European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 07, Issue 02, 2020

Crop Value Forecasting using Decision Tree Regressor and Model s

AkshayPrassanna S, B A Harshanand, B Srishti, Chaitanya R, KirubakaranNithiyaSoundari, SwathiSriram, V Manoj Kumar, VarshithaChennamsetti, Venkateshwaran G, Dr.Pramod Kumar Maurya

VIT University, Vellore, Tamil Nadu, India

Abstract – Machine Learning is an emerging research field which can be used for the analysis of crop price prediction and accurately provide solutions for the same. We can use this system as a backhand while we decide what a farmer should plant while considering factors such as annual rainfall, WPI and so on which is provided from the dataset and produce a logical conclusion on which products would give a more reliable outcome. The performance between Random forest ensemble learning and decision tree regressor is compared and it has been observed that the Random Forest Ensemble learning method gives a higher accuracy. In this system there are 23 crops whose information can be accessed upon for deciding collaborated with a simple user friendly UI

Keywords: Decision Tree, Ensemble Learning, Accuracy, Random Forest, Crop Prediction, Flask, Python

1. INTRODUCTION

Considering the contemporary world and all the changes that are happening, Farmers face a hard time in choosing the crops that they have to grow. In order to obtain a good amount of profit which helps them sustain the whole year, assisted by the upcoming Technology being Artificial Intelligence can give a big boost in their production. Hence we wish to provide an accurate solution to this current day problem by building a website that runs on flask hosting our model, easily accessible to any person anywhere. On hosting such a model a farmer who can access it can gain statistics of the price that each crop will sell and accordingly calculate which products he wants to plant and grow. This is done by using Python, Flask package for python, scikit-learn package for python, MaterializeCSS, Chart.js.

Crop Value Forecasting using Decision Tree Regressor and Model Boosting using Random Forest Ensemble Learning aims to solve crop value prediction problem in an efficient way in order to ensure guaranteed benefits to the poor farmers. India is a country where 52% of the population is engaged in farming out of which 82% farmers are small and marginal, it is our sole responsibility to provide a centralized system to solve the problem of irregular crop value forecasting and provide equal opportunities to our farmers. Our model will contain data of around 23 commodities (including all kind of crops). It will provide crop detailed forecast up to next 12 months. We tend to provide a more efficient mechanism that will predict the crop value with more accuracy than already existing models. The crop's annual Rainfall, WPI (Wholesale Price Index) datasets will be used for training the model. A user-friendly UI will be developed to project the forecasted Top Crop Gainers and Losers of current time.

2. LITERATURE REVIEW

Combined reinforcement learning and Q-learning to create Deep Q Networks.Uses the conventional training regime to train the hybrid model and finally uses it for crop yield prediction. Due to the use of RI

the system makes it easy for the against to predict the yield of crop using self-exploration and analysis.Also the ability of capturing the time dependencies makes the model more general that could be used widely. Use of RNN causes gradient explosion problems at times depending upon the data. Hence appropriate gradient clipping techniques should be used.[1]

Uses satellite images with labels regarding the yields of the period 2003 to 2016. Used the 3d convolutional neural network which consisted of both the 2d and the 3d convolutional layers. With the batch size of 30 images the neural network was run for 16 epochs. They came up with a conclusion that proves the theory of showing that the neural network shows the result differently based on the different area coverage percentages of crops. The error is almost 10% which should be decreased. They also found that after 20% of crop coverage area of dataframes, the error becomes very huge.[2]

Different classification models such as decision trees, support vector machine models, clustering methods and neural networks for different crops were used and compared. Instead of showing only one algorithm, they explained that each crop could perform better with a different type of algorithm and classified them. This showed amazing comparative results for each crop and their corresponding type of data. Accuracy can be improved more.[3]

Uses Random Forest Algorithm and Decision Tree to predict the annual crop yield. For the training purposes it uses data collected from Kaggle. Due to the relative simplicity of the algorithm implemented the system demands less computational resources and can also give out results fastly. It could only work on structured dataset. Hence it is not universally applicable to different kinds of data.[4]

It uses IOT based sensors to collect data from New Brunswick and Prince Island. Once the data is collected it uses different ML algorithms such as Linear Regression, KNN, Elastic-Net, Support Vector Regression and Linear Regression. The SVR method has the highest accuracy compared to the other methods. It was able to explain the influence of internal as well as external factors such as Climate, Environment etc in influencing the crop yield. Was tested on a small dataset. Testing it on a much larger dataset may make it more reliable.[5]

Combines CNN, RNN and FC layers to create a hybrid Neural Network Model. Two types of CNN used in the proposed system are W-CNN and S-CNN. Using the hybrid model it uses the conventional training methodology and later uses it for prediction. Due to the use of RNN it was able to capture time dependent inter-dependencies which in turn helped the model to achieve high accuracy in untested environments. The primary limitation of the paper is their black box property.[6]

Based on the remote sensing data collected it trains ML algorithms such as Regression models, SVM, Random Forest and Nearest Neighbour. Finally it estimates their performances by comparing their accuracies relatively to one another. Uses the past production also predicts the demand for each type of crop along with yield. This would help the farmers to decide which crop to cultivate. Considers only a limited set of factors such as weather and geographical area. The probability of success of such a system where other factors influence the yield of crops is less.[7]

The algorithm uses Decision Tree to classify and recommend the crops the farmers should grow. The dataset consisted of the recent year so that the results are more relevant. They also built a user friendly GUI that also has nlp algorithms that converts text to speech and vice versa. The GUI that contains nlp algorithms is very helpful for the usage of farmers. With the dataset always being updated to the most recent year's data, the results will be more relevant. There are not enough features that are to be considered in determining the crop. Adding more features is needed in order to get more accurate results.[8]

The research was done using an economic approach. It used the demand and supply curves to predict the demand for the production for the next year. It uses the DPFM model which has map and reduce functions to forecast the demand. They took a different approach and showed how hadoop is also used with the DPFM model. It uses parallel processing which reduces the time taken for execution greatly. The accuracy results shown by it are very bad. It has an RMSE of 46.7% and MAPE of 70.3%. This is very bad model for forecasting.[9]

AgMERRA (The meteorological data collected through the NASA satellites) is used and it also shows the precipitation and temperature of the place based on a grid system. The weighted regression methods such as Random Forest, Lasso and avg-5 models were used in the prediction. They took in the possibility that areas that have the features of a cultivable land needed more priority and hence used the weighted methods. Regrouping data was also done so that the variability within the year is also taken into account. It is an inconvenient model as it requires the knowledge of the cultivated area of the present year which is not realistic. The mean absolute error is 5.5% which can be improved.[10]

Uses the support vector machines approach (SVM). It uses the SMO algorithm. For the dataset, they collected information on the government website using the WEKA tool. The preprocessing methods in their methodology were very thorough such that no unwanted columns and rows were present. Hence the noise was removed thoroughly. SVM is not a suitable model for predicting as it only gives about 78.76% accuracy.[11]

Implemented and compared the performance of autoregressive integrated moving average (ARIMA), the partial least square (PLS), artificial neural network (ANN), and PLS further integrated with response surface methodology (RSM) called RSMPLS. RSMPLS can be used for instigating the non-linear relationship between historical prices. Also, The developed service is integrated in the smart agri-management platform, providing an interface for historical price retrieval and future price forecast. Model only uses one parameter (prices). More features such as climate, location of the market, planting area will give more accurate results.[12]

The Notations are first directed which is basically the dataset which is pre processed and this is sent to the algorithms which are Nearest Neighbour,Inverse Distance Weighting, Krigning Method, this is then Processed via an Artificial Neural Network and the output is then collected. Kriging approach is also recommended for the development of forecasting service and is also found to be the most efficient algorithm. Other spacial and temporal features, such as the climate in different regions, location of the planting area, and historical trading amounts are not considered.[13]

