Agricultural Data Analysis using Machine Learning Algorithms

Appa Rao Bobbili[1], Dr. Sreedevi M[2]

[1]Student, [2]Professor
Amrita Sai Institute of Science and Technology
Autonomous NAAC with A Grade, Andhra Pradesh, India,

***Abstract*: Agriculture is undoubtedly the largest livelihood provider in India and also contributes a significant figure to the economy of our Country. The technological factors affecting the crop production includes practices used and also managerial decisions. So, predicting the crop yield prior to its harvest would help farmers to take appropriate steps. We attempt to resolve theissue by building a user-friendly prediction system. The results of the prediction are suggested to the farmer such that suitable changes can be made in order to improve the produce. There are different techniques or algorithms which help to predict crop yield. By analyzing all the parameters like location, soil nutrients, pH value, rainfall, moisture a potential solution can be obtained to overcome the situation faced by farmers. This paper focuses on the analysis of the agriculture data and finding optimal yield to provide an insight before the actual crop production using data mining techniques and Machine Learning algorithms.**

***Keywords*: Yield, Random Forest regress or, Decision Tree regress or, GDP, Digitalisation.**

I. INTRODUCTION

Today, India is one of the leading producers across the world in the agriculture sector[1]. Agriculture is the broadest economic sector and plays an outstanding role in the socio-economic part of India. Agriculture is an eccentric business crop production which is influenced by many climate and economic factors. Andhra Pradesh, basically being an agro-Based economy contributes more than 29% of the GDP as against 17% in the country's GDP. Periodical advice to the farmers either in terms of improved agricultural strategies or advancements in factors affecting the production of crops may strengthen the state in the agriculture sector. Yield prediction is one among the agricultural advancements. Due to these kinds of innovations agriculture is driving the interest of modern man. In the past farmers used to predict their yield from previous experiences[2]. Digitalisation in farming gives awareness about the cultivation of the crops at the right time and at the right place even to young farmers. These kinds of advancements need the use of data analytics. This is one such system that can be used to address yield prediction. The main objectives are:

1) To analyse different parameters (soil nutrients, rainfall, area etc)
2) To use machine learning techniques to predict crop yield.
3) To provide an easy to use User Interface

II. HOW DATA MINING IS USED IN AGRICULTURE SECTOR

Data mining techniques are used in performing several activities in the agricultural sector such as pest identification, detection and classification and prediction of crop diseases. It can also be used in yield prediction, input management (planning of irrigation and pesticides), fertilizer suggestion and predicting soil. In a world full of data , data mining is the computational process for discovering new patterns[3]. Data mining techniques provide a major advantage in agriculture for detection and prediction for optimizing the pesticides. Techniques for agriculture related activities provide a lot of information. The yield of agriculture primarily depends on diseases, pests, weather conditions, planning of various crops for the harvest productivity are the results.

Crop production for reliable and timely requirements for various decisions for agriculture marketing. Predictions are very useful for agriculture data. For instance, by applying data mining techniques, the government can fully benefit from data about farmers' buying patterns and also to achieve a superior understanding of their land to achieve more profit on the farmer's part.

Data mining techniques followed in two ways[4]:

1) Descriptive data mining.
2) Predictive data mining.

Descriptive data mining tasks characterize the final properties of the info within the database while predictive data mining is employed to predict the direct values supported patterns determined from known results. Prediction involves using some variables or fields within the database to predict unknown or future values of other variables of interest.

As far as data mining techniques are concerned, in most cases predictive data mining approaches are employed. Predictive data mining techniques are employed to predict future crop, forecasting, pesticides and fertilizers to be used, revenue to be generated and so on. These techniques are used for pre-harvest forecasting for the agriculture field and are able to provide a lot of data on agricultural-related activities. Data of agriculture in data mining can be presented in the form of datasets.

III. PROPOSED SYSTEM

The main objectives of proposed work is to analyse the agricultural parameters using data mining algorithms and predict the yield.In our proposed work, agriculture data has been collected from various sources which include:

Dataset in agricultural sector[5] ,Crop wise agriculture data:[6], Soil data of different districts:[7]

In this proposed system , we mainly focussed on Andhra Pradesh State in India. As the state has two major rivers flowing , it has a

diversity in factors useful for agriculture at district level. Periodical data about the crop , soil and water a particular region is the major focus of this study.The final dataset has been tabulated as in table-1:

S. No	Feature	Description
1	Year	The year in which the crop will be cultivated.Generally, the upcoming year
2	Season	One among Kharif,Rabi and Whole Year.
3	Crop	Name of the crop
4	District	Name of the district
5	pH Level	This describes the nature of the soil
6	Nitrogen	Amount of nitrogen present
7	Potassium	Amount of potassium present
8	Phosphorus	Amount of phosphorus present
9	Rainfall	Expected rainfall in millimeters
10	Area	Area of field in hectares

Table-1: Description of Input data

The below diagram depicts the system architecture of our proposed system. Our whole system can be divided into 2 modules as a whole i.e., one model predicts the optimal yield and the other model analyses the patterns in the dataset. The operation of these models as a whole is specified clearly in the below diagram.

