AATCC
Review

20 October 2022: Received
18 January 2023: Revised
24 January 2023: Accepted
11 February 2023: Available Online

http://aatcc.peerjournals.net/

Research Article Open Access

A Study on Castor Oil seed crop in Tamil Nadu Using Machine Learning and Non-Linear Mathematical Models

Kalpana· M[1*], Sivasankari. B[2*], Vasanthi .R[1], Pangayar Selvi. R[1] and Gitanjali. J[1]

[1]Agricultural Engineering College and Research Institute, Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu-India
[2]Agricultural Engineering College and Research Institute, Tamil Nadu Agricultural University, Madurai, India

Abstract

Castor is a non-edible industrial oilseed crop. Castor seeds are used for domestic, medicinal, and industrial purposes. Castor oil is used in machinery and particularly high-speed engines and airplanes. The present study has been undertaken to identify the best Non-Linear model and yield prediction model using machine learning techniques for castor oil seed crops in Tamil Nadu. 30 years data was collected from Season and crop Report of Tamil Nadu for Castor oil seed crop from 1990-2020. The best-fitted model was chosen based on model selection criteria like the highest coefficient of determination (R2), and with the least MAPE, RMSE, and MAE values. The four nonlinear models are fitted for the Area & Production of Castor oil seed crops. The results indicate that the Sinusoidal model is found to be the best fit for Area and Production since it has a high R2 value (0.91) and low RMSE value. According to the yield prediction model for castor oil seed crop, the Machines Learning models such as the Random forest model, Logistic Regression model, and Support Vector classifier models are considered. The study indicates that the Random Forest model is found to be the best-fitted model based on model performance metrics and the Actual vs predicted graph also clearly indicate the coincidence between the actual and predicted yield. Using these fitted models one could be able to study the trend for the Area and Production of oilseed crops in Tamil Nadu

Keywords: *Castor oil, Machine learning models, MAE, MAPE, RSME, Sinusoidal model.*

Introduction

India is regarded as a paradise of oilseed crops having 19.0 % of the total world's oilseeds area and 10.0 % of the world's oilseeds production. India is the fourth largest producer in the world in terms of output and occupies second place in an area under oilseeds.

Castor (*Ricinus communis L*) is a non-edible oilseed crop used in industry. The non-edible oil found in castor seeds is utilized for home, medical, and industrial uses. All moving elements of machinery, especially high-speed engines, and airplanes, are lubricated with castor oil. In terms of acreage (5.4 lakh hectares) and production (2.6 lakh tonnes), India is the world's leader. It contributes around 28% of global acreage and 36% of overall output. Gujarat, Rajasthan, and Andhra Pradesh are the top three castor-producing states in India, accounting for 84 percent of total production.

Tamil Nadu is an important castor-growing state in India, with an area of 15000 hectare. Major castor-producing districts are Salem, Namakkal, Erode, Dharmapuri, and Perambalur. In Tamil Nadu major seasons for castor cultivation are June-July and November-December. The productivity of castor hybrid as a pure crop under a rainfed ecosystem is 1800 kg/ha and 3000 kg/ha as a pure crop under irrigated ecosystem [1].

This study is mainly focused on computing a suitable Non-linear model for the Castor oil seed crop in Tamil Nadu. To predict the yield using Machine learning models. This study helps to know about

**Corresponding Author: Pranjal Kr Kaman and Sivasankari B*
E-mail Address: kalpusiva@gmail.com, bsankari2007@rediffmail.com
DOI:https://doi.org/10.58321/AATCCReview.2023.11.02.01

the trend analysis in the area, and the Production of Castor oil seed crops and to know about the predicted yield using different machine learning techniques in Tamil Nadu.

Materials and Methods

To assess the nonlinear models and yield prediction using machine learning tools, Data was collected from Season and crop Report, Tamil Nadu for Castor oil seed crop from 1990-2020 (30 years) [2]. There are four nonlinear models used for trend analysis [3]. Yield prediction is calculated using Machine Learning techniques such as Random Forest Regressor, SVM (Support Vector Machine), and Logistic Regression. The Nonlinear models are given below

Exponential model is $Y = Ae^{bx}$
Logistic Model of the form Y = $Y = \frac{a}{1+be^{-cx}}$
Sinusoidal model $Y = A + B\cos(CX - D)$
Hoerl Model $Y = AB^xX^c$
The machine learning Techniques are as follows,

Random Forest Regressor

A random forest is a meta-estimator that fits several classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. Random Forest Regression algorithms are a class of Machine Learning algorithms that use the combination of multiple random decision trees each trained on a subset of data. The use of multiple trees gives stability to the algorithm and reduces variance. The algorithm creates each tree from a different sample of input data. At each node, a different sample of features is selected for splitting and the trees run in parallel without any interaction. The predictions from each of the trees are then averaged to produce a single result which is the prediction of the Random Forest.

Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of the target or dependent variable is three possible classes.

Support Vector Machine

Support vector machines classification model gives good performance on unknown data. This model provides a solution to the most fundamental classification issue. This model finds hyperplane with the maximal margin. In the support vector machine, some slack variables are established to manage the nonlinear separable cases. Some training errors could be handled using this phenomenon. This reduces the effect of noise in training data. Through the selection of the uppermost probability, classification is executed. This classifier involves a penalty metric that permits a definite amount of misclassification. This is mainly imperative for non-separable training sets [4]

Steps to execute the model

1. Importing necessary libraries
2. Importing the dataset
3. Separating the features and the target variable
4. Splitting the data into a train set and a test set

The dataset is split into two Training and testing. We can also select the proposition of their division metric. In this model, the training set is 80% of the dataset and 20% is the test set. The training of a model also depends on this proportion as more training of data more chances of better accuracy. We will be using 20% of the available data as the testing set and the remaining data as the training set.

Model Performance Metrics

(i) The goodness of Fit of a Model

The goodness of fit of a model is assessed by computing the coefficient of determination R^2, Root Mean square error (RMSE), Coefficient of determination is defined as the proportion of the variance in the dependent variable that is predictable from the independent variable(s). The following prioritized equation is used to determine the R^2 value [5].

$$R^2 = 1- \frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2} \quad 0<R^2< 1$$

Root Mean Square Error (RMSE):

RMSE is defined as the square root of the average of squared errors [6].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n}}$$

Mean Absolute Error:

The lower value of these statistics is considered as the best fitted model.

$$MAE = \frac{\sum_{i=1}^{n}|Y_i-\widehat{Y_i}|}{n}$$

The criteria for deciding the best nonlinear model are one which has low RMSE value and High R^2 value. The best yield prediction of the three machine learning techniques is the one, which has low MAE, MSE and RMSE value.

Results and Discussion

There are four nonlinear models were used for studying the trend of the Castor oil seed crop in Tamil Nadu Exponential, Logistic, Hoerl, and Sinusoidal models [7]. Among the four non linear models it could be observed from Table.1 & Table.2, the Sinusoidal model is found to be the best fit for both Area and Production of Castor oil seed crop, since it has a high R^2 (0.91) value and low RMSE value.

According to the yield prediction of Castor oil seed crop using machine learning techniques it could be observed from the Table.3 that Random Forest Regressor has low MAE, MSE and RMSE values.

The Fig.1 shows that the actual yield with the predicted yield for Random Forest Model. In this graph it clearly explains that the actual yield has high coincidence with the predicted yield. The model fitting coincides with early study with the selected models for groundnut cultivation [8]

Summary and Conclusion

An attempt has been made to study the trend for the Area and Production of Castor oil seed crop in Tamil Nadu. In this study, four nonlinear models were fitted for the Area & Production of Castor oil seed crop. Secondary data was collected over a period from 1990 to 2020 (30 years). Among the four nonlinear models as Exponential, Logistic Sinusoidal, and

Table 1: Estimation of parameters for fitted nonlinear models of area (ha) in Castor (1990-2020)

Parameters	Exponential model	Logistic model	Sinusoidal model	Hoerl model
A	43986.9	2.27E-05	16628.8	24827.3
B	-0.089	1.094	14084.6	0.843
C	-	-	0.18	0.71
D	-	-	-0.88	-
R^2	0.834	0.834	**0.917**	0.897
RMSE	5991.71	5991.71	**3242.96**	3621.52

Table 2: Estimation of parameters for fitted nonlinear models of production (ha) in Castor (1990- 2017)

