IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AN LSTM BASED CROP PRICE PREDICTION SYSTEM

[1]Vishal Kasa, [2]Karan Mahajan, [3]Shweta Dhautre, [4]Suresh Kapare

[1,2,3]Student at MIT ADT School of Engineering, [4]Professor at MIT ADT School of Engineering
[1]Department of Computer Science Engineering,
[1]MIT School of Engineering, Pune, India

Abstract: India is an agriculture-based country, with about 70% of households reliant on agriculture, and farmers dedicating their entire lives to the country's economy. They have a huge problem when the crop isn't worth the price and they aren't aware of the crop's marketplace price. Price changes in agricultural commodities have a negative impact on a country's GDP. Farmers are hurt emotionally and financially as their complete hard work for the crop cycle is for nothing. Price prediction has become a highly critical agricultural problem that can only be solved with the data available. The purpose of this study is to forecast crop prices for the next rotation. Price forecasting could aid the agriculture supply chain in making the appropriate decisions to reduce and manage the risk of price variations. This work is based on identifying appropriate data models that aid in achieving high price forecast accuracy and prediction. This paper describes a system that employs data analytics approaches to forecast crop prices. The suggested system will use machine learning algorithms and Sequence modeling to forecast crop prices based on a variety of characteristics such as harvested area, yield, and so on. This gives a farmer an idea of the future price of the crop that he has planted. This model entails crop price forecasting. Crop price forecasting is a difficult task since it involves so many variables.

Index Terms* - Agriculture; Machine learning; *Price Prediction; economic factors; Random Forest Regressor, MLP, CNN, LSTM

I. Introduction

Agriculture is the foundation of all various economies. Agriculture has long been regarded as the primary and most important enterprise performed in various regions. There are numerous methods for increasing and improving agricultural output, quality and revenue. Data mining can also be used to evaluate crop prices. Data mining, in general, is the process of examining data from various angles and synthesising it into valuable knowledge.

Getting an idea of the crop price so as to maximise profits is a significant agricultural dilemma. Every farmer wants to know how much money he will get based on his expectations. Price predictions can be calculated by studying a farmer's previous experience with specific crops. Accurate information regarding agricultural yield history is critical for making decisions about crop price forecasting. As a result, this research presents a method for predicting agricultural prices. Data analytics is the act of examining data collections to derive conclusions about the information they contain, with the use of specialised systems and software becoming more common.

Farmers in India largely produce crops based on fixed crop cycles or traditional methods. Given the current circumstances, many of them are unaware of the potential losses and are unaware of the benefits they receive by cultivating them. Farmers are confronted with these challenges because they are unaware of the economic situation. To reduce these risks, it's critical to choose crops that will yield a fair profit when harvested. While most farmers stick to traditional cropping plans, it's critical to plan crops based on economic conditions or market conditions as well.

Earlier [11] we had conducted a thorough research and review of various crop recommendation systems, crop price prediction systems, etc. which formed the base for the implementation of this study. The scope of this study [11] was to understand different factors that affect the price of crops and the types of systems used.

In this manuscript we are trying to predict the price of a few crops where we have the location and a tentative sowing date. We have compared different approaches one is using machine learning algorithms and second is a hybrid sequential modeling approach.

II. LITERATURE REVIEW

The objective of this manuscript [1] is to build a system which provides efficient and effective price prediction features. The aim is to propose a new framework and develop a system to make some advances towards a more efficient price prediction. This paper includes a decision-making support model that can be helpful for farmers to predict prices. This model includes a portal in which farmers are required to login to their account with the credentials (username and password) which can be their name and mobile number as it is easy for them to remember.

The main goal is to establish the new predictive model based on the Hybrid Association [2] rule-based Decision Tree algorithm (HADT). The applications and techniques of data mining as well as Big Data using agriculture data is considered in this paper. In particular, the farmers are more concerned about estimating how much profit they are about to expect for the chosen crop. As with many other sectors the amount of agriculture data is increasing on a daily source. In this work, agriculture crop price dataset of Virudhunagar District, Tamilnadu, India is considered and for the price prediction model based on data mining decision tree techniques

Back Propagation (BP) neural network forecasting model [3] and Autoregressive Integrated Moving Average (ARIMA) forecasting model of Hainan vegetables price are set up. Based on the above two models, linear combination and nonlinear combination forecasting models of vegetable price is established by linear programming method and BP neural network method.