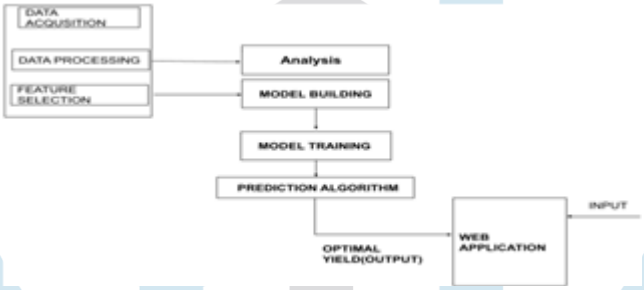

Fig.1 The blueprint of the proposed system

IV. METHODS

In the implementation of this yield prediction system Regression Analysis is used.Regression Analysis is considered as one of the oldest,and widely used multivariate analysis techniques in the social sciences. Unlike others regression stands as an example of dependence analysis in which the variables are treated asymmetrically. In regression analysis, the object is to obtain a prediction of one variable, based on given the values of the others[8].Random Forest and Decision Tree algorithms are generally used in classification problems but these can also be used in regression problems as well.

A. Decision Tree Regression

The Decision Tree algorithm comes under supervised machine learning techniques. A decision tree arrives at an outcome by asking a series of questions to the input data , each question narrows down the possible outcomes until the model gets enough potential to make a unique prediction[9]. The order of the questions as well as their contents are being determined by the model.All the questions that are raised have their answer as either true or false.

B. Random Forest Regression

Random Forest algorithm comes under the family of ensemble algorithms. This is also a supervised learning algorithm. This can be implemented in classification and regression as well. Random forest algorithm basically works on Decision Tree principle by constructing a number of decision trees having different sets of hyper-parameters for tuning and training on different subsets of data[10].

V. EXPERIMENTAL RESULT

A. Decision Tree Regression

Decision Tree algorithm on applying on the dataset resulted 100% on data and 82%(approx.) on test data.Fig-2 shows the accuracy of Decision tree algorithm on data:

DecisionTree Regression:

```
[ ]  from sklearn.tree import DecisionTreeRegressor
     dt_obj =DecisionTreeRegressor(random_state=1)
     dt_obj.fit(X_train,y_train)
     print('Train Score DT:',dt_obj.score(X_train,y_train))
     print('Test Score DT:',dt_obj.score(X_test,y_test))

Train Score DT: 1.0
Test Score DT: 0.8162150580308982
```

Fig-2: Result of Decision Tree algorithm

B. Random Forest Regression

Random Forest algorithm on applying on the dataset resulted 98% on data and 90%(approx.) on test data.Fig-3 shows the accuracy of Decision tree algorithm on data:

RandomForest Regression:

```
[ ]  from sklearn.ensemble import RandomForestRegressor
     rf_obj =RandomForestRegressor(n_estimators = 10,random_state=0)
     rf_obj.fit(X_train,y_train)
     print('Train score RF:',rf_obj.score(X_train,y_train))
     print('Test score RF:',rf_obj.score(X_test,y_test))

Train score RF: 0.9859482658364138
Test score RF: 0.8994486718857367
```

Fig-3: Result of Random Forest algorithm

	index	production var	price var
0	Barley	-129.610556	51.277056
1	Jute	-43011.146430	125.248918
2	Niger seed	-144.009443	249.534632
3	Safflower	-1031.251935	122.445887
4	Sunflower	-4511.467222	223.906926

Fig-4:Crop with reduced production and increased price

	index	production var	price var
0	Arhar/Tur	5237.572915	260.551948
1	Groundnut	1758.235163	199.339827
2	Jowar	3648.323679	234.956710
3	Jute	-43011.146430	125.248918
4	Moong	941.797058	308.993506
5	Niger seed	-144.009443	249.534632
6	Safflower	-1031.251935	122.445887
7	Sesamum	1494.291172	279.404762
8	Sunflower	-4511.467222	223.906926
9	Urad	2670.533797	284.469697