A decision making support model has been implemented that can be helpful for farmers to predict prices. This model includes a portal in which farmers are required login their account with the credentials.Farmers have to enter commodity name and previous selling price of the crop. Based on the previous prices, this model will be able to provide average prices for a particular crop which will be beneficial for farmers to make better decisions and predict prices. The data they are using to predict the crop price is not efficient as they are using the previously logged in prices by users and basically just taking an average. It is not taking any other parameter into consideration and hence, this model will not be always yielding correct accurate results.[14]

This study incorporates an ARIMA as the FSM for computational intelligence (CI) models to predict three major food. the components of the proposed integrated forecasting models include artificial neural networks (ANNs), support vector regression (SVR), and multivariate adaptive regression splines (MARS). They have compared the performance of various models and also for ARIMA models, their

structures will remain unchanged over time. The techniques for identifying the correct ARIMA model from the variety of possible models may be unintuitive and computationally expensive.[15]

Hybrid Association Decision Tree algorithm (HADT)is used with Rule creation. Multiple rules are generated using classification algorithm.Reduce Phase is then attained finally accompanied by the Pattern and model Creation. The performance of the proposed algorithm is measured and compared with the existing algorithms for data mining. The performance is measured by accuracy, rate of error, and time of execution. And it can be seen that the HADT classifier gives the best results in all aspects of measures. Also, the proposed approach is suitable for large datasets. Implementing priceforecasting in the suggested scheme continues step-by-step,leading to time-consuming procedures.[16]

Deep learning architectures, including CNNs and LSTMs, are trained on the histograms, developed from raw images, to predict crop yields. This model outperforms traditional remote-sensing based methods by 30% in terms ofRoot Mean Squared Error (RMSE), and USDA national- level estimates by 15% in terms of Mean Absolute Percentage Error (MAPE). It does not explicitly account for spatio-temporal dependencies between data points[17]

Conjectures the climate, yield and cost of significant harvests of Andhra Pradesh and Uttar Pradesh dependent on recorded information. Particularly, for Srikakulam, in light of the fact that they are the biggest maker of Millets in Andhra Pradesh. This method guides ranchers to settle on the yield they might want to plant for the imminent year, which causes them to acquire extreme cost for their items. Accuracy of prediction is less.[18]

A Back propagation neural network models is developed on the basis of 4 vegetation indices. These indices includes NVDI, GVI, SAVI and PVI. In the models, the input and output parameters were not transformed, i.e., the actual gray value statistics obtained from image processing were used. The accuracies over the years had a maximum of 96% and the average was 93%. PVI based models produced better results compared to other indices. The accuracy on the model for each year has high inflation. This proves the inconsistency on collection of data. [19]

This project is undertaken using machine learning and evaluates the performance by using Random forest, Polynomial Regression and Decision Tree algorithms. The proposed approach in this research paper, aims at predicting the best yielded crop for a particular region by analyzing various atmospheric factors.This paper analyzed that proposed model has got more efficiency than the existing model for finding crop yield. It may not work well when there are non-linear relationships between dependent and independent variables.[21]

Use of computer construction models to identify the crop production potential based on temperature and irradiance as parameters for carbon-3 crops corresponding to CO2 digestion factors. Can be used to integrate temperature , soil CO2 levels and irradiance as parameters for other crop datasets to estimate crop yield will give a more comprehensive model. Acquiring pre-processed parameters from other datasets may not be compatible to integration.[21]

This paper identifies 3 gaps in the existing applications and tackles each of them using a 3D CNN model with data set from satellite images. Solves standardized training protocol that specifies the optimal time frame, both in terms of years and months of each year, to be considered in the training set, verified applicability to developing countries under the condition of scarce data and effective utilisation of spatial features in remote sensing images. This model should be trained based on specific regions. Meaning not a generalised model.[22]

Random Forest Algorithm integrated with MapReduce programming on Hadoop system. Will provide efficiency in handling large amounts of data in order to improve the scale of the process. May not be a modular and platform independent solution for crop yield prediction based problems.[23]

Different MLR , data mining techniques are being used to make the model for prediction.The results so obtained were verified and analyzed using the Data Mining technique namely Density-based clustering technique Due to the use of clustering and data mining techniques, the accuracy range of the result is precise. When tested in Andrapradesh, this model should accuracy to only a specific crop. Need to build a model which is used for any type of crop.[24]

Support Vector Machine, Back Propagation Neural Network. Use of big data models integrated with an ensemble set can handle larger scales of data. Maintaining the accuracy and efficiency of the ensemble while integrating big data techniques.[25]

The multivariate enhancement of previous crop yield, soil moisture and rainfall data as parameters can be used to generate a crop yield condition for each area in India. The non-linear loop wise generation of the data factors may be used as a principle for the ensemble algorithm used. The crop yield condition may show high fluctuation when incorporated to an ensemble machine due to its iterative process.[26]

This paper uses Machine learning and supervised and unsupervised learning and proposes a software application to predict crop yield from past data.It focuses on the creation of a prediction model which is used to predict the maximum production rate of the crop . Using polynomial regression, it finds relationships between the independent and dependent variables. Since this methodology used supervised learning ML, it has shown very less errors. Though there are very few errors, the accuracy is only 80%. Need to improve the accuracy.[27]

Supervised Machine Learning and Artificial Neural Networks such as Kohenon'sSOM,Back Advertising Network. May be integrated and configured with the existing ensemble set to improve the iterative efficiency. Configuring an ANN to a specific dataset based on previous forms of the ensemble set.[28]

Remote sensing and GIS techniques were employed, in this study, to predict potato tuber cropyield.Potato yield samples were collected 2–3 days prior to the harvest time and were correlated to the adjacent NDVI and SAVI, where yield prediction algorithms were developed and used to generate prediction yield maps. The highest correlation was observed for the field number 67-S, where both cumulative SAVI (Landsat-8) and single-date SAVI (Sentinel-2) resulted in a similar R2 value of 0.65. Algorithm needs to be improvised to get an better accuracy.[29]

They talk about the approach to yield modeling that uses a semi-parametric variant of a deep neural network.They describe a semi-parametric approach that fuses the two and works better than either alone in terms of predictive performance. In order to get more accurate , they used a another technique called dropout, which almost gives an accurate result after iterations. It is a time consuming procedure to get the results.[30]

The Dataset is a custom made dataset that was made out of images that were captured using the Draganfly X4P quad-copter A DCNN based architecture called LodgedNet is proposed which has 3 main components: a CNN backbone, atexture feature extraction module, and a classification module with which, Feature Extraction, Data Augmentation and Training LodgedNet is a very efficient approach to Detection and prediction for the crops which is evidently because of 2 feature detectors. One limitation of the transfer learning used for the 10 DCNN-based models is that they have been trained using RGB images[31]

Using clustering , built a final model is a varying-coefficient model where the temperature and rainfall predictors are replaced by the scores of the functional principal components which explain most of the variability in temperature and rainfall. Heir model is easy to interpret and implement, and therefore, easy to be incorporated in the existing software for 3 agribusiness decisions. Accuracy has to be improved.[32]

This Paper uses Copula Models which has a dependence structure between the random variables used.Using Skylar's theorem and Treating Rainfall as censored Variables an Estimation Procedure is done for copula parameters for each site. 90% confident intervals of ACCESS-S raw, QM and ECPP forecastsIs shown in the results of this paper and thereby it is considered a pretty sharp one in approach. The raw forecasts are the least reliableforecasts for both non-accumulated and accumulated dailyrainfall.[33]

They are implementing ANN which is used for classification and prediction as most of intense computation takes place during training phase only, there is no requirement of testing phase. ANN is very helpful in giving accurate results. This process is time consuming. Only if after the iteration, if the outcome is accurate, it will be used further. Elese, the process will continued to till they get an accurate result.[34]

The Daily Fire pixel counts were first calculated and preprocessed as data.Then a BPNN ensembles the fire pixel counts. A three layer structure for each BPNN member is Constructed with One input layer, one hidden layer and one Output layer. The use of BPNN or other machine learning techniques in forecasting pollutant emissions and air quality shows promise and warrants further investigation. This model is trained for a Very Large dataset and thereby might produce less accurate results when fed with a small dataset which has lesser inputs, for the case of processing.[35]

