International Journal of Modern Agriculture, Volume 9, No.4, 2020
ISSN: 2305 - 7246

Analysis and High Accuracy Prediction of Coconut Crop Yield Production Based on Principle Component Analysis with Machine learning Models

Sasmita Kumari Nayak [1]

[1] Computer Science and Engineering, Centurion University of Technology and Management, Odisha, India

Corresponding author*:*

Sasmita Kumari Nayak
Computer Science and Engineering, Centurion University of Technology and Management, Odisha, India

Abstract

In analysis of crop yield production, an emerging research field is the Data Mining. Crop yield is a highly complex trait in agriculture. Basically, data mining is a method for analysing data from varied viewpoints and summarized the same into important information. For crop yield prediction, Machine learning is also a significant decision support tool that includes supporting decisions upon, which crops to cultivate and what actions should be taken while growing season of the yields. The results of the prediction will be made available to the farmer. For the research of crop yield prediction, various machine learning models have been employed. In this article, the prediction has been done for coconut crop. This paper applied four different supervised techniques like Random Forest, Gradient Boosting, Support Vector Machine, and Decision Tree Regression techniques to get the accuracy of coconut crop yield. This study proposed and implemented all the models to predict coconut crop by using the previous data. The outcomes of simulation illustrates that the proposed work efficiently for prediction of coconut crop.

Keywords: Coconut Crop Yield Prediction, Data Mining, PCA, Machine Learning Models.

Introduction

In global food production, the much attention has given to Crop yield prediction (Khaki et al., 2019). Farmers as well as Growers have got benefitted from prediction of yields for making financial and management decisions (Khaki et al., 2019; Horie et al., 1992). Still, crop yield prediction is highly complex because of varied complex factors like temperature, weather condition, soil condition and so on. Each crop has various attributes or parameters to get predictions with the help of various models and these models could be examined by doing many studies (Medar et al., 2019). Several Machine Learning (ML) models shall apply for getting the maximal production of coconut crop, which is the main objective of this article. Coconut crop production depends upon weather conditions (cloud, temperature, rainfall and humidity) as well as geographical conditions (depth areas, hill areas, river ground) (Medar et al., 2019).

Data mining is an analytical tool used by users to analyze data and to show the relationships among them. Data mining covers research for feature selection, feature extraction and prediction of problems of agriculture, Healthcare, financial data analysis, retail industry. From the various studies of agriculture field, I got multiple ways of increasing the economic growth of India. In analysis of crop yield production, an emerging research field is used named as Data Mining. The issue of prediction of crop has a great importance in agriculture field. All farmers have only one interest i.e. the expectation of high quantity of crop production. Earlier, the crop prediction can be done depending on the experience of farmers with specific crop and field. And this is a major issue, which can be solvable with the help of available data and Data mining techniques. Basically, data mining is a method for analysing data from varied viewpoints and summarized the same into important information.

In machine learning, supervised learning is a method to deal with all fields issues. There are various techniques that can be used for predicting the crop like Artificial Neural Network (ANN) (Medar et al., 2019; Siti et al., 2014; Gour et al., 2016; Karan et al., 2016; Nishit et al., 2017), multiple linear regressions(Medar et al., 2019; D Ramesh et al., 2015), Decision Tree Algorithms (DST) (Medar et al., 2019; Gour et al., 2016; Karan et al.,

2016), Regression analysis (Medar et al., 2019; Raorane et al., 2015; Gour et al., 2016; Karan et al., 2016), Bayesian Belief Network, Clustering (Medar et al., 2019; Karan et al., 2016), (Medar et al., 2019; Anshal et al., 2015; Karan et al., 2016), Support Vector Machine (SVM) (Medar et al., 2019; Nishit et al., 2017).

This study proposed and implemented all the models to predict coconut crop by using the previous data. The outcomes of simulation illustrates that the proposed work efficiently for prediction of coconut crop.

The organization of the paper is as follows: the related work of crop yield production and proposed methodology is presented in section II and III respectively. Section IV deals with the experimental outcomes followed by the conclusion and future work in section V.