Fig-5: Crops with lower increase in production but are increasing in price

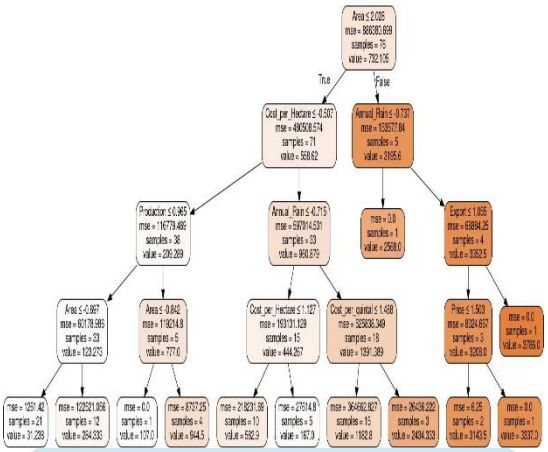

Fig-6: Decision Tree for Suicide Prediction

S.No.	File name	Name of DB in program	Details or Comments
1	apy	Crop Production Statistics(crop prod)	Main DB that has crop production info from 2000 to 2014.For different district of each State it includes, what are various crops produced, their area of production and total produc- tion,and what type of crop is it.(Kharif,Rabi)
2	Crops price	Crop prices(crop price)	Prices of some crops year wise change till 2013.For different commodities it has data for its price in rupees per quintal.
3	area cult	Crop cultivation area(crop cult)	Area of land a crop is produced on year by year for major crops from 2000 to 2009.
4	culti cost	Crop cultivation cost(culti cost)	State wise cost of cultivation of crops per hectare and per quintal.Three variant of cost are there(actual paid out cost plus imputed value of family labour(A2+FL),comprehensive cost including imputed rent and interest on owned land and capital(C2) and cost per quintal)
5	Mean Temperatures	Mean Temperatures(temperature)	Data of mean temperature from 2000-2012 for whole year and over interval of two months. This is used to determine effect of temperature on various crops
6	rainfall cleaned	Rainfall Statistics(rainfall)	State wise rainfall statics from year 2000-2015 annually and monthly in millimeter per square meter(area)
7	Avg annual Growth Rate Major Crops	Crops Growth rate(growth)	Growth rate of various crops from 1997 to 2012 over a interval of five years.Growth rate represent increase in size, mass or number of crops over a period of time. It is used in anal- ysis of preference of one crop over other.
8	suicides 10 14	Suicide Statics(suicides)	Data of no. of total cases of suicides in vari- ous states from year 2010-2014.While analysis it is taken into account to predict responsible factors.
9	IndiaExport	Exports(exports)	Data regarding the amount of export of vari- ous materials and its price from 2003 to 2015.
	File name	**Name of DB in program**	**Details or Comments**
10	data set	Data(data)	Combined data of various states from 2000 to 2014. This is combined representation of all data in one table. All other tables mentioned above are combine using the common features and merged.

VI. CONCLUSION

Both Decision tree regression and Random Forest regression techniques are implemented on the input data to assess the best performance yielding method. These methods are compared using performance metrics. According to the analyses of metrics both the algorithms work well , but Random Forest regression gives a better accuracy score on test data than Decision tree regression. The proposed work can also be extended to analyse the climatic conditions and other factors for the crop and to increase the crop production.

VII. REFERENCES

[1] https://www.investopedia.com/articles/investing/100615/4-countries-produce-most-food.asp
[2] http://www.fao.org/fileadmin/templates/ess/documents/meetings_and_workshops/GS_SAC_2013/Improving_methods_for_crops_estimates/Crop_Yield_Forec asting_Methods_and_Early_Warning_Systems_Lit_review.pdf
[3] https://ijaers.com/uploads/issue_files/3%20IJAERS-MAY-2017-60-Different%20Types%20of%20Data%20Mining%20Techniques.pdf
[4] https://docs.oracle.com/cd/B12037_01/datamine.101/b10698/4descrip.htm
[5] https://data.gov.in/
[6] http://www.apagrisnet.gov.in/2018/weekly/October/weekly_report_(Rabi)_06_21-11-18.pdf
[7] http://dataverse.icrisat.org/file.xhtml?fileId=1185&version=RELEASED__version=.3
[8] https://www.sciencedirect.com/topics/medicine-and-dentistry/regression-analysis
[9] https://gdcoder.com/decision-tree-regressor-explained-in-depth/
[10] https://www.analyticsvidhya.com/blog/2020/12/lets-open-the-black-box-of-random-forests/
[11] https://journalofbigdata.springeropen.com/articles/10.1186/s40537-017-0077-4