and some other decisions can be made based on the data provided by the precision agriculture. It takes in various inputs and gives a best output. Back propagation neural networks are used for the excellent long-term memory and best results of the future data.

Dakshayini Patil at all 2017 [6] studied and proved that the rice crop is significant to the economy. Rice crop is the most sustainable and the highest produced crop in India accounting to about 40% of India’s total production. Machine learning is again used for the prediction of the price of rice crops. Yield is directly based on the best conditions for the crop to grow. A better strategy to know when and where to plant which crop can greatly increase the yield which in turn increases the profits. Various data can be used for the successful prediction of this yield of the rice crop. Here, they used data of several regions across Maharashtra. Weighted contributions have been used. This is also known as artificial neurons. Which is connected to the neural systems. This means that the failed or inaccurate results are send backwards which are then replicated in a reverse manner. The system is built around information from each of these individual neurons forming a neuron layer. There would also be concealed neurons in the middle. This neuron layer of system is widely popular and used in a lot of places since it is highly beneficial in several Artificial Intelligence fields.

T. Mhudchuay et.al. [7] Concentrated on downpour took care of rice where the fundamental activities are when to begin development and when to collect. The objective is to locate the ideal development and collect period to such an extent that ranchers' salary is amplified. This paper speaks to a use of a Deep Q-learning in the rice crop development practice, where the ideal activities are resolved.

Shivi Sharma et.al., [8] proposed a technique utilized, in that dirt and condition highlights for example normal temperature, normal stickiness, all out precipitation and creation yield are utilized in anticipating two classes in particular: great yield and awful yield.

Suhas S Athani et.al. [9] Presents the data relating to the harm of harvests as of late because of the development of weeds. Weeds are one of the significant hazards to the genuine home and mankind. Right now, thought, Support Vector Machine (SVM) Classifier is used to make out whether plant is harvest or weed. The maize crops are consistently observed by catching pictures utilizing camera. So as to group a plant as a yield or weed, different highlights are removed which among them are shape, surface, shading.

Ranjini B Guruprasad et.al.,[10] introduced a contextual analysis of climate and soil information-based yield estimation demonstrating for paddy crop at various spatial goals (SR) levels, to be specific, at the area and taluk levels in India. We give a point by point investigation of precision of the yield estimation models across changed arrangements of highlights and diverse AI systems.

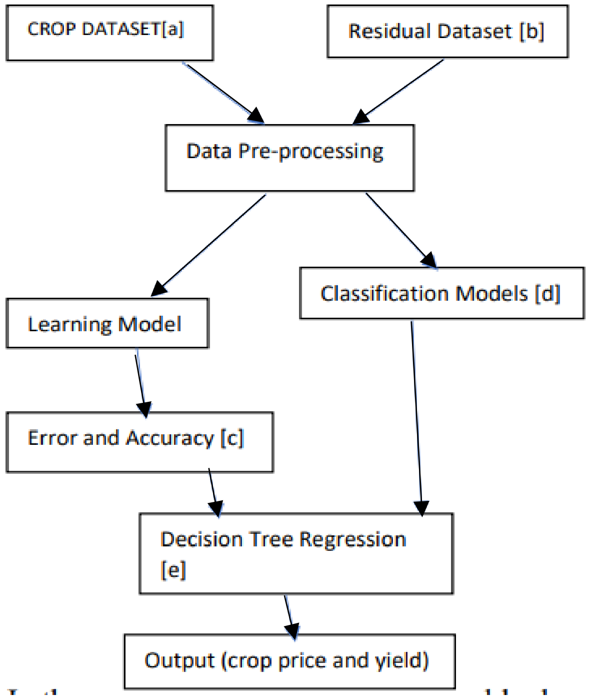
Neural system works in a feed forward type of network, where each neuron sends messages forward. In such system neurons are linked together by a forward mechanism. Each layer can have several types of links and systems in place but mostly it's connected forward, meaning there is no backward transmission. When utilizing these layers and networks, the system must calculate from unforeseen events and inputs. The foreseen yields are compared against the real yields for given information. Utilizing the unforeseen yields can greatly change the way one can predict the output. In such systems it is better to keep the input at a certain amount as calculated to give out the best prediction results and accuracy as much as that is possible from such a situation. The end result of such good neural networks would create an excellent predicting mechanism and results when accurate number of inputs is used on these layers of neurons in an orderly and well-mannered style. Doing the opposite would however lead to the prediction underperforming and such cases should be avoided by always considering the input and making sure it's from any errors before feeding it to the system. Thus, we can conclude that in fact neural networks do predict the yield of crops with excellent accuracy.