Literature Survey

A huge number of crop yield prediction works are completed by different researchers with different models and found their accuracies by comparing the several ML methods in the prediction of leaf disease. A few of them are outlined in Table I.

Table I: Summary of crop yield prediction by using ML models.

Authors	**Applications**	**Algorithms**	**Remarks**
Siti Khairunniza-Bejo, Samihah Mustaffha, Wan Ishak Wan Ismail (Medar et al., 2019; Siti et al., 2014)	Providing results to some problems of farmers for finding good yield	ANN	Time consuming process
Anshal Savla, Himtanaya Bhadada, Vatsa Joshi, Parul Dhawan (Medar et al., 2019; Anshal et al., 2015)	Based on parameters, understand and analyze crop yield rate for zones	Classification, Clustering, Normalization	only provides framework
Raorane A.A, Dr. Kulkarni R.V (Medar et al., 2019; Raorane et al., 2015)	Rain fall estimation and reason investigation to get lower yield	Regression analysis	Not specified any specific method
B Vishnu Vardhan, D Ramesh	Analyze and verify the existing data based on multiple linear regression method	Multiple Linear Regressions	Less accuracy
Subhadra Mishra, Debahuti Mishra, Gour Hari Santra (Medar et al., 2019; Gour et al., 2016)	Forecast and increase crop yield rate	ANN, Regression analysis Decision Tree	Not specified any clear method
Karan deep Kauri (Medar et al., 2019; Karan et al., 2016)	Increasing the farming sector in the countries	ANN, Bayesian Belief Network, Decision Tree, Clustering, Regression analysis.	Less accuracy

Nishit Jain, Amit Kumar, Sahil Garud, Vishal Pradhan, Prajakta Kulkarni (Medar et al., 2019; Nishit et al., 2017)	Predict crop sequences and maximize yield rate and make benefits to the farmers. Also predict crop diseases, study crop simulations, different irrigation patterns.	ANN, SVM.	Exact accuracy is not specified.
Elavarasan et al. (Klompenburg et al., 2020; Elavarasan et al., 2018).	A survey of crop yield prediction based on climatic parameters.	ML models	looking broad to get more attributes that represent crop yield
Liakos et al., (Klompenburg et al., 2020; Liakos et al., 2018).	A review on applications of ML in the agriculture.	ML models	Analysis on soil, water, livestock and crop management
Li, Lecourt, Bishop (Klompenburg et al., 2020; Li et al., 2018).	a review to determine the ripeness of fruits	ML models	decide the optimal yield prediction and harvest time
Beulah (Klompenburg et al., 2020; Beulah, 2019).	a survey on prediction of crop yield	Data mining techniques	solved crop yield prediction based on data mining techniques

Proposed Methodology

This research had been done using several Machine Learning algorithms, namely, SVM (Sasmita et al., 2020; Sasmita Kumarai Nayak et al., 2020; Sripada et al., 2020), Decision Tree (Sasmita et al., 2020; Sripada et al., 2020), and Random Forest (Sasmita et al., 2020; Tapas et al., 2020; Sripada et al., 2020). ML techniques have been traditionally applied to large, highly dimensional databases. Machine learning (ML) is a subset of computer science, whereby a computer algorithm learns from prior experience. The steps of machine learning model as shown in Figure 1 (Sasmita et al., 2020). The most essential part of machine learning model is the collection and pre-processing of data. This model has been applied to clean, normalize and pre-process the collected data called as crop yield dataset.

It is a time consuming process but also an essential for the model is need to understand the process of collecting, storing, transforming, reporting the data.