[4] The system gives detailed forecasts up to the next 12 months. The methodology we use in the system is decision tree regression which is Machine Learning Regression technique. The parameters considered for prediction are: - rainfall, wholesale price index (minimum support price, cultivation cost). Accurate prediction of crop price; plays an important role in crop production management. Such predictions will also support the allied industries for strategizing the logistics of their business.

The aim of this paper is to predict the crop price for the next rotation [5]. This work is based on finding suitable data models that help in achieving high accuracy and generality for price prediction. For solving this problem, different Data Mining techniques were evaluated on different data sets. This work presents a system which uses data analytics techniques in order to predict the price of the crop. The proposed system will apply machine learning algorithms and predict the price of the crop based on multiple factors like Area harvested, Area planted etc. This provides a farmer with an insight of the future price of the crop that he is going to harvest.

The proposed system [6] has been created to assist farmers make better choices with regard to which time is most appropriate amid their wanted time of sowing and the area. The framework predicts the yield and cost of the crop of choice, giving the agriculturist valuable information well some time recently beginning the method of cultivation. Numerous expectation calculations can be utilized for edit surrender and cost expectation such as choice trees, neural systems, SVM etc. The decision tree is prepared on a few Kharif and rabi crops (like paddy, arhar, bajra, grain, etc) giving great accuracy.

The paper mentions data of 15 years on the basis of a daily monthly and annually for major crops of China. A SVR model [7] was designed to predict the wholesale agricultural products price. Fruit was chosen as the research object.

The objective of this paper [8] is to predict crop price for next rotation. It demonstrates techniques to estimate crop price using the current data. The model is implemented using linear regression neural networks and root mean square error is calculated. This work presents a framework that employs information analytics methods in order to anticipate the cost of the trim. The framework is implemented employing a python programming dialect within the jupyter notebook.

Deep Learning, Machine Learning and Visualization are used in the proposed system [9]. The paper describes an android mobile application that predicts the crop price. The Model is made utilizing LSTM RNN for vegetable forecast and ARIMA for price prediction. This proposed framework contains four primary components such as crop prediction, price prediction, visualization and optimization.

The paper [10] endeavours to figure the costs of vegetables from changes within the cost of rough oil. The paper describes data mining techniques such as K-Means, K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN) and Support Vector Machines (SVM) which help to predict crop prices. They studied the location of Coimbatore and showcased the cost of tomato as an illustration and simulated the result utilizing MATLAB.

III. IMPLEMENTATION:

The steps we followed to develop our system are shown below in (Figure).

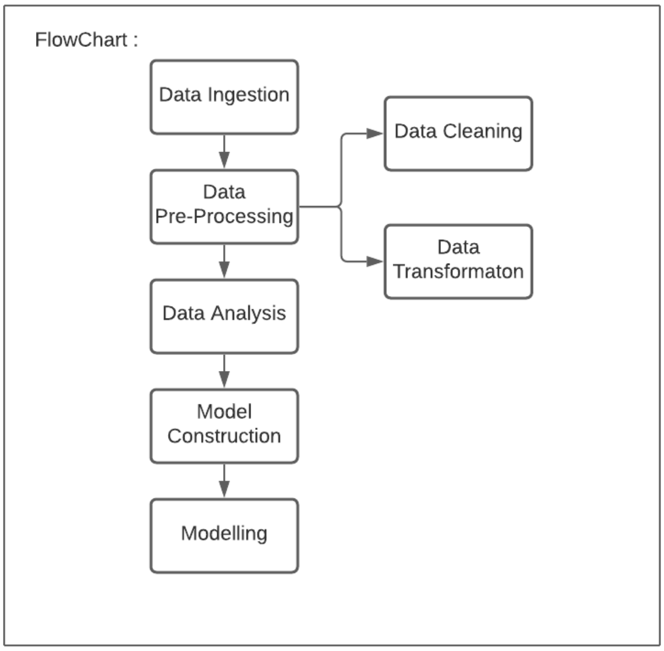


Figure 1: Workflow

3.1 Selection of factors:

We started our research by visiting different websites like data.gov.in, data.world, Enam.gov.in, Rajyasabha.nic.in, agmarknet.gov.in by using keyword such as crop supply, production, crop demand, Farmgate price, etc. and we found a lot of economic factors that affect the price of a crop such as demand, supply, area, production, price inflation, Government stored food grain amount, intangible social factors etc. But most of them can't be modeled and we don't even have corresponding data for them. So, for this research we have taken the factors like Area harvested, yield of crop and time series data of past price of crop and crop season.