This paper is using optical and microwave imagery, along with weather data, and priors generated by the HSB model in 2016 to train the CNN model.This HSB Model used to generate The prior for the CDL was a simple stochastic Markov chain that learnt cropping dynamics from historical crop sequences for every pixel. Accuracy is High as much as 85 - 95%

More research is needed, especially in using (RNNs), to better model the temporal dynamics of this process.[36]

This paper suggests how stepwise multiple linear regression techniques can be successfully used for pre-harvest crop yield forecasting. This model is very consistent and can be applied on zone or state level. The study also shows that use of de-trended yield data in model development gets most appropriate pre-harvest forecast models. Develops preharvest forecasting models and these forecasts have significant value in agricultural planning. This model is consistent and can be applied only on zone or state level. Otherwise, it is less accurate.[37]

Moderate Resolution Imaging Spectroradiometer (MODIS) imagery was obtained from theDescartes Labs Satellite Platform Interaction with the Descartes Labs Satellite Platform is done through a python console. MOD09 Aqua and Terra surface reflectance data points with associated coordinatesareinterpolated onto a grid in the form of an image. Model has been implemented in Illinois and Africa hence reliable and tested wide scale Model has been trained only for large scale datasets.[38]

This paper presented a machine learning approach for crop yield prediction using large datasets of corn hybrids. The carefully designed deep neural networks were able to learn nonlinear and complex relationships between genes and environmental condition. The feature selection approach successfully found important features,and revealed that environmental factors had a greater effect on the crop yield

than genotype. Complex model structure makes it hard to produce testable hypotheses that could potentially provide biological insights[39]

This paper proposes how to simulate the crop yield under a plausible change in climate for the coastal areas of South India through the end of this century. The crop simulation model, the Decision Support System for Agrotechnology Transfer. Continuous field observations have been carried out to understand the real time situation of the agronomical practices and the management for the major crops rice, groundnut, and sugarcane. We need to Improve the model due to vulnerability issues[40]

(Computer and electronics in agriculture SVM, K-means. Kmeans was able to accurately produce required results. K means accuracy reduced as the classification count increased.[41]

In this paper they have provided Neural Network models that predicted weekly yield using all above data plus analysis individual data in comparison with total yield in each week to realize the most and least important data set on our prediction in order to address for future research that would optimize weekly crop yield consequently. The paper shows how advanced greenhouse monitoring systems, coupled with the ability to collect physiological data about the plants, provide us with potentially useful information that can be used to create accurate yield prediction models. Model is less accurate.[42]

Logistic regression model with a selected feature selector. Predicts order and the type of crop to plant. The proposed method does not increase crop yeild.[43]

C-4.5 algorithm. User friendly webpage. Complex and unheard of algorithm[44]

Ada SVM, Ada Naive. Benefits: Shows improvement over classic naive bayes and SVM. Very vulnerable to noise so any corruption in the dataset could lead to errors[45]

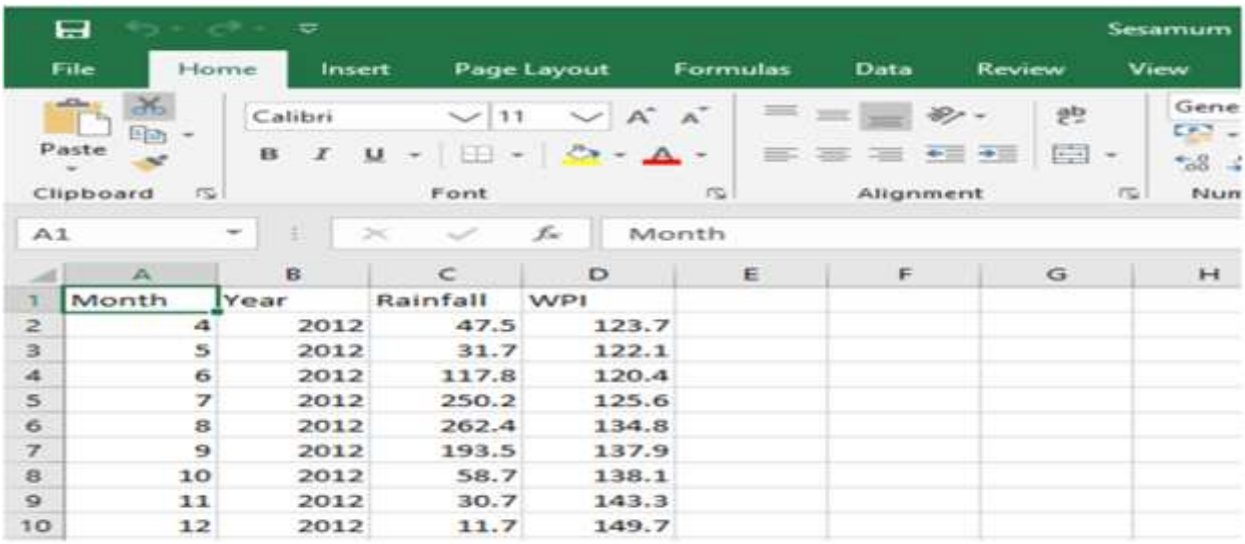
Methodology: Hierarchical Clustering. Dataset was developed using satellite data which makes the dataset unique and error free. The algorithm is very dataset centric.Such datasets obtained from satellite is hard to acquire for this project. As the original dataset was done on canadian grounds, it wouldn't be ideal to consider it as a dataset in india.[46]

3. Preliminaries

Requirements

The main requirements are python and flask. Python is a high level language and the system models are defined in python. Thus, a fully functioning python 3.5+ environment with sklearn, numpy and pandas libraries is required. The html-python framework flask is another necessity. Flask is called a web framework, it is designed in python. It has no particular tools or library requirements of it's own. It is also a backend based microframework. HTML CSS are required for creating the front end of the website.

So the requirements are: HTML, CSS for the front end, python for defining and running the system models and flask for integrating the front end with python.



	A	B	C	D	E	F	G	H
1	Month	Year	Rainfall	WPI				
2	4	2012	47.5	123.7				
3	5	2012	31.7	122.1				
4	6	2012	117.8	120.4				
5	7	2012	250.2	125.6				
6	8	2012	262.4	134.8				
7	9	2012	193.5	137.9				
8	10	2012	58.7	138.1				
9	11	2012	30.7	143.3				
10	12	2012	11.7	149.7				

Dataset

The dataset used is called crop dataset. This crop dataset is an authentic dataset got from https://data.gov.in. There are 23 crops in the dataset which are - Arhar, Bajra, Barley, Copra, Cotton, Gram, Groundnut, Jowar, Jute, Maize, Masoor, Moong, Niger, Paddy, Ragi, Rape, Safflower, Sesamum, Soyabean, Sugarcane, Sunflower, Urad and Wheat. Each crop has its own specific file/dataset. Each file contains details about Annual Rainfall and Wholesale Price Index (WPI) from April 2012 to December 2018.

For example, the picture attached below is from the file called sesamum which is the dataset for the crop sesamum.

4. System Model

Decision Tree Iterative Dichotomiser 3 Algorithm

A decision tree is a data structure that is built using nodes (containing values or conditions) and edges (connecting all nodes). The decision tree is constructed based on a dataset containing attributes or features which classify the raw data for each record. Each node in the tree can be either a decision node that makes decisions or a leaf node that gives the outcome.

Basically, the ID3 Algorithm which is a precursor to the C4.5 algorithm attempts to create the smallest possible decision tree. It uses entropy of each attribute to decide which edge is to be followed.