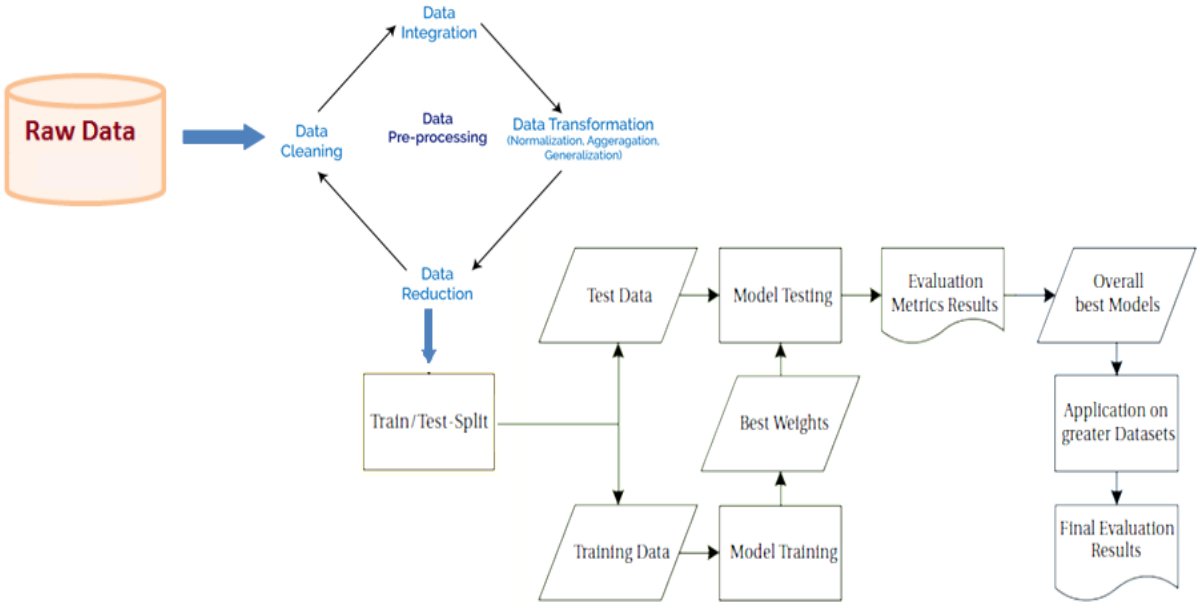

Figure 1: Steps of Machine Learning Model (Sasmita et al., 2020).

Dataset collection

In this article, we have collected the crop production in India data from the website https://www.kaggle.com. It has around 15 years (from 2000 to 2015) of different crop yield productions from different states and districts. In total, the raw data consists of 7 columns or attributes and 246091 numbers of instances. Table II shows the description of crop yield data.

Table II: Description of Crop Production in India Data

Attribute Name	Type of Attribute	Description of Attributes
State_Name	String	Name of the State
District_Name	String	Name of the District
Crop_Year	Numerical	Year of the crop production
Season	String	Current season of the crop production
Crop	String	Type of crop
Area	Numerical	Area of the agricultural field
Production	Numerical	Total crop production

Data Pre-processing

The first step of machine learning model is the data pre-processing. In this case, we will clean, transform and normalise the data. Data pre-processing is required to get the better accuracy for coconut crop prediction. In the data cleaning phase, remove the missing vales from raw data. In this study, no transformation phase is required. After pre-processing of data, the number of instances will be 242361.

Principal component analysis (**PCA)**

It is a statistical procedure. It is allowed to provide an outline of information content in huge data tables using a small set of "summary indices", which will be make easier visualization and analysis. Due to this retentive trends and patterns, the simplification of high-dimensional data complexity is referred to as Principal component analysis (PCA). This high-dimensional data is converted into less number of dimensions that serve as summarised features. PCA reduces various challenges faced by High-dimensional data like increasing of error rate and computing expense, because of multiple test correction during testing of every feature to get the result (Lever et al., 2017).

It is a very flexible tool. It is allowed to analyse the datasets which contains, like, imprecise measurements, categorical data, missing values and multicollinearity. The main objective is to extract the required information from the data and express it as a group of summary indices, referred as, principal components.

Proposed Machine Learning (ML) Models

In this paper, we have given more emphasis to machine learning model i.e. conventional machine learning models. This model is of two kinds of learning models, such as, Supervised and Unsupervised ML model. In supervised learning the predicted value of an attribute or attributes is provided to train the data and it is utilised for solving the classification and regression problems. Whereas in unsupervised learning the predicted variable is not assigned (Sasmita et al., 2020).