3.2 Data Ingestion

Data Ingestion is the movement of data from a variety of sources to a storage medium where it may be accessed, used, and analysed by an organization. Any analytics architecture's backbone is the data ingestion layer. Data consistency and accessibility are essential for downstream reporting and analytics systems. Data can be ingested in a variety of ways, and the design of a data ingestion layer can be based on a variety of models or architectures.

Since the price of crops fluctuates from location to location, we have selected the location of Pune for our study. We have also selected a few crops viz. Moong, Gram, Bajra, Jowar, Rice, Maize and Wheat for our study. We selected two datasets in which one has time series data of crop, area and crop production found on data.gov.in [16] and another has time series data of crop and past prices found on agmarknet.gov.in [15]. We have also considered Kaggle for data but most of the data has no description so that makes the data hard to understand and use.

We have also approached some agriculture universities for data and also gathered some insights from farmers and also noted some points.

3.3 Data Processing

Data preprocessing is a data mining approach for transforming raw data into a format that is both useful and efficient. The information presented here goes through two stages.

3.3.1 Data Cleaning:

It is critical that data be error-free and free of unnecessary information. As a result, the data must be cleaned before proceeding to the next stage. Checking for missing values, duplicate records, and improper formatting, as well as eliminating them, is part of data cleansing. The dataset had some null values, special characters, most of them had different formats, etc. so we had to perform data cleaning operations on the dataset.

3.3.2 Data Transformation:

Data transformation is the mathematical manipulation of datasets; data is changed into relevant formats for data mining. This allows us to have a better understanding of the data by arranging the hundreds of records in a logical sequence. We have two types of dataset and we merged them on the basis of "Crop Name", "Year" and "Month". Then we must perform the Data Cleaning step again. We had also performed normalization and Attribute Selection on the dataset.

3.4 Data Analysis

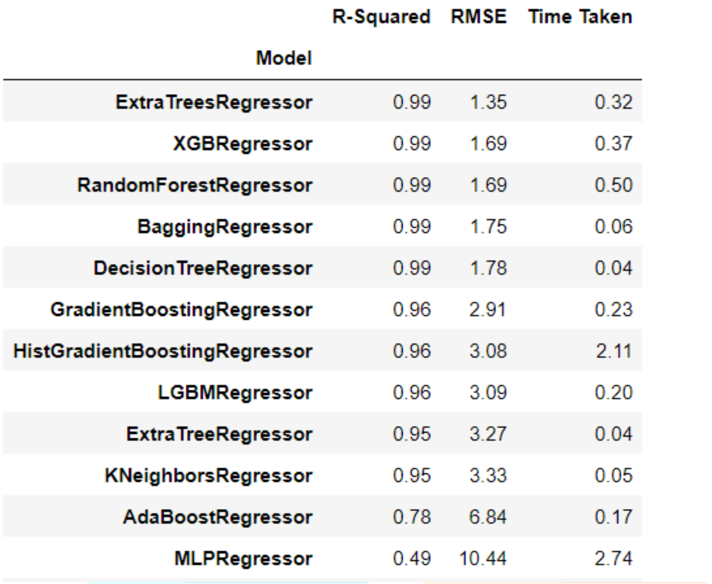
Data analysis is a technique for better understanding datasets using visual features such as scatter plots and bar charts. This helps us to identify data trends and conduct analysis more properly as a result. The analysis always starts with a question of what we are expecting from the dataset so after getting the dataset we approach government websites to evaluate that the data we had collected is correct and unbiased. Now we are sure that the data is perfect for our work.

3.5 Model Training and Testing

The relationship between the components and the goal variable is explained by correlation graphs. A correlation coefficient with a magnitude of 0 implies that there is no association between the selected qualities, while a correlation coefficient with a magnitude of 1 suggests that the selected variables are at best connected. We found out whether this data has any kind of correlation and we removed the parameters that had less correlation.