III. Flow Diagram

In the block diagram above, there are two datasets used. The first dataset (labelled [a]) is used for the crop data which contains the crop type, soil, location, data of crop sprouting, quality of seeds etc. The second dataset (labelled [b]) is used for the other data, which is required like the weather, humidity, rainfall, type of soil, temperature fertilizers used etc.

The first dataset is fed to various models and the error and accuracy of the model on the dataset was tested. Meanwhile, the second dataset was tested on classification models and various and errors and accuracy were tested. After this the final model was decided to be decision tree regression as the best fit model.

Therefore, this model was used for the final implementation of the project. This gave the best output which was the yield and from yield we can calculate the price of the crop. If the model gave poor performance, then it would have been intercepted at the point (labelled [c] and [d]) and the dataset would have again been tested to get better results. If better results were obtained, then at the point (labelled [e]) this model would have been used for the final implementation. And thus, from that model we would attain the best result.



IV. Proposed Methodology

A. Proposed Algorithm

The algorithm used in this project is mainly random forest. It also has decision trees and linear regression. The final algorithm used was random forest because it gave the best accuracy. Before that the data was tested on both linear regression and decision trees, but it did not produce much accuracy. As a result, new models were tried, and decision trees was finally chosen for implementation. Decision trees and linear regression are from the 'sklearn' library of python. Both are regression models. The decision tree is a type of classifier which works by the use of decision trees and creates a classification model. The decision tree classifier creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute.

B. Techniques

The linear regression model was trained using the datasets explained in the next section. Then it was used to predict the results using the test dataset. The coefficient determinant score and mean squared error of the linear regression model was found out. The results were not ideal, so it was decided to use a different model.

The decision tree regression model was used next. After training it with the dataset given, the model was tested. The testing was again done using mean squared error and coefficient of determination. The model gave good results and predictions was nearly accurate. Hence this model was decided as the final one.

V. Experimental Setup

A. *Dataset Description*

The dataset used contains some features which are useful for the crop yield predictions. The dataset contains previous year data of crops, 'year' is one column in the dataset which defines which year data is present. It also contains the location which is the region or the local area where the crop was harvested. Having information about the region is important for finding out various factors. The dataset also contains a column containing the precipitation amount which was present in the region.

The fourth, fifth and sixth columns contain the minimum, mean and maximum temperatures in degree Celsius which was recorded in the region. The seventh column is about the crop evapotranspiration in millimeters. Evapotranspiration is defined as the sum of all forms of evaporation plus transpiration. The next column is about the area in hectare which is basically the amount of area which was used for planting the crops. Then next column is the production of the crop which was obtained in unit of Tones. Finally, the last column is the target feature which is the Yield in the unit of Tones per Hectare.

B. *Data Preprocessing*

For data pre-processing, the first step we did was to check for the presence of any missing values. Some columns like the temperature and the evapotranspiration had some missing values. To handle the missing values present, mean imputation was performed. After this, the data was normalized using some of the normalization techniques. Data splitting is commonly used in machine learning to split data into a train, test, or validation set. Each algorithm divided the data into two subset, training/validation. The training set was used to fit the model and validation for the evaluation. Data splitting is the act of partitioning available data into two portions, usually for cross-validatory purposes. One portion of the data is used to develop a predictive model and the other to evaluate the model's performance. The dataset should be divided into two parts. One will be the training set and other will be the testing set.

C. *Feature Extraction*

Finally, some of the feature extraction/ selection processes were done. The categorical values were transformed into continuous values. After this again, some of the missing data checks and normalization were done.

After carrying out all these steps, the initial dataset was transformed into a preprocessed dataset and this dataset can now be used for training and testing purposes. The dataset was split into four categories as x_train, y_train, x_test and y_test with the help of sklearn library of python.

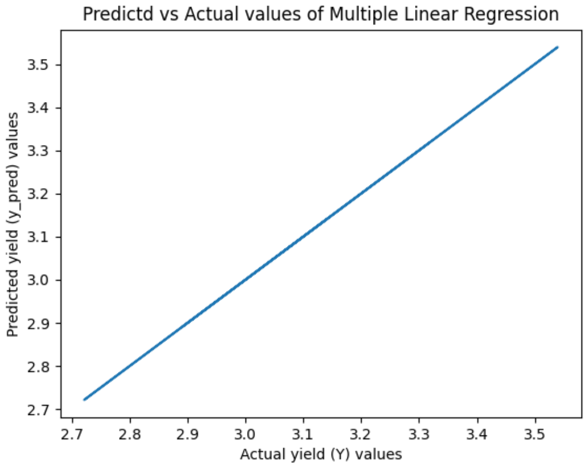
VI. Result Analysis

The results we got from different models used are as follows. For linear regression it gave a Coefficient determination score of 0.1 which is the least among all other models used and this score isn't ideal. The coefficient score of the decision tree regression model came to around 0.95 which is good but still there was some room for improvement. For the final model, that is the random forest the score came to 0.98 which is great. Because of this random forest was used as the final model.

