Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 51 (2) 183-188 (2019)

http://doi.org/10.26692/sujo/2019.6.32

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Regression Analysis of Rice Data for Yield Prediction Using Python Programming Language

I.A. SUPRO, J. A. MAHAR, A. MAITLO

Department of Computer Science, Shah Abdul Latif University, Khairpur, Pakistan

Received 22nd July 2018 and Revised 5th May 2019

Abstract: The interdisciplinary domain Data Science exists ubiquitously for helping to filter out status of the passive data existing over the internet through analytics techniques on Big Data. In fact, it is intricate procedure of exploring different data set to disclose facts including hidden pattern, unidentified correlations and market trend that could assist organizations make business verdicts by predicting. A number of experts are working on vegetables and fruits yield prediction, the analysis of rice yield prediction using regression analysis with Python language is presented in this paper. The rice data of District Larkana is collected from Agriculture Statistic Department, Islamabad with three factors: Area under Cultivation, Production and Yield. The linear regression technique is applied to calculate the relationship between the Area under Cultivation (Independent) and its effect on Yield (Dependent). The positive, moderate and significant relationship is observed between the dependent and independent variables. This study can helps to researchers for knowing the worth of analytics techniques for prediction of harvest.

Keywords: Regression Analysis, Rice Data Set, Yield Prediction, District Larkana

1. INTRODUCTION

Agriculture is important profession across the world that serves humans. Approximately, 70% Pakistanis lives in rural areas and attached with agribusiness. Agriculture sector is the core part and contributes approximately 23.4% in Pakistan economy (Sellam, 2016). It is observed that most of the farmers uses traditional production method and earn less profit (Usman, 2016).

Pakistan's economy mainly relies on four sectors i.e. Food, Fiber crops and Horticulture, Livestock and Dairy, Fisheries and Forestry. Pakistan produces its most of the economy through agriculture sector. As per agribusiness indication, the cultivating is engaged in various process of business development in the form of formation and distribution of agriculture products, manufacturing of farmstead hardware, equipment and supplies along with their supervision and management. In order to improve the efficiency and enhancement of the agribusiness, the research scholars propose modern and advanced hardware and software technologies (Zambon, 2019).

Various statistical and data mining techniques have been used to solve the problems and issues pertaining to data analytics and prediction. Seven big data analytics are found during the literature review (Vivekananth, 2015). These techniques are: Association Rule Learning, Sentimental Analysis, Regression Analysis (RA),Machine Learning (ML), Genetic algorithms, Classification tree analysis, Social network analysis. The RA is frequently used by the researchers (Usman, 2016) (Altaf, 2015) for yield prediction so that initially this analytics technique is selected in this study for rice yield of District Larkana however, other data analytics techniques will be implemented on the same rice data set used in the experiments.

Crop yield prediction is necessary for rapid decision making. This critical issue has been solved with various statistical, artificial intelligence, remote sensing and machine learning approaches (Chlingaryan, 2018) using the data sets of different harvests. The yield prediction systems are also available and help to the farmers in reducing the losses (Sellam, 2016). This paper presented the analysis of the rice data of District Larkana using RA technique to predict the yield. The Python programming language is used for the experiments.

2. RICE PRODUCTION

Worldwide, Pakistan has 4th position in the rice export. (**Table 1**) shows the rice exports in the years 2015-16 and 2016-17. Approximately, 25% rice produced in six districts of province Sindh i.e. Thatta, Badin, Dadu, Larkana, Shikarpur and Jacobabad. Among the all districts, approximately 20% rice produced in Larkana (Altaf, 2015) hence, data of this district is preferred for experiments.

Table-1 Pakistan Rice Exports

Months	Years	
	2015-16	**2016-17**
November	547,286	438,399
December	475,346	391,161
January	390,323	390,690
Total	**1,724,690**	**1,220,250**

The facts and figures published in the development statistics of Sindh 2017 report. In Larkana district, the rice crop is cultivated in 105,223 hectares Area and produces 409,470 Million Tons Production, which is 3891.45 Yield per Hectare in KG.

3. RELATED WORK

Since long time, RA is used to solve the statistical and data analytics problems. This technique is also used in the prediction of crop yield. Regression models are experimented by (Altaf, 2015) on the data of rice and strong relationship was observed between the dependent and independent variables. Three environmental factors and their effects on the rice yield were analyzed by (Sellam, 2016) using the RA. The calculated results proved that annual rain factor is directly influenced on the rice yield.