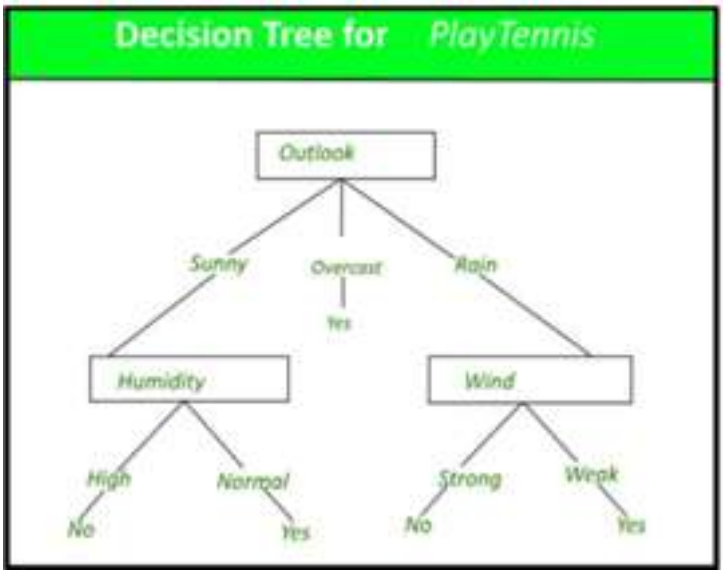
Since ID3 is a greedy top down approach it identifies attribute with extreme values (Highest Information Gain attributes). This helps to identify which attributes create the most homogenous branches as shown below.[33][34]

Random Forest Algorithm

The random forest algorithm is a supervised learning algorithm which can apply for both classification and regression type datasets. It will randomly select a certain set of features from the given dataset attributes and create a set of decision trees by finding the root nodes and splitting the attributes.

After creating the forest then the best decision is chosen based on highest votes acquired among the predicted targets as the final prediction from the classifier.

The random forest classifier accuracy increases with increase in number of trees in the forest.



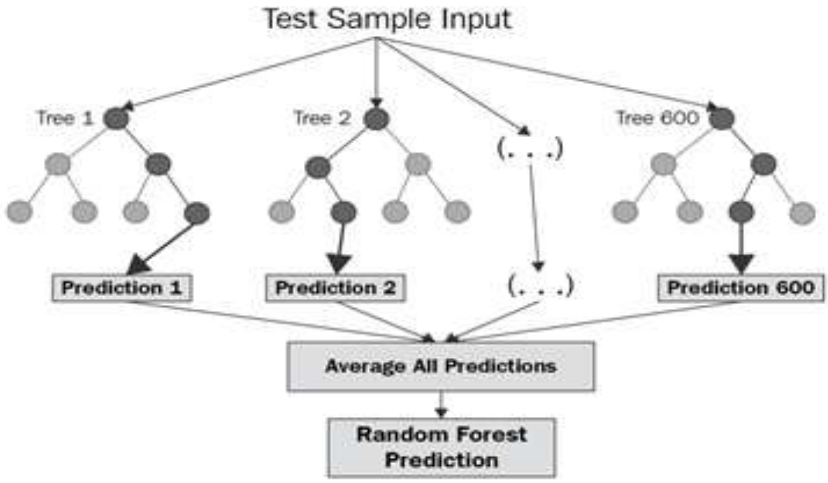
Ensemble Classifier

Since Random Forest may involve overfitting (classification into too many features) it is generally advised to use two or more models in conjunction with it by ensemble methods.[23]

Even in ensemble methods the most suitable for increasing accuracy will be by parallelising two or more models since the independence between the learners can be used to reduce error by law of averages.

Furthermore, we can use models of diverse nature to further increase efficiency compared to the individual base learning models. This involves heterogenous ensemble methods.

In case of the most randomised forests splitting thresholds of each tree is again randomised by selecting from each candidate feature to identify the extreme threshold splitting rule. This may help in reducing the variation in results for the mode



.

5. Process

The problem statement of the project is to Crop value forecasting using Decision tree regressor and model boosting using random forest ensemble learning. The basic process of this project is that we will preprocess the data provided to us, then it is used to prepare the model for the backend and using flask to connect it to the UI interface to show the full and final output.

The dataset in this project was taken from the website https://data.gov.in/.It contains Annual Rainfall and Wholesale Price Index (WPI) from April 2012 to December 2018 for 23 different crops namely – Arhar, Bajra, Barley, Copra, Cotton, Gram, Groundnut, Jowar, Jute, Maize, Masoor, Moong, Niger, Paddy, Ragi, Rape, Safflower, Sesamum, Soyabean, Sugarcane, Sunflower, Urad and Wheat. Each one contains attributes like, 'Month', 'Year', 'rain', 'WPI'.The Dataset is preprocessed to get the dataset will work with.

In this project there are two machine learning algorithms used. One of them is Decision tree regressor and Random forest. The decision tree regressors's role in model building is to build regression and classification models in the form of a tree structure. What it does is to break down a dataset into a smaller subset and simultaneously an associated decision tree is developed. The tree consists of nodes and leafs which represents values for attributes tested. This tree handles categorical and numerical data. ID3 is the core algorithm. It uses a greedy top-down search through the space of possible branches with no backtracking. The other algorithm is Random forest which is a Supervised Learning Algorithm and is a bagging technique which uses Ensemble learning. Ensemble combines the predictions obtained together to make accurate predictions. It will be used as a meta-estimator which aggregates many decision trees, with some helpful modifications.[34][23]

This generated 600 decision trees where we got in total of 600 predictions, one from each tree. Then all the predictions are gathered and an average is calculated. It is then feeded into the random forest predictor which eventually gives us an output. This is done by the ensemble so that we will get an exact accuracy.

A python package index(PPI) is used here called Flask.It is used to deploy the project. An instance for the flask is being created. It acts as a Central configuration object which sets up the pieces of the application required. Flask Framework does not have built in database facilities. It does have a package which helps in connecting the UI interface to a SQL database.

The User interface is made with HTML/CSS to display the details of each crop. The increase and decrease in the price of each crop according to the months and a graph depicting the same. The UI is connected to the model with the help of Flask framework.

6. Comparison

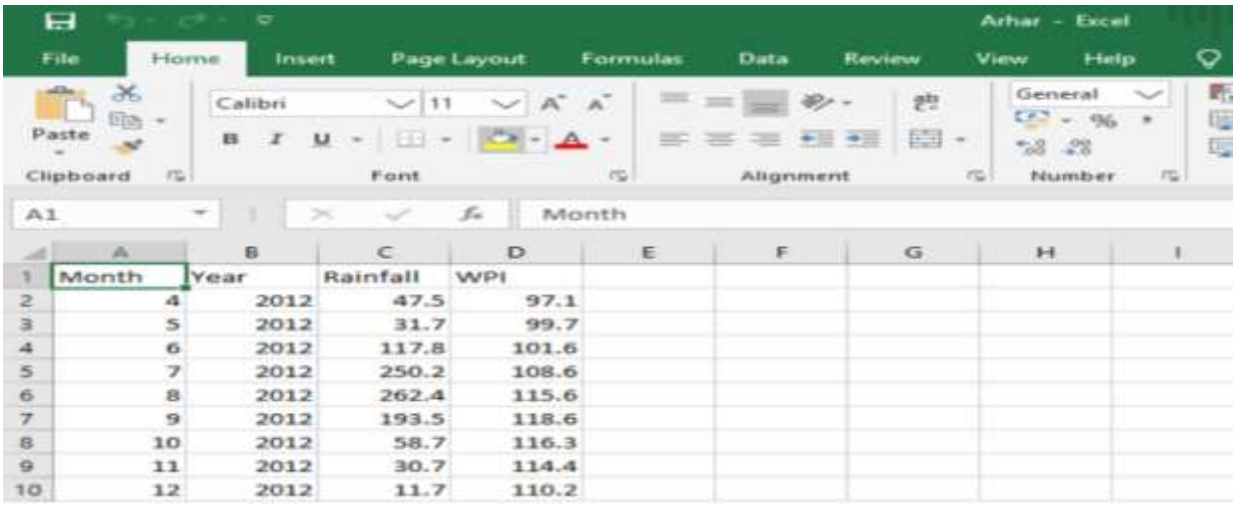
S.No.	Random Forest	Decision Tree
1.	A Supervised learning algorithm	A Supervised learning algorithm
2.	Multiple decision trees are trained parallelly for prediction	Trains only one decision tree
3.	Subsets are created from data set to train multiple decision trees	Data set is not divided
4.	Better accuracy	Accuracy is comparatively low
5.	Possibility of overfitting is low	Overfitting can happen
6.	Does not require normalization or scaling	Does not require normalization or scaling
7.	Relatively slower to build	Easily built
8.	Hard to interpret	Easy to interpret
9.	Computationally expensive	Computationally less expensive
10.	Can work with large datasets	Large datasets are not well suited