We had considered the parameters viz. 'Crop Year', 'Crop', 'Area', 'Yield', 'month', 'Modal Price (Rs. /Quintal)'. We have used a library called Lazy Predict [17]**:** This library helps us to fit our dataset with several algorithms like RandomForestRegressor, DecisionTreeRegressor, KNN, MLPRegressor with basic parameter tuning.

We can see different models RMSE and R-Squared in the figure below.



	R-Squared	RMSE	Time Taken
Model			
ExtraTreesRegressor	0.99	1.35	0.32
XGBRegressor	0.99	1.69	0.37
RandomForestRegressor	0.99	1.69	0.50
BaggingRegressor	0.99	1.75	0.06
DecisionTreeRegressor	0.99	1.78	0.04
GradientBoostingRegressor	0.96	2.91	0.23
HistGradientBoostingRegressor	0.96	3.08	2.11
LGBMRegressor	0.96	3.09	0.20
ExtraTreeRegressor	0.95	3.27	0.04
KNeighborsRegressor	0.95	3.33	0.05
AdaBoostRegressor	0.78	6.84	0.17
MLPRegressor	0.49	10.44	2.74

Figure 2: RMSE and R-Squared for Models

3.6 Final model selection (Modelling):

The process of data modelling entails building a data model for the data that will be kept in the database. Modelling entails training a Machine Learning Algorithm to predict labels from features, adjusting it for business requirements, and testing it on holdout data. The result of modelling is a trained model that can be used to infer new data points and make predictions.

Modelling is separate from the other steps in the Machine Learning process and uses standardised inputs, allowing us to change the prediction issue without having to rewrite all of our code. We can produce new label timings, create associated features, and insert them into the model if the business requirements alter. Models are built and then tested for accuracy using a variety of methods.

Crop price prediction is a difficult undertaking, especially for newcomers with no prior experience. When confronted with such circumstances, most people attempt to forecast crop values based on prior data, either using a machine learning, neural network or manually.

We also tried to make a model using machine learning algorithms but the results were not accurate and hence we concluded that machine learning algorithms are not useful in this case as most models are overfitting as you can see in the table below.

Table 1: Accuracy and Error

Algorithm	Accuracy		Error	
	Train Accuracy	Test Accuracy	RMSE	MAE
Random Forest Regressor	97.68	92.42	17884.6	7030.9
MLP	95.5	82.4	21.875	18.55
Extra Trees Regressor	99.4	90.8	169.58	106.64

So now, to tackle this problem, we think that it has to be solved by using hybrid models instead of typical Machine Learning Algorithms.

We have used Time series analysis or simply sequential modelling for Crop price Prediction by taking the last 20 years of price data and we have used Stacked LSTM model for Analysis and prediction

3.6.1 Stacked LSTM Model

The Stacked LSTM network as shown in Figure 3, provides good precision, as it is computationally intensive. The output of the neural network is a thick layer with stacked LSTM layers. The model becomes more sophisticated as a result of the stacked LSTM network, allowing it to read and create complicated features.

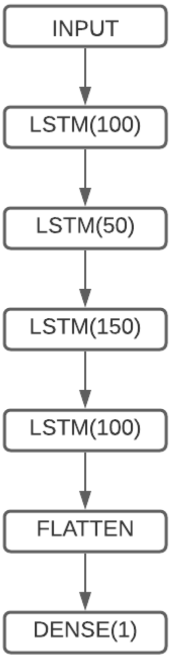


Figure 3: Stacked LSTM Architecture

In this Stacked LSTM, the size of the sliding window is very important. The size of the window changes a lot when it comes to predicting the price. All neural networks can be tested on sliding windows of various widths. This aids in finding the optimal forecast based on the evaluation metrics through comparison. This neural network is trained with Adam optimizer for 500 epochs, learning rate of 0.0001 with 256 batch size.

3.7 Model validation

Model validation is the process of comparing model outputs to independent real-world observations (systematically) to determine whether they are quantitatively and qualitatively consistent with reality.

For model validation we have used Accuracy, RMSE, MAE and MSE.