Model	Mean Squared Error	Coefficient Determination score
Multiple Linear Regression	0.95	0.1
Decision Tree Regression	0.13	0.95
Random Forest	0.12	0.98

Fig. 1. Comparing accuracy of different models used

Fig. 2. Graph of Accuracy of Multiple Linear Regression



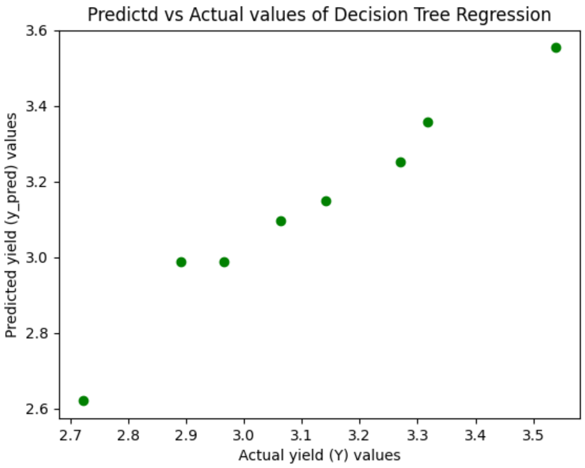


Fig. 3. Scatter plot of accuracy of Decision Tree Regression

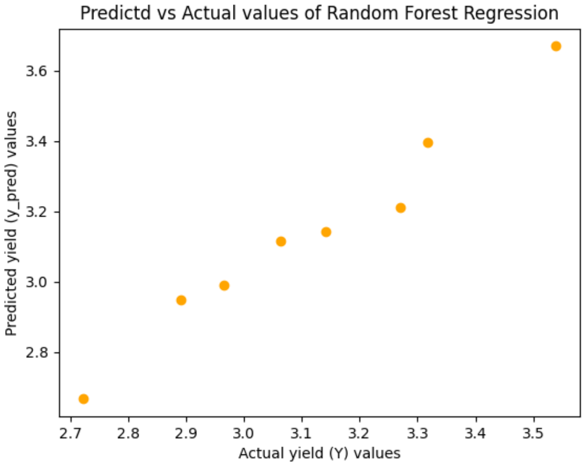


Fig. 4. Scatter plot of accuracy of Random Forest

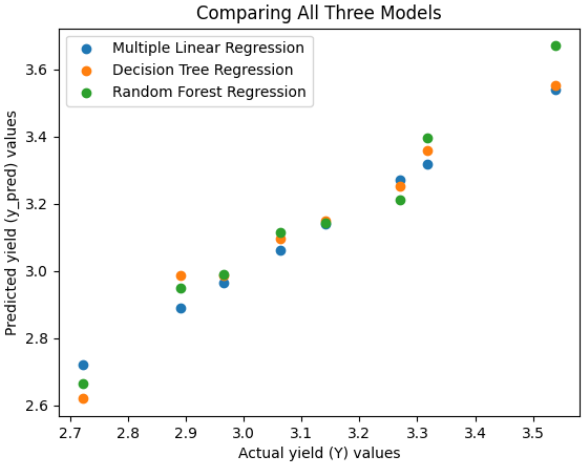


Fig. 5. Graph comparing the accuracy of all the three model used

VII. Conclusion

The project aimed at predicting the yield of crops with the help of various other features. It is crucial and helpful to know the amount of yield that can be received trough the year. It is helpful to the farmers and would also help the farmers in deciding what type of crop to grow during a certain frame of time. Different algorithms were used to predict the yield of crops. Certain factors that were identified as important were used in the dataset for training of these different models. The regression models were used because it is a classification type of problem. After identifying the important features, the data set was chosen. The dataset contains all the features as explained earlier. Later on, the data was preprocessed like feature extraction, normalization etc. were performed on the dataset to create a good, preprocessed dataset. On this dataset some of the regression models were tested out and the best three regression models are used in the final model. The models are linear regression, decision trees and random forests.

Out of all the models used, the random forest gave the best accuracy. For making predictions this model can be used as it is dependable which has been proven by its good accuracy on the training samples. With that the final model was decided and the predictions were done. After making such a project based on predictions, the shortcomings of the past related works were fixed as explained earlier and a new better prediction system had been implemented for the prediction of the yield of the crops.

VIII. Future Scope

Several new features could be built upon this project in the future. Some of those features are as explained below. The most important thing in this project is the prediction system. A prediction system with poor accuracy is nothing but useless. Even though in this project, the past drawbacks have been overcome and better accuracy was obtained with the help of better models, there is a lot of room for further improvement. Using better models or even changing the dataset by preprocessing it further by adding new features or dropping existing features which have no impact on the final predictions are some of the ways by which the accuracy of the project could be improved upon. Better models could also play a huge role in predicting with better accuracy. This all could be done in the future.