Weather conditions are mainly focused by (Lee, 2014) for the development of prediction system that is purely based on the agricultural data sets taken with real-time approach. A model is proposed by (Lizumi, 2018) for yield prediction of rice crop. The rice yield prediction system is used as a simulator by (Ghosh, 2015) that is based on the measurement of resource environment synthesis. The variables area of cultivation and weight of the rice are used to compute the relation between the variables.

The machine learning technique is used by (Ghadge, 2018) for scalable system of the yield prediction. Available literature proved that various yield prediction systems are claiming the appropriate results from the different vegetables and fruits including rice but research efforts are still continued to get the more accurate results from the yield prediction systems. Furthermore, various software tools are available to analyze and predict the data using the RA technique. Nowadays, Python language offers various statistical tools and functions for data analysis with easy access and preparation approach of data sets.

4. RICE DATA COLLECTION

The rice data of District larkana is mandatory component of this research so that the statistical data of crop rice is taken from the Agriculture Statistics Department, Islamabad. It is noted that the data is available at country and province level. The rice data set of district Larkana is manually prepared from the year 1981-82 to 2015-16 according to the district-wise percentage available in the literature (Altaf, 2015) (Noonari, 2016).

Two Talukas of Larkana i.e. Dokri and Bakrani are mainly considered for the cultivation of the crop rice.The rice data is collected and prepared with three variables: Area (In Hectares, Production (In Million Tons) and Yield (Per Hectare in KGs). These variables are used to perform RA on the prepared rice data sets. The collected and prepared rice data of district Larkana from the year 1981-82 to 2015-16 is shown in (**Table 2**).

5. LINEAR REGRESSION ANALYSIS

Regression is a statistical device specifically designed to compute the average relationship between the selected dependent and independent variables in terms of different units of data. The RA is a process of fitting any model or function to given data. The values of dependent variable from the independent variables can be estimate or predict using this analysis technique (Narayanan, 2011). In this study, the factor Area under Cultivation (AUC) is considered to contribute to rice Yield (Y) and found the relation and influence of selected factor on the yield of crop rice. The Y is selected as a dependent variable and AUC is considered as independent variable to experiment the RA on the rice crop because this crop is commonly cultivated in various Talukas of district Larkana.

The response variable Y is analyzed using the RA technique which alters with value of the intervention variable X (Narayanan, 2011) (Hair, 2006). The relationship between the AUC and Y is analyzed and calculated using the linear regression model. Three conditions i.e. Linearity, Nearly Normal Residuals and Constant Variability are associated with the linear model. Furthermore, the strong relationship of linear model is evaluated using the coefficient of determination (R^2) (Sellam, 2016).

Table-2 Rice Data of Larkana from 1981-82 to 2015-16

YEAR	AREA	PRODUCTION	YIELD
1981-82	186600	489300	2622.19
1982-83	177100	447400	2526.26
1983-84	177300	484100	2730.40
1984-85	168000	443600	2640.48
1985-86	168700	414400	2456.43
1986-87	208200	597400	2869.36

1987-88	207900	596300	2868.21
1988-89	208100	597500	2871.22
1989-90	188800	539200	2855.93
1990-91	204900	591800	2888.24
1991-92	205100	585700	2855.68
1992-93	189300	525000	2773.38
1993-94	209300	706200	3374.10
1994-95	191900	583000	3038.04
1995-96	190400	552900	2903.89
1996-97	210400	660400	3138.78
1997-98	211900	630300	2974.52
1998-99	214400	671500	3132.00
1999-00	213700	714600	3343.94
2000-01	162800	566500	3479.73
2001-02	130800	376900	2881.50
2002-03	152500	477600	3131.80
2003-04	170700	515100	3017.57
2004-05	161800	508800	3144.62
2005-06	194300	636900	3277.92
2006-07	81,439	283,351	3479.30
2007-08	96,200	327,939	3408.93
2008-09	95,088	370,165	3892.87
2009-10	97,823	383,600	3921.37
2011-12	98,935	388,447	3926.28
2012-13	93,944	362,100	3854.42
2013-14	98,009	322,418	3289.68
2014-15	102,134	295,284	2891.14
2015-16	105,223	409,470	3891.45

6. IMPLEMENTATION

Anaconda is the quick way to do data science and ML with R and Python programming language on different OS. In addition to, it is a standard industry product developed for testing and training on a single device or machine. It comes with several versions including, enterprise, distribution and consulting. Thus, the Python PL is used for getting results by using ML technique including regression. Few libraries are required to call the other functions of regression for prediction that are given in **(Fig.1).**

```
%matplotlib inline

import numpy as np
import pandas as pd
import scipy.stats as stats
import matplotlib.pyplot as plt
import sklearn
```