3.7.1 Root Mean Square Error:

The root mean square error (RMSE) is the residuals' standard deviation (prediction errors). The residuals are a measure of how far away the data points are from the regression line, and the RMSE is a measure of how spread out these residuals are. In other words, it indicates how tightly the data is clustered around the line of best fit.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{pred,i} - y_i)^2}{n}} \qquad (1)$$

3.7.2 MAE –

The Mean Absolute Error(MAE) is the average of all absolute errors.

$$\text{MAE} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\text{test set}} |\underbrace{y_i}_{\text{predicted vaue}} - \underbrace{\hat{y}_i}_{\text{actual value}}| \qquad (2)$$

The fundamental goal is to compare and generate better predictions on time series data using a Hybrid conventional stacked LSTM model. The model's predictions are significantly influenced by the varying sizes of sliding windows. The sliding window of size n-1, for example, employs t1, t2, t3,..., (tn-1) time steps to estimate crop value at tn and tn+1 values t2, t3,..., tn, and so on.

For forecasting, this unique approach compares neural networks with varying window sizes. The best window size is chosen to be employed in the procedure and obtain the best prediction.

The RMSE is the most important assessment parameter for evaluating the performance of a model and deciding on a certain window size to obtain the final crop price prediction. The model with the smallest RMSE value is completed with a given window size.

IV. RESULTS AND CONCLUSION

4.1 Results:

Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE) are the evaluation criteria for each window size (n). The conclusions of the model are summarised in the observation tables below, which differ by crops. In Figure 4, the Maize crop data is used.

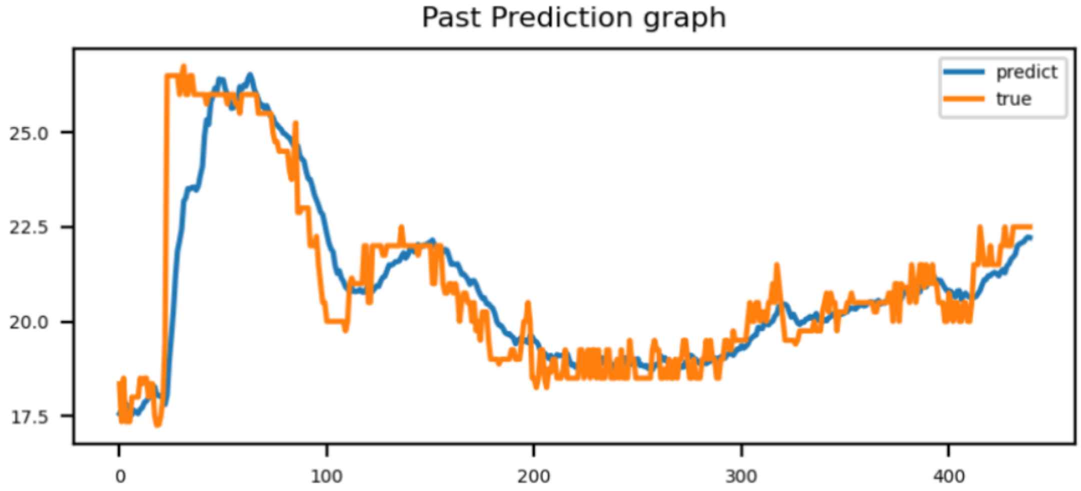


Figure 4: Actual Crop Price vs Predicted Crop Price for Maize

Table 2: Observation with Different Window sizes

Crop Name	Window Size	RMSE	MSE	MAE
Maize	30	1.322	1.747	0.896
	45	1.09	1.188	0.643
	60	1.184	1.403	0.696
Wheat	30	5.281	27.890	3.66
	45	5.122	26.232	3.393
	60	5.335	28.463	3.862
Bajra	30	1.2499	2.247	1.060
	45	1.551	2.405	1.102
	60	1.608	2.584	1.173
Rice	30	15.847	251.143	10.486
	45	15.931	253.806	10.681
	60	16.245	263.903	10.961
Jowar	30	3.334	11.119	3.023
	45	3.186	10.152	2.882
	60	4.554	10.152	2.882
Moong	30	1.954	3.818	1.395
	45	2.067	4.274	1.480
	60	2.132	4.546	1.540
Gram	30	2.484	6.173	1.699
	45	2.558	6.543	1.735
	60	2.641	6.973	1.815

As expected, changing the window size (n) has a considerable impact on crop forecast. We have manually tuned the system based on the best window size for better predictions. For this model to work well the dataset needs to be updated in regular time intervals because then the system will predict well as we are doing time series analysis using the latest number of window sizes of previous data.

4.2 Conclusion and Future Scope

In this manuscript we have developed a system that predicts the price for the crops at a particular location on a particular day.