An agro-based country depends on agriculture for its economic growth. When a population of the country increases dependency on agriculture also increases and subsequent economic growth of the country is affected. In this situation, the crop yield rate plays a significant role in the economic growth of the country. So, there is a need to increase crop yield rate. Some biological approaches (e.g., seed quality of the crop, crop hybridization, strong pesticides) and some chemical approaches (e.g., use of fertilizer, urea, potash) are carried out to solve this issue. In addition to these approaches, a crop sequencing technique is required to improve the net yield rate of the crop over the season. One of existing system we identified is Crop Selection Method (CSM) to achieve a net yield rate of crops over the season. Depending on these factors some changes can be made to this existing project.

One other work that could be done is one the UI part, UI is a very important aspect for users. Users love better looking and simple to use UI, the ease of access of various features or the better in-depth features offered could make the UI part much better. This is why it's crucial to have a good UI. In this project, the UI used is a basic one which any average website uses. But to make it stand out, the UI part can be further modified or developed upon. Using new visualizations like graphs could make the users understand what exactly the point conveyed by the project is. Explanations with easy-to-read graphs and complex analysis of the background work done but yet keeping it simple for normal users to understand would improve the project much more.

Another work of area where the project can be improved upon is the actual accessibility of the project. Projects like these are only available on GitHub or certain other repositories like those. In reality, the project is developed to give users the access to better applications that is easily available. This can be achieved by hosting this project as an application or an actual website. As of now the project is developed similar to a website. By providing the backend support and the database, the website could be hosted and linked to the internet. After this, the website which hosts this prediction system would be available to access by anyone who has access to the internet. Now, users would be able to make use of this prediction system very easily. Furthermore, the project can be developed in the form of an application. Doing so would enable anyone to download a physical copy of the entire project and run it on the user's device to make use of the prediction system.

We have to collect all required data by giving GPS locations and information of a land and by taking access from Rain forecasting system of by the government, we can predict crops and fertilizers by just giving GPS location. Also, we can develop the model to avoid over and under crisis of the food. Complex neural network models like the CNN will be used to check for the higher accuracy and performance metrics.

These are some of the methods in which the project could be worked on in the future. Implementing more features like these would make the project much better and more beneficial for everyone.

REFERENCES

[1] https://en.wikipedia.org/wiki/Agriculture_ in_India

[2] Niketa Gandhi et al," Rice Crop Yield Forecasting of Tropical Wet and Dry Climatic Zone of India Using Data Mining Techniques",IEEE International Conference on Advances in Computer Applications (ICACA) , 2016.

[3] K.E. Eswari. L.Vinitha. " Crop Yield Prediction in Tamil Nadu Using Baysian Network ", International Journal of Intellectual Advancements and Research in Engineering Computations, Volume-6, Issue2, ISSN: 2348-2079.

[4] Shruti Mishra, Priyanka Paygude, Snehal Chaudhary, Sonali Idate "Use of Data Mining in Crop Yield Prediction" IEEE Xplore Compliant - Part Number: CFP18J06- ART, ISBN:978-1-5386-0807-4; DVD PartNumber: CFP18J06DVD, ISBN:978-1- 5386-0806-7.

[5] Anna Chlingaryana, Salah Sukkarieha, Brett Whelanb — Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review, Computers and Electronics in Agriculture 151 (2018) 61–69, Elisver,2018.

[6] Dakshayini Patil et al,"Rice Crop Yield Prediction using Data Mining Techniques:An Overview", International Journal of Advanced Research in Computer Science and Software Engineering ,Volume 7, Issue 5, May 2017.

[7] Mhudchuay T, Kasetkasem T, Attavanich W, Kumazawa I, Chanwimaluang T. Rice Cultivation Planning Using A Deep Learning Neural Network. In2019 16th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON) 2019 Jul 10 (pp. 822-825). IEEE.

[8] Sharma S, Rathee G, Saini H. Big data analytics for crop prediction mode using optimization technique. In2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC) 2018 Dec 20 (pp. 760- 764). IEEE.

[9] Athani SS, Tejeshwar CH. Support vector machine-based classification scheme of maize crop. In2017 IEEE 7th International Advance Computing Conference (IACC) 2017 Jan 5 (pp. 84-88). IEEE.

[10] Guruprasad RB, Saurav K, Randhawa S. Machine Learning Methodologies for Paddy Yield Estimation in India: a Case Study. InIGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium 2019 Jul 28 (pp. 7254-7257). IEEE.