Random forest regressor is one among the most accurate models available. The main feature of this model is that it reduces the possibility of overfitting. This happens through the bagging technique, in which the subsets are created randomly and trained parallelly without any interaction between them. SVM, ANN, Naïve Bayes and Gradient Boosting are also used as crop value prediction models. But Random forest regressor is the fittest for this scenario. Our dataset is a categorical one, in which the collinearity of the random forest regressor is better than that of SVM. Random forest regressor is less computationally expensive than ANN and also requires a low amount of data. Naïve Bayes also requires a high amount of data to work better than Random forest. Random forest regressor beats Gradient boosting on the possibility of overfitting in our model. As far as this dataset is concerned we believe Random forest regressor provides better results.

7. Result Screenshots

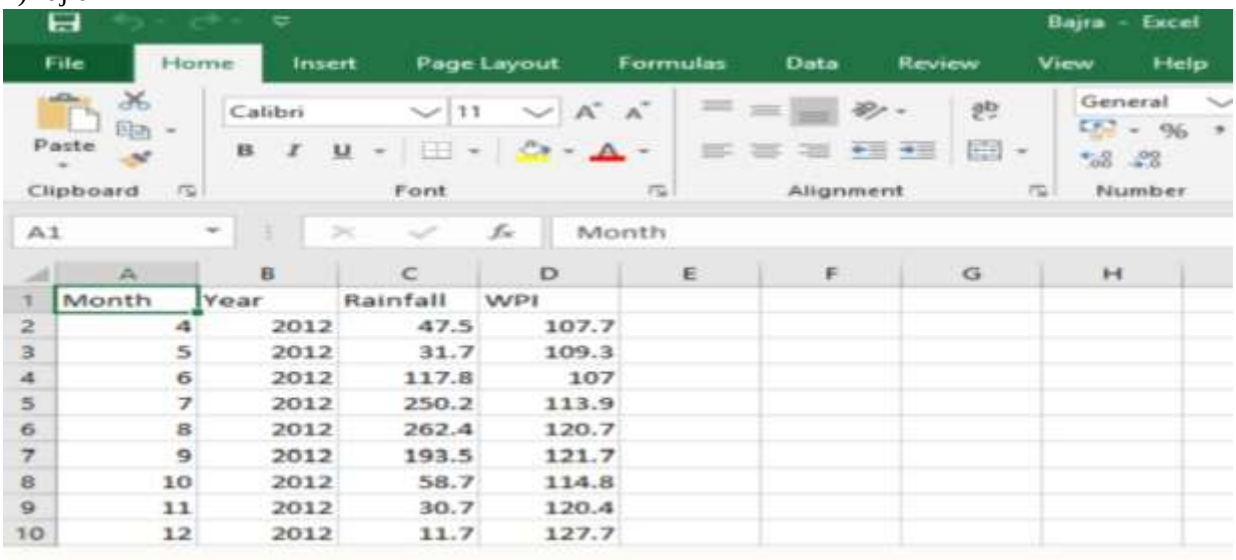
Some of the crops:

1)Arhar



	A	B	C	D	E	F	G	H	I
1	Month	Year	Rainfall	WPI					
2	4	2012	47.5	97.1					
3	5	2012	31.7	99.7					
4	6	2012	117.8	101.6					
5	7	2012	250.2	108.6					
6	8	2012	262.4	115.6					
7	9	2012	193.5	118.6					
8	10	2012	58.7	116.3					
9	11	2012	30.7	114.4					
10	12	2012	11.7	110.2					

2)Bajra



Bajra - Excel

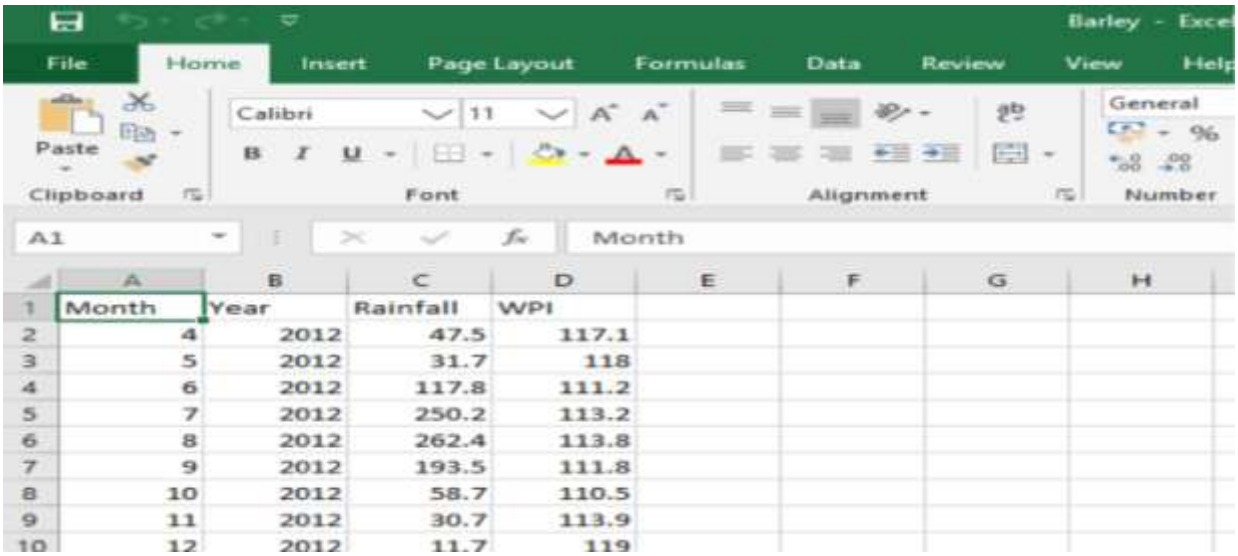
File Home Insert Page Layout Formulas Data Review View Help

Paste Clipboard Calibri 11 Font Alignment General Number

A1 Month

	A	B	C	D	E	F	G	H
1	Month	Year	Rainfall	WPI				
2	4	2012	47.5	107.7				
3	5	2012	31.7	109.3				
4	6	2012	117.8	107				
5	7	2012	250.2	113.9				
6	8	2012	262.4	120.7				
7	9	2012	193.5	121.7				
8	10	2012	58.7	114.8				
9	11	2012	30.7	120.4				
10	12	2012	11.7	127.7				

3)Barley



Barley - Excel

File Home Insert Page Layout Formulas Data Review View Help

Paste Clipboard Calibri 11 Font Alignment General Number

A1 Month

	A	B	C	D	E	F	G	H
1	Month	Year	Rainfall	WPI				
2	4	2012	47.5	117.1				
3	5	2012	31.7	118				
4	6	2012	117.8	111.2				
5	7	2012	250.2	113.2				
6	8	2012	262.4	113.8				
7	9	2012	193.5	111.8				
8	10	2012	58.7	110.5				
9	11	2012	30.7	113.9				
10	12	2012	11.7	119				

Implementation:

Sample Code for Decision Tree:

```
commodity_list = []

class Commodity:

    def __init__(self, csv_name):
        self.name = csv_name
        dataset = pd.read_csv(csv_name)
        self.X = dataset.iloc[:, :-1].values
        self.Y = dataset.iloc[:, 3].values

        #from sklearn.model_selection import train_test_split
        #X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.1, random_state=0)

        # Fitting decision tree regression to dataset
        from sklearn.tree import DecisionTreeRegressor
        self.regressor = DecisionTreeRegressor(max_depth=10, random_state=0)
        self.regressor.fit(self.X, self.Y)
        #y_pred_tree = self.regressor.predict(X_test)
        # fsa=np.array([float(1),2019,45]).reshape(1,3)
        # fask=regressor_tree.predict(fsa)

    def getPredictedValue(self, value):
        if value[1]>=2019:
            fsa = np.array(value).reshape(1, 3)
            #print(" ",self.regressor.predict(fsa)[0])
            return self.regressor.predict(fsa)[0]
        else:
            c=self.X[:,0:2]
            x=[]
            for i in c:
                x.append(i.tolist())
            fsa = [value[0], value[1]]
            ind = 0
            for i in range(0,len(x)):
                if x[i]==fsa:
                    ind=i
                    break
            #print(index, " ",ind)
            #print(x[ind])
            #print(self.Y[i])
```

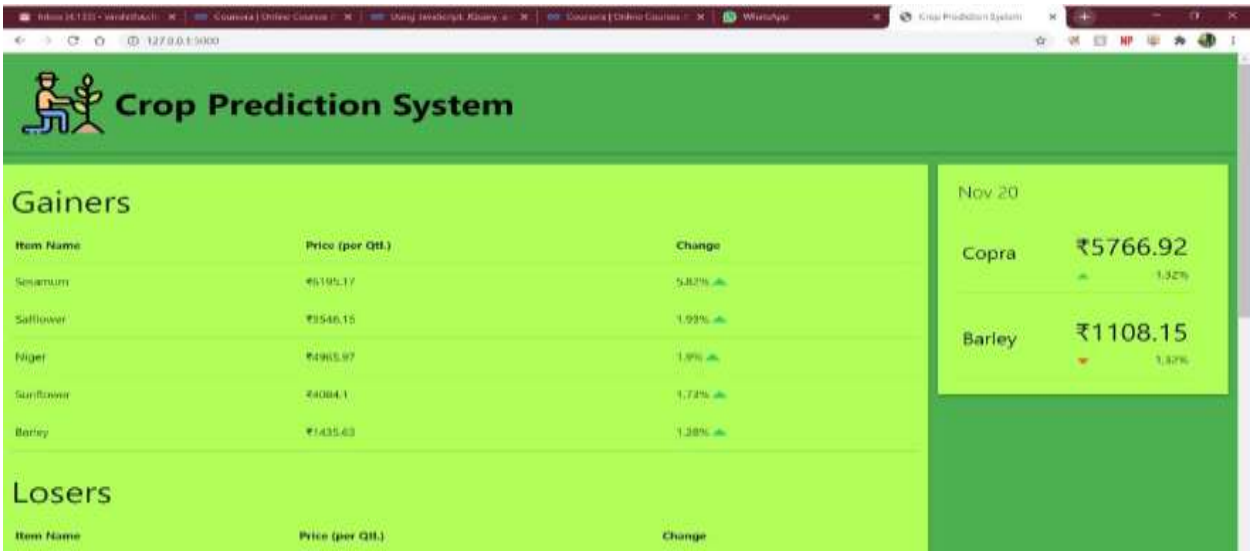
Sample Code for Random Ensemble:

```
commodity_list = []

class Commodity:

    def __init__(self, csv_name):
        self.name = csv_name
        dataset = pd.read_csv(csv_name)
        self.X = dataset.iloc[:, :-1].values
        self.Y = dataset.iloc[:, 3].values

        #from sklearn.model_selection import train_test_split
        #X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.1, random_state=0)

        # Fitting random forest regressor to dataset
        from sklearn.ensemble import RandomForestRegressor
        self.regressor = RandomForestRegressor(max_depth=2, random_state=0)
        self.regressor.fit(self.X, self.Y)
        #y_pred_tree = self.regressor.predict(X_test)
        # fsa=np.array([float(1),2019,45]).reshape(1,3)
        # fask=regressor_tree.predict(fsa)

    def getPredictedValue(self, value):
        if value[1]>=2019:
            fsa = np.array(value).reshape(1, 3)
            #print(" ",self.regressor.predict(fsa)[0])
            return self.regressor.predict(fsa)[0]
        else:
            c=self.X[:,0:2]
            x=[]
            for i in c:
                x.append(i.tolist())
            fsa = [value[0], value[1]]
            ind = 0
            for i in range(0,len(x)):
                if x[i]==fsa:
                    ind=i
                    break
            #print(index, " ",ind)
            #print(x[ind])
            #print(self.Y[i])
```

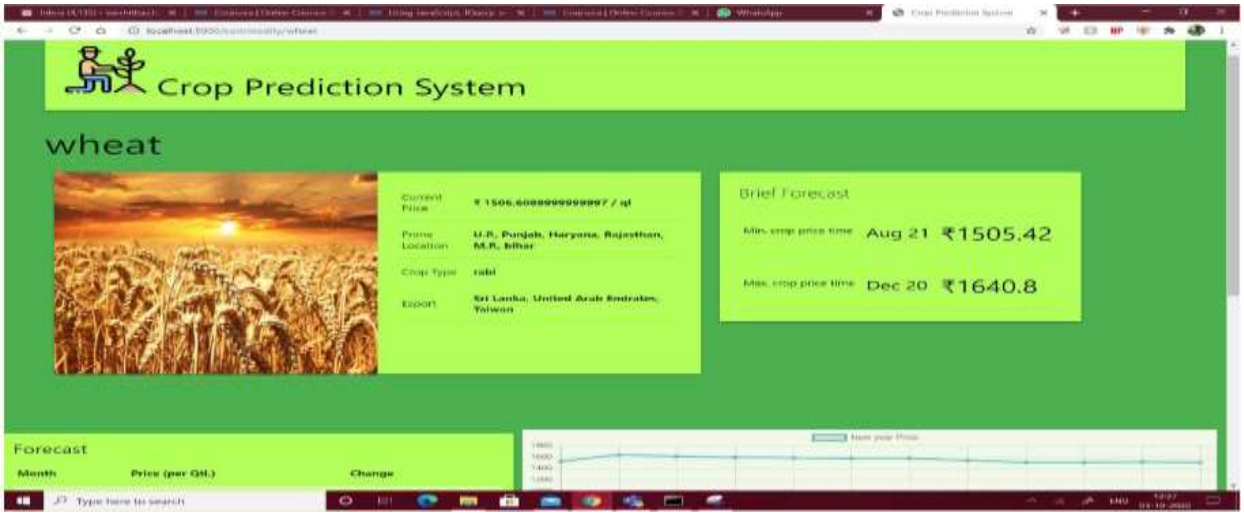
Sample Code running Server:

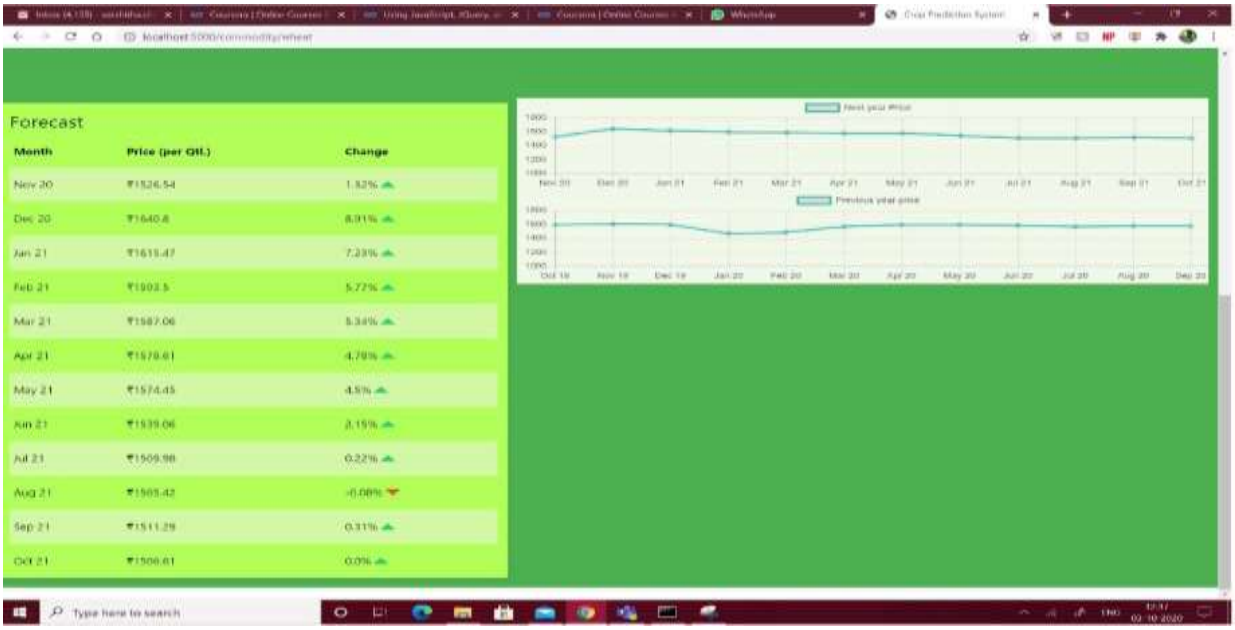
```
if __name__ == "__main__":
    arhar = Commodity(commodity_dict["arhar"])
    commodity_list.append(arhar)
    bajra = Commodity(commodity_dict["bajra"])
    commodity_list.append(bajra)
    barley = Commodity(commodity_dict["barley"])
    commodity_list.append(barley)
    copra = Commodity(commodity_dict["copra"])
    commodity_list.append(copra)
    cotton = Commodity(commodity_dict["cotton"])
    commodity_list.append(cotton)
    sesamum = Commodity(commodity_dict["sesamum"])
    commodity_list.append(sesamum)
    gram = Commodity(commodity_dict["gram"])
    commodity_list.append(gram)
    groundnut = Commodity(commodity_dict["groundnut"])
    commodity_list.append(groundnut)
    jowar = Commodity(commodity_dict["jowar"])
    commodity_list.append(jowar)
    maize = Commodity(commodity_dict["maize"])
    commodity_list.append(maize)
    masoor = Commodity(commodity_dict["masoor"])
    commodity_list.append(masoor)
    moong = Commodity(commodity_dict["moong"])
    commodity_list.append(moong)
    niger = Commodity(commodity_dict["niger"])
    commodity_list.append(niger)
    paddy = Commodity(commodity_dict["paddy"])
    commodity_list.append(paddy)
    ragi = Commodity(commodity_dict["ragi"])
    commodity_list.append(ragi)
    rape = Commodity(commodity_dict["rape"])
    commodity_list.append(rape)
    jute = Commodity(commodity_dict["jute"])
    commodity_list.append(jute)
    safflower = Commodity(commodity_dict["safflower"])
    commodity_list.append(safflower)
    soyabean = Commodity(commodity_dict["soyabean"])
    commodity_list.append(soyabean)
    sugarcane = Commodity(commodity_dict["sugarcane"])
    commodity_list.append(sugarcane)
    sunflower = Commodity(commodity_dict["sunflower"])
    commodity_list.append(sunflower)
    urad = Commodity(commodity_dict["urad"])
    commodity_list.append(urad)
    wheat = Commodity(commodity_dict["wheat"])
    commodity_list.append(wheat)

    app.run(debug=False)
```

Website:

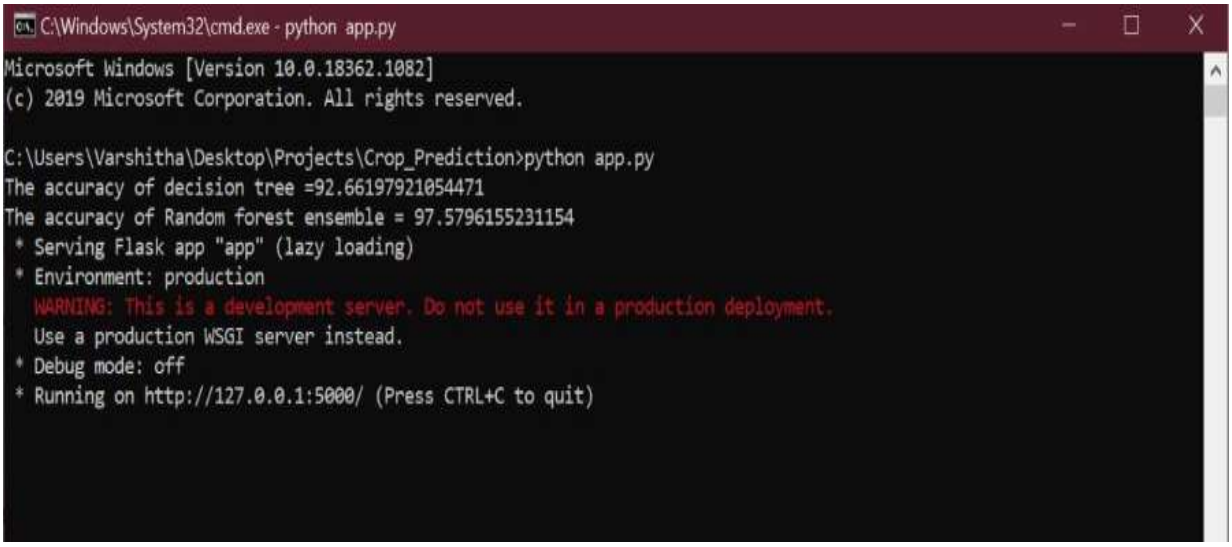



Explore By Commodity
Paddy
Wheat
Barley
Soya Bean
Cotton
Coconut
Ground Nut Seeds
Mustard Seed
Gingelly Seed(Sesamum)
Gram
Sugarcane
Arhar
Ragi
Maize
Moong
Masoor
Urad
Raw Jute
Niger Seed
Kardi Seed
Jowar
Bajra


Crop Prediction System
wheat
Brief Forecast
Min. crop price time Aug 21 ₹1505.42
Max. crop price time Dec 20 ₹1640.8
Forecast
Month
Price (per Qtl.)
Change



Result Screenshot:



8. Results

Accuracy Comparison:
Decision Tree –92.66%
Random Forest Ensemble Learning –97.57%

- When using the Decision Tree Model, the accuracy produced was around 92.66%. The top crop gainers predicted are Maize with +3.32% increase and Sunflower with 2.43% increase. The crop losers predicted are Niger with -7.81% decrease, Moong with -4.2% decrease and Masoor with -2.4% decrease.

- In order to increase performance the new method used was Random Forest Ensemble Method which provided an accuracy of around 97.57%which is more than that produced by the Decision Tree Model. The top crop gainers predicted are Rape with +0.92% increase and Groundnut with +0.445 increase. The top crop losers are Sunflower with -0.54% decrease and Arhar with -0.2% decrease. Hence the comparison clearly shows that ensemble methods are always better to improve performance and efficiency by giving high accuracy.

REFERENCES

[1] Elavarasan, D., & Vincent, P. D. (2020). Crop Yield Prediction Using Deep Reinforcement Learning Model for Sustainable Agrarian Applications. IEEE Access, 8, 86886-86901.

[2] Terliksiz, A. S., &Altýlar, D. T. (2019, July). Use Of Deep Neural Networks For Crop Yield Prediction: A Case Study Of Soybean Yield in Lauderdale County, Alabama, USA. In 2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics) (pp. 1-4). IEEE.

[3]Gandge, Y. (2017, December). A study on various data mining techniques for crop yield prediction. In 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT) (pp. 420-423). IEEE.

[4] Kumar, Y. J. N., Spandana, V., Vaishnavi, V. S., Neha, K., & Devi, V. G. R. R. (2020, June). Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector. In 2020 5th International Conference on Communication and Electronics Systems (ICCES) (pp. 736-741). IEEE.