Table 3: Final Observation

Crop Name	Best Window Size	RMSE
Maize	45	1.09
Wheat	45	5.122
Bajra	30	1.249
Rice	30	15.84
Jowar	45	3.186
Moong	30	1.954
Gram	30	2.484

By Comparing Table 1 and 2, It is clear that Hybrid Models are more exact and efficient for crop price prediction, particularly short-term prediction. To achieve more precision, it is critical to pay attention to numerous hyperparameters and model tweaking with model architecture.

As seen in this experiment, determining the window size, or the number of time steps to look at, is an important parameter for prediction and from Table 3 we can say that less window size gives better results as most recent data will mostly fit well. As a result, while hybrid models perform better, they require hyperparameter adjustment, which varies depending on the crop price variations.

Finally, all of the strategies discussed in this paper are merely tools to assist farmers in making decisions. Because the forecasting of price cannot be forecast with 100 percent accuracy, a great deal of effort must be spent into optimising existing algorithms and developing new ones that can attempt to close the gap between calculations and reality.

We would focus on building new algorithms employing the latest technology and tweaking them to produce the best results possible to further decrease the gap between on-paper calculations and real numbers. We'd like to add multi-cropping from global marketplaces in the next analysis instead of simply Indian market prices.

To make it accessible to the users, we have also created a simple frontend and API endpoints. Because of the API endpoints, this system can also be complemented with a crop recommendation system.
With this, we plan to increase the number of crops and the locations.

Abbreviations

Table 4

Abbreviations	
MLP	Multilayer Perceptron
HADT	Hybrid Association rule-based Decision Tree algorithm
BP	Back Propagation
ARIMA	Autoregressive Integrated Moving Average
SVM	Support Vector Machine
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
KNN	K-Nearest Neighbor
RMSE	Root mean square error
MAE	Mean Absolute Error
ANN	Artificial Neural Networks
MSE	Mean Square Error

Acknowledgment

We'd like to thank Mrs. Kanchan Bhonde of TechMahindra Maker's Lab, as well as Prof. Jyoti Malhotra of MIT ADT University, for allowing us to work on this project. We'd also like to thank Mr. Saket Apte of TechMahindra Maker's Lab for their unwavering support and advice.

REFERENCES

[1] Aman Vohra, Nitin Pandey, S.K. Khatri ; Decision Making Support System for Prediction of Prices in Agricultural Commodity

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

[3] Lu Ye, Xiaoli Qin, Yuping Li, Yanqun Liu and Weihong Liang; Vegetables Price Forecasting in Hainan Province Based on Linear and Nonlinear Combination Model

[4] Rohith R, Vishnu R, Kishore A, Deeban Chakkarawarthi; Crop Price Prediction and Forecasting System using Supervised Machine Learning Algorithms

[5] Pandit Samuel,B.Sahithi , T.Saheli , D.Ramanika , N.Anil Kumar; Crop Price Prediction System using Machine learning Algorithms

[6] Sadiq A Mulla, Dr.S.A.Quadri; Crop-yield and Price Forecasting using Machine Learning

[7] Wang Shengwei, Li Yanni, Zhuang Jiayu, and Liu Jiajia; Wang Shengwei, Li Yanni, Zhuang Jiayu, and Liu Jiajia

[8] Gangasagar HL, Jovin Dsouza, Bhagyashree B Yargal, Arun Kumar SV, Anuradha Badage; Crop Price Prediction Using Machine Learning Algorithms

[9] Thayakaran Selvanayagam, Suganya S, Puvipavan Palendrarajah, Mithun Paresith Manogarathash, Anjalie Gamage, Dharshana Kasthurirathna; Agro-Genius: Crop Prediction using Machine Learning

[10] Manpreet Kaur, Heena Gulati, Harish Kundra; Data Mining in Agriculture on Crop Price Prediction: Techniques and Applications

[11] Vishal Kasa, Karan Mahajan, Shweta Dhautre, Suresh Kapare; Crop Selection Based on Economic Factors: A Review; irjet.net volume 8 issue 5.

[12] https://data.world/

[13] https://enam.gov.in/web/

[14] https://rajyasabha.nic.in/

[15] https://agmarknet.gov.in/

[16] https://data.gov.in/

[17] https://pypi.org/project/lazypredict/

[18] https://www.kaggle.com/