In this article, we have considered the following machine learning models to predict the coconut crop. They are Decision Tree (Sasmita et al., 2020; Sripada et al., 2020), Random Forest (RF) (Sasmita et al., 2020; Tapas et al., 2020; Sripada et al., 2020), Support vector Machine (SVM) (Sasmita et al., 2020; Sasmita Kumarai Nayak et al., 2020; Sripada et al., 2020), and Gradient boosting regression models. These four machine learning models are supervised machine learning models. We have implemented the above four regression models for experimentally and compared with each other to find the highest accuracy.

Decision Trees (DST)

Decision tree (DST) is a tree structure, which works on the principle of conditions. It is a robust and effective model for predictive analysis. DST has terminal nodes, branches, and internal nodes. Terminal nodes i.e. leaves are representing the target variable; branches are representing the possible values of the attributes i.e. conclusion of the test dataset and each "test" attributes are represented by internal nodes. It is a supervised model, which is used for both classifications as well as regression. Hence, the name is "CART" i.e. Classification And Regression Tree. Because of reliability and stability, always preferred the Tree algorithms

Random Forest (RF)

Random forest (RF) is a supervised machine learning model based on ensemble learning. Ensemble learning is a more powerful prediction model. The combination of same model several times or different types of model is called as an ensemble learning model. RF model combining multiple model of the same type i.e. multiple decision trees, which results a forest of trees. Therefore, it is named as "Random Forest". This model could also be a "CART" model. This study follows the regression model. The main steps of RF model are as follows:

i) From the dataset, choose N number of records randomly.

ii) Create a decision tree of these records.

iii) Pick the number of trees for the model and repeat the steps i and ii.

iv) For a new record, predict the target value for every tree in the forest. The final value can be calculated by taking the average of all the values predicted by all the trees in forest. Then take the average of all predicted values of trees in forest, which represents the final value of the model.

Support vector Machine (SVM)

Regression problems involve the task of approximating a mapping function from input variables to a continuous output variable. The approach of using SVMs to solve regression problems is called Support Vector Regression (SVR). This regression method maintains all the main features, which characterize the model (maximal margin). SVR supports both linear and non-linear regressions. This model is similar to the principle of SVM for classification. From a few exceptions, this model tries for getting the curve of given data points. However, since it's a regression model, which uses the curve to get the match among the vector and position of the curve, rather to using the curve as a decision boundary. Support Vectors helps to determine and represent the nearest match among the function and data points. But, the objectives are same: minimizing the error, individualizing the hyperplane that maximizing the margin.

Gradient Boosting Regression (GBR)

It is an ensemble decision tree regression model. GBR computes the difference between the actual target value and the current prediction, referred to as residual. After that, this model trains a weak model, which maps features to that residual. This residual is added to the input of existing model, which nudges the model into correct target. Repeat this step to improve the prediction. In every step, a new tree is trained with respect to negative gradient of the loss function that similar to the residual error. The main steps of GBR model are as follows:

i) Choose a weak learner

ii) Apply an additive model

iii) Defining the loss function

iv) Minimizing the loss function

Result

This experiment is conducted on crop yield production data of India from 1998 to 2015. The complete data is not available for the year 2015. Hence we can consider here that up to 2014 the dataset is available. In this dataset, contains 246091 numbers of instances or records with 7 attributes of each. In this experiment, removed the null value data and chosen only 242361 instances for prediction. Hence, we have applied all the steps of data pre-processing through the implementations of ML models. At first picked the highest production of crops i.e. Coconut, from the various crops. All these ML models are behaved as regression. Regression accuracy is the level of perfectly analysing the instances by the regression model, which provides the performance measure of the regression model. At the end, results are compared among all the ML models, such as, GBF, Decision Tree, RF and SVM for the prediction of coconut crop yield production. Analysis of coconut crop yield production is shown in Figure 2, 3 and 4.