Parameters	Exponential model	Logistic model	Sinusoidal model	Hoerl model
A	13573.16	7.37E-05	5170.16	7629.54
B	-0.088	1.092	4345.44	0.845
C	-	-	0.181	0.711
D	-	-	-0.913	-
R^2	0.831	0.831	**0.919**	0.898
RMSE	1855.762	1855.762	**990.326**	1110.681

Table 3. Model performance metric measures for yield prediction Models

Model	Mean Absolute Error	Mean Squared Error	RMSE
Random Forest model	381.428	234221.928	**483.965**
Logistic Regression model	1756.142	3843686.428	**1960.532**
SVC classifier model	3067.857	19320308.142	**4395.487**

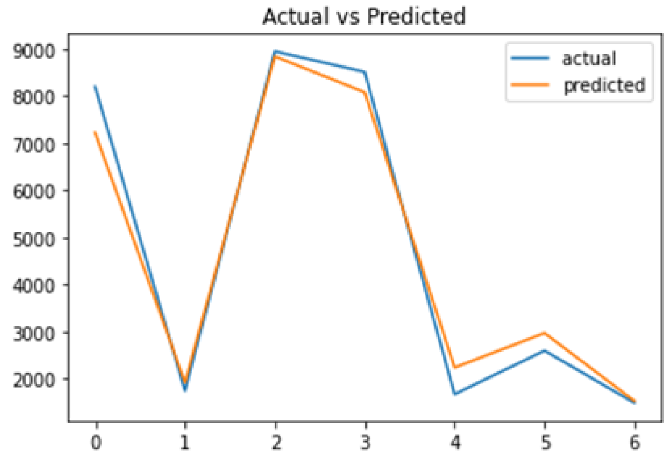

Fig.1: Actual and Predicted yield castor

Hoerl models, the best model is selected based on high R^2 and low RMSE values. The Sinusoidal model is found to be the best-fitted model for both Area and Production of Castor oilseed crops. According to yield prediction, Random Forest model is found to be the best-fitted model based on model performance metrics and the Actual vs predicted graph also clearly indicate the results. Using these fitted models, one could be able to study the trend for Area and Production of oilseed crops in Tamil Nadu

Future Scope of the Study

This R^2 and RSME values can result in strengthening the model validation for predicting the future outputs. These models can be used to study the trends of castor crop cultivation and yield in any part of the world.

Conflict of Interest

The authors declare that there is no Conflict of Interest. The authors had full access to all set of data, with an explanation of the nature and extend of access to all of the data in this study and authors take complete responsibility for the integrity of the data and accuracy of the data analysis.

Acknowledgement

I acknowledge Tamil Nadu Agricultural University, Coimbatore for providing the necessary facility to carry out the work.

References

[1.] https://agricoop.nic.in/ accessed on 27.01.2023

[2.] Basu, M.S. and Ghosh, P.K., 1995. The Status of Technologies used to Achieve High Groundnut Yields in India", In Achieving High Groundnut Yields, Patancheru, India, ICRISAT

[3.] Das, P. K. (2000). Growth models for describing state-wise wheat productivity. *Indian Journal of Agricultural Research*, *34*(3), 179-181.

[4.] Camps-Valls G, Gomez-Chova L, Calpe-Maravilla J, Soria-Olivas E, Martin-Guerrero J D, Moreno J, "Support Vector Machines for Crop Classification using Hyper Spectral Data", *Iberian Conference on Pattern Recognition and Image Analysis*, pages: 134-141, January 2003.

[5.] Karthik, V., 2010. An Economic Analysis of Production, Processing, and Marketing of Turmeric in Dharmapuri District, Unpublished PG. (Ag) a thesis submitted to the Department of Agricultural Economics, Tamil Nadu Agricultural University, Madurai.

[6.] Shapiro, S. S., Wilk, M. B. and Chen, H. J., 1968. A comparative study of various tests for normality. *Journal of the American Statistical Association*, 63(324): 1343-1372.

[7.] Venugopalan, R. and Shamasundaran, K.S., 2003. Nonlinear Regression: A realistic modeling approach in Horticultural crops research. *J.Ind. Soc. Ag.Statistics,* 56(1):1-6.

[8.] Paul, K.S.R., Faruk, M. and Rambabu, V.S. 2013. Trend, Growth, and Variability of Groundnut Crop in Andhra Pradesh. *J. Res. Arts Edu.* 2(6): 74-78.