[5] Abbas, F., Afzaal, H., Farooque, A. A., & Tang, S. (2020). Crop yield prediction through proximal sensing and machine learning algorithms. Agronomy, 10(7), 1046.

[6] Khaki, S., Wang, L., &Archontoulis, S. V. (2020). A cnn-rnn framework for crop yield prediction. Frontiers in Plant Science, 10, 1750.

[7] Suganya, M. (2020). Crop Yield Prediction Using Supervised Learning Techniques. International Journal of Computer Engineering and Technology, 11(2).

[8] Raja, S. K. S., Rishi, R., Sundaresan, E., &Srijit, V. (2017, April). Demand based crop recommender system for farmers. In 2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR) (pp. 194-199). IEEE.

[9] Prabhu, B. B., &Dakshayini, M. (2017, November). Demand-prediction model for forecasting AGRI-needs of the society. In 2017 International Conference on Inventive Computing and Informatics (ICICI) (pp. 430-435). IEEE.

[10] Chen, X., Bayol, B., &Cournède, P. H. (2018, November). Application of Weighted Regression for the Prediction of Soft Wheat Production in France. In 2018 6th International Symposium on Plant Growth Modeling, Simulation, Visualization and Applications (PMA) (pp. 141-146). IEEE.

[11] Gandhi, N., Armstrong, L. J., Petkar, O., &Tripathy, A. K. (2016, July). Rice crop yield prediction in India using support vector machines. In 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 1-5). IEEE.

[12] CPeng, Y. H., Hsu, C. S., & Huang, P. C. (2015, November). Developing crop price forecasting service using open data from Taiwan markets. In 2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI) (pp. 172-175). IEEE.

[13]Peng, Y. H., Hsu, C. S., & Huang, P. C. (2015, November). An investigation of spacial approaches for crop price forecasting in different Taiwan markets. In *2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI)* (pp. 176-179). IEEE.

[14] Vohra, A., Pandey, N., & Khatri, S. K. (2019, February). Decision Making Support System for Prediction of Prices in Agricultural Commodity. In 2019 Amity International Conference on Artificial Intelligence (AICAI) (pp. 345-348). IEEE.

[15]Shao, Y. E., & Dai, J. T. (2018). Integrated feature selection of ARIMA with computational intelligence approaches for food crop price prediction. *Complexity, 2018*.

[16]Rajeswari, S., &Suthendran, K. Developing an Agricultural Product Price Prediction Model using HADT Algorithm.

[17]You, J., Li, X., Low, M., Lobell, D., &Ermon, S. (2017, February). Deep gaussian process for crop yield prediction based on remote sensing data. In *Thirty-First AAAI conference on artificial intelligence*.

[18] van Klompenburg, T., Kassahun, A., &Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, 105709.

[19]Panda, S. S., Ames, D. P., &Panigrahi, S. (2010). Application of vegetation indices for agricultural crop yield prediction using neural network techniques. *Remote Sensing*, *2*(3), 673-696.

[20]Sangeetha &Shruthi G (INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 8, ISSUE 01, JANUARY 2020)

[21]Versteeg, M. N., & Van Keulen, H. (1986). Potential crop production prediction by some simple calculation methods, as compared with computer simulations. *Agricultural systems*, *19*(4), 249-272.

[22]Russello, H. (2018). Convolutional neural networks for crop yield prediction using satellite images. IBM Center for Advanced Studies.

[23]Sahu, S., Chawla, M., &Khare, N. (2017, May). An efficient analysis of crop yield prediction using Hadoop framework based on random forest approach. In *2017 International Conference on Computing, Communication and Automation (ICCCA)* (pp. 53-57). IEEE.

[24] Ramesh, D., &Vardhan, B. V. (2015). Analysis of crop yield prediction using data mining techniques. International Journal of research in engineering and technology, 4(1), 47-473.

[25]Palanivel, K., &Surianarayanan, C. (2019). An Approach for Prediction of Crop Yield Using Machine Learning and Big Data Techniques. *International Journal of Computer Engineering and Technology*, *10*(3), 110-118.

[26]Prasad, A. K., Singh, R. P., Tare, V., &Kafatos, M. (2007). Use of vegetation index and meteorological parameters for the prediction of crop yield in India. International Journal of Remote Sensing, 28(23), 5207-5235.

[27]HardikJoshi , Monika Gawade , Manasvi GanuProf.PriyaPorwal (International Journal of Engineering science and research technology (April, 2018)

[28]Ghadge, R., Kulkarni, J., More, P., Nene, S., &Priya, R. L. (2018). Prediction of crop yield using machine learning. Int. Res. J. Eng. Technol.(IRJET), 5.

[29]Al-Gaadi, K. A., Hassaballa, A. A., Tola, E., Kayad, A. G., Madugundu, R., Alblewi, B., &Assiri, F. (2016). Prediction of potato crop yield using precision agriculture techniques. *PloS one*, *11*(9), e0162219.

[30]Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environmental Research Letters, 13(11), 114003.

[31]Mardanisamani, S., Maleki, F., HosseinzadehKassani, S., Rajapaksa, S., Duddu, H., Wang, M., ...& Vail, S. (2019). Crop lodging prediction from UAV-acquired images of wheat and canola using a DCNN augmented with handcrafted texture features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 0-0).

[32]Kantanantha, N., Serban, N., & Griffin, P. (2010). Yield and price forecasting for stochastic crop decision planning. Journal of agricultural, biological, and environmental statistics, 15(3), 362-380.

[33]Li, M., &Jin, H. Development of a postprocessing system of daily rainfall forecasts for seasonal crop prediction in Australia.

[34]Chaudhary, K., &Kausar, F. PREDICTION OF CROP YIELD USING MACHINE LEARNING.

[35]Feng, X., Fu, T. M., Cao, H., Tian, H., Fan, Q., & Chen, X. (2019). Neural network predictions of pollutant emissions from open burning of crop residues: Application to air quality forecasts in southern China. *Atmospheric Environment*, *204*, 22-31.

[36]Chakrabarti, S., Braswell, R., Malizia, N., Sulla-Menashe, D., Cormier, T., &Friedl, M. (2019, July). In-Season Prediction of Crop Types In The us Great Plains Using Sequence Based Stochastic Models and Deep Learning. In IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium (pp. 5836-5839). IEEE.

[37]Garde, Y. A., Singh, S., Mishra, G. C., & Singh, T. (2012). Weather based pre-harvest forecasting of wheat at Ghazipur (UP). *International Journal of Agricultural Sciences*, *8*(2), 325-328.

[38]Petersen, L. K. (2018). Real-time prediction of crop yields from MODIS relative vegetation health: A continent-wide analysis of Africa. Remote Sensing, 10(11), 1726.

[39]Crop Yield Prediction Using Deep Neural Networks

[40] Praveen, D. (2017). Spatiotemporal analysis of projected impacts of climate change on the major C3 and C4 crop yield under representative concentration pathway 4.5: Insight from the coasts of Tamil Nadu, South India. PloS one, 12(7), e0180706.

[41]Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and electronics in agriculture*, *151*, 61-69.

[42]Mohammadhossein, Hajiyan School of Engineering, University of Guelph,(School of Engineering University of Guelph, May 2012)

[43]Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and electronics in agriculture*, *151*, 61-69.

[44]Veenadhari, S., Misra, B., & Singh, C. D. (2014, January). Machine learning approach for forecasting crop yield based on climatic parameters. In *2014 International Conference on Computer Communication and Informatics* (pp. 1-5). IEEE.

[45]Balakrishnan, N., &Muthukumarasamy, G. (2016). Crop production-ensemble machine learning model for prediction. *International Journal of Computer Science and Software Engineering*, *5*(7), 148.

[46] Johnson, M. D., Hsieh, W. W., Cannon, A. J., Davidson, A., &Bédard, F. (2016). Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. Agricultural and forest meteorology, 218, 74-84.