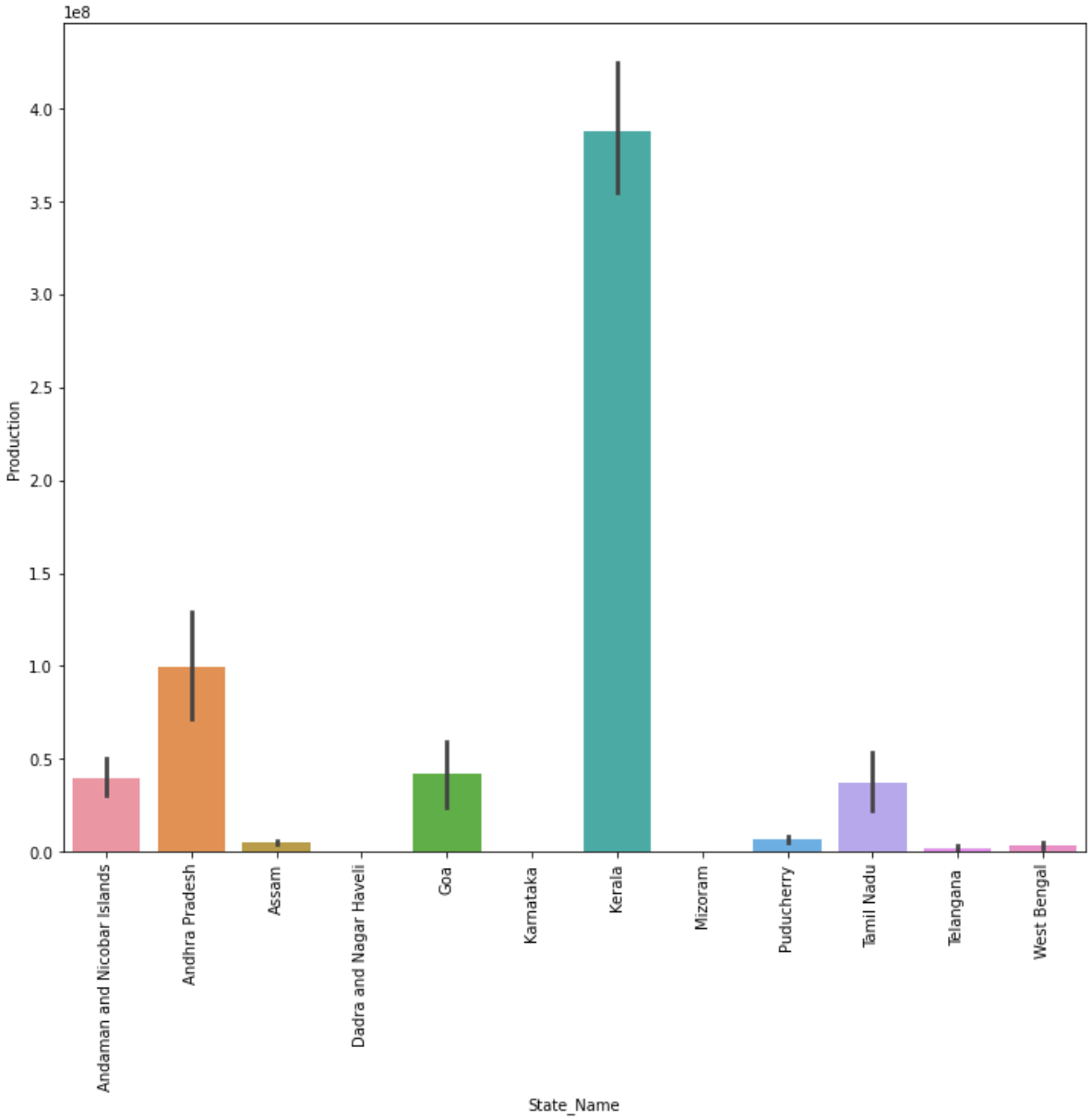

Figure 2: State wise Coconut crop yield production

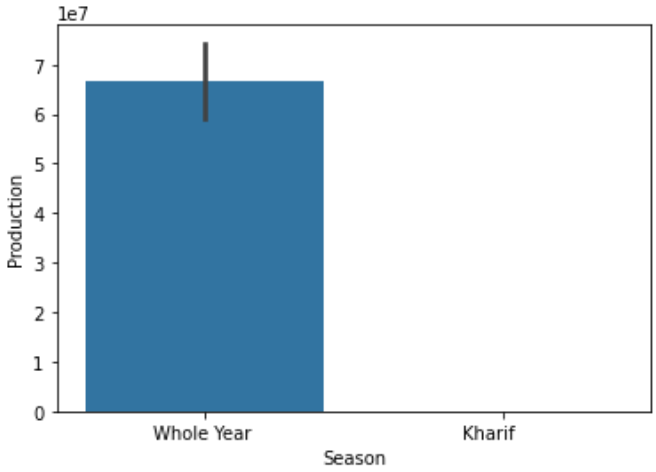

Figure 3: Season wise Coconut crop yield production

From these visualizations I found that the production of coconut crop is high in Kerala and the production is increasing gradually from 1998 to 2015 as shown in Figure 2 and Figure 4 respectively. Figure 3 shows that the production of coconut crop does not depend on any season.

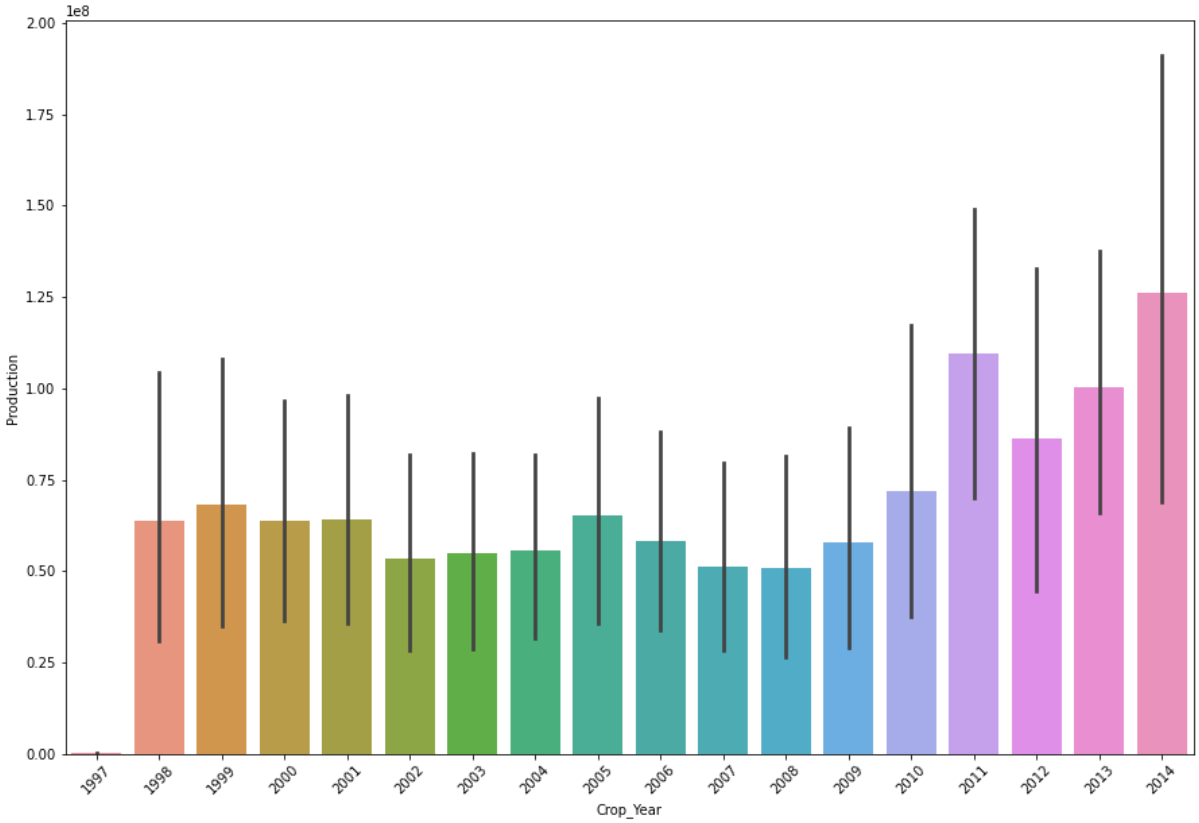

Figure 4: Cconut crop yield production from 1998 to 2014

The outcomes of varied machine learning models are contrasted based on accuracy prediction with the coconut crop yield production data of India. These outcomes have found better after applying the PCA method. The graphical observation and its prediction accuracy values are shown in Figure 5 and Table III respectively. From the outcomes, it has been found that GBR is the best prediction model as compared to the other models.

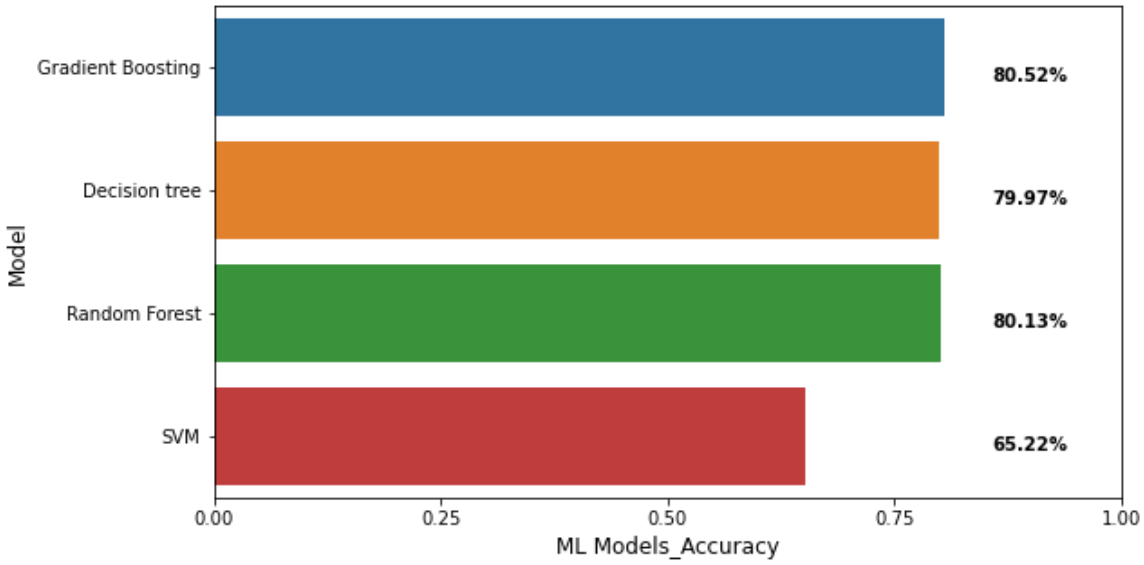

Figure 5. Accuracy of coconut crop production using Machine Learning Models

Table III. Accuracy of Machine Learning Models for the coconut crop

Machine Learning Model	**Accuracy**
GBR	80.52%
DST	79.97%
RF	80.13%
SVR	65.22%

Conclusion and Future Work

In this work we carried out an experimental work to compare popular machine learning models for coconut crop yield prediction using various accuracies over crop yield production of India. We have observed that Gradient Boosting model results best coconut crop yield prediction with an accuracy of 80.52% ≅ 81%. For this crop yield prediction, Gradient Boosting model shows a proficient as well as an acceptable model. The percentage of accuracy as well as prediction is highly determined by the data being utilized as input for prediction and regression. All models have its own benefits along with limitations and the hardest part is to decide the best model. After analyzing all above mentioned models of supervised learning, the Gradient Boosting Regressor (GBR) model has considerable level of accuracy and acceptance of our used crop yield production of coconut crop.

The accuracy of the prediction for the model may be better by implementing a hybrid prediction model where various machine learning models are assembled to work. In our future work, we are planning for implementing the hybrid prediction model to get the better and higher accuracy.

Reference

1. Anshal Savla, Himtanaya Bhadada, Parul Dhawan, Vatsa Joshi, Application of Machine Learning Techniques for Yield Prediction on Delineated Zones in Precision Agriculture, May 2015.

2. Beulah, R., 2019. A survey on different data mining techniques for crop yield prediction. Int. J. Comput. Sci. Eng. 7 (1), 738–744. https://doi.org/10.26438/ijcse/v7i1.738744.

3. D Ramesh, B Vishnu Vardhan, Analysis Of Crop Yield Prediction Using Data Mining Techniques, International Journal of Research in Engineering and Technology, Jan-2015.

4. Elavarasan, D., Vincent, D.R., Sharma, V., Zomaya, A.Y., Srinivasan, K., 2018. Forecasting yield by integrating agrarian factors and machine learning models: a survey. Comput. Electron. Agric. 155, 257–282. https://doi.org/10.1016/j.compag.2018.10.024.

5. Gour Hari Santra, Debahuti Mishra and Subhadra Mishra, Applications of Machine Learning Techniques in Agricultural Crop Production, Indian Journal of Science and Technology, October 2016.

6. Horie, T., Yajima, M., and Nakagawa, H. (1992). Yield forecasting. Agric. Syst. 40, 211–236. doi: 10.1016/0308-521X(92)90022-G.

7. Karan deep Kauri, Machine Learning: Applications in Indian Agriculture, International Journal of Advanced Research in Computer and Communication Engineering, April 2016.

8. Khaki, Saeed & Wang, Lizhi. (2019). Crop Yield Prediction Using Deep Neural Networks. Frontiers in Plant Science. 10. 10.3389/fpls.2019.00621.

9. Klompenburg, Thomas & Kassahun, Ayalew & Catal, Cagatay. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture. 177. 105709. 10.1016/j.compag.2020.105709.

10. Lever, J., Krzywinski, M. & Altman, N. Principal component analysis. Nat Methods 14, 641–642 (2017). https://doi.org/10.1038/nmeth.4346

11. Li, B., Lecourt, J., Bishop, G., 2018. Advances in non-destructive early assessment of fruit ripeness towards defining optimal time of harvest and yield prediction—a review. Plants 7 (1). https://doi.org/10.3390/plants7010003.

12. Liakos, K.G., Busato, P., Moshou, D., Pearson, S., Bochtis, D., 2018. Machine learning in agriculture: a review. Sensors (Switzerland) 18 (8). https://doi.org/10.3390/s18082674.

13. Medar, R., Rajpurohit, V. S., & Shweta, S. (2019, March). Crop Yield Prediction using Machine Learning Techniques. In 2019 IEEE 5th International Conference for Convergence in Technology (I2CT) (pp. 1-5). IEEE.

14. Nishit Jain, Amit Kumar, Sahil Garud, Vishal Pradhan, Prajakta Kulkarni, Crop Selection Method Based on Various Environmental Factors Using Machine Learning, Feb -2017.

15. Raorane A.A, Dr. Kulkarni R.V, Application of Data mining Tool to Crop Management System, January 2015.

16. Sasmita Kumarai Nayak, Swati Sucharita Barik & Mamata Beura. (2020).” Weather Forecasts Based on Rainfall Prediction Using Machine Learning Methodologies,” Adalya Journal 9 (6), Page No : 72 – 80. https://doi.org/10.37896/aj9.6/009

17. Sasmita Kumarai Nayak, Swati Sucharita Barik, Mamata Beura,” Analysis of Infectious Hepatitis Disease with High Accuracy Using Machine Learning Techniques,” TEST Engineering & Management 83 (Vol. 83: May/June 2020), 14294-14302.

18. Siti Khairunniza-Bejo, Samihah Mustaffha and Wan Ishak Wan Ismail , Application of Artificial Neural Network in Predicting, Journal of Food Science and Engineering, January 20, 2014.

19. Sripada Swain, Sasmita Kumari Nayak, Swati Sucharita Barik. (2020).” A Review on Plant Leaf Diseases Detection and Classification Based on Machine Learning Models,” Mukt shabd 9 (6), 5195-5205. DOI:09.0014.MSJ.2020.V9I6.0086781.105023

20. Tapas Ranjan Jena, Swati Sucharita Barik, Sasmita Kumarai Nayak. (2020).” Electricity Consumption & Prediction using Machine Learning Models,” Mukt shabd 9 (6), 2804-2818. DOI:09.0014.MSJ.2020.V9I6.0086781.104774