Fig.1. Libraries used for Regression in Python

Matplotlib is the library that helps to visual data. Similarly, others are also libraries used for different functionality but the important one is the sklearn because it gives the regression method functionality. The **(Fig.2)** is given that shows the code used for the prediction.

```
lm = LinearRegression()
lm.fit(X_train, Y_train)
pred_train = lm.predict(X_train)
pred_test = lm.predict(X_test)
```

Fig.2.Training and Testing Code for Prediction

In **(Fig.2),** lm is the variable assigned to Linear Regression method. It is the built-in method that produces result on the behalf of train and test variables. Fit is another method that provide us to fit the trained data for the testing and likewise both models are given to that are already defined in CSV format and inserted for prediction. In fact, lm is the responsible to predict the data because it is already assigned to a built-in method Linear Regression.

7. RESULTS

Regression analysis technique is used to calculate and analyzed the relationship between the Y (Dependent) and AUC (Independent) that helps to accomplish a decision to predict the rice yield. The selected factors are considered from 1981-82 to 2015-16. The regression technique is applied on the data set given in (**Table 2**).

The influence of AUC on Y can be analyzed with the value of R^2. The R^2 value of 0.375133255 and R value (multiple R) is 0.612481228 are received that means there is positive, direct and moderate relationship between variable Y and AUC and clearly identify the Yield is influenced by Area Along with this the Adjusted R Square is 0.355606169; it means 35% of variation in yield can be traced to explain by area. It is because of only one independent variable is used in the

experiments. If independent variable increases, the Adjusted R Square will also improve.

The results of ANOVA table (displayed by Python) show that the F is significant because significance of F or p-value of F is below 0.05. Now by this significance we can conclude our result as reliable. It means independent variable area significantly impact the Yield. Hence, research is reliable. Now the intercept is 4044.391 and slop of independent variable is -0.00559. Further the coefficient clearly states that there is negative change by 0.0055 at p-value 0.0001 means it is less than 0.05. It states that with change of β coefficient, there will be fractional/low change in yield, but in negative direction.

Every data point has one residual and the mean and sum of the residual=0. The disparity between the observed value and the predicted value can be analyzed with the residual. The residual plot of the given data is depicted in (Fig.3) shows that there is a random pattern specify that a linear model gives a decent fit to the data. It is because of the first value is negative, some point are positive then some are negative and some positive points are plotted at the end of the graph. The moderate and significant relationship between the independent and dependent variables is visually presented in (Fig.4).

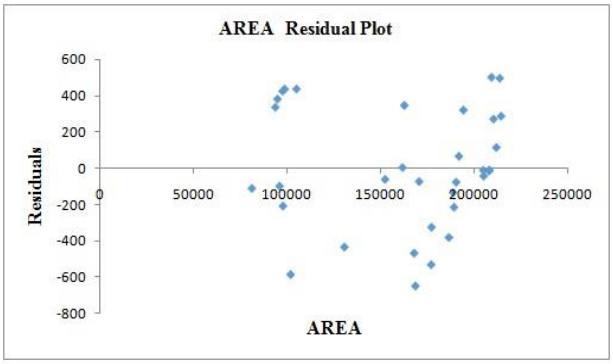

Fig.3. Residual Plot of Collected Rice Data

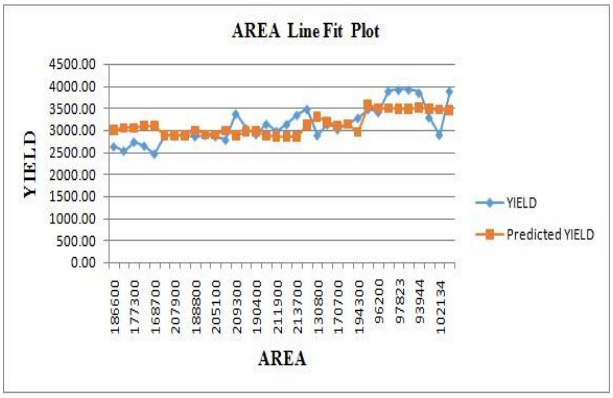

Fig.4.Line Fit Plot of Collected Rice Data

8. DISCUSSIONS

During the experiments in term of regression analysis, the calculated results of R^2 value of 0.375133255, R value is 0.612481228 and Adjusted R^2 value is 0.355606169. Due to the use of the only one independent variable in experiments, achieved results are not considered at acceptable level. Increase in number of independent variables will improve Adjusted R^2. To do so, various independent variables such as annual rainfall, humidity, temperature and wind speed should be added in to the developed data set. The source of information of these variables is Pakistan Meteorological Department, Karachi.

Prediction on unreliable and uncertain input parameters is the key challenge particularly in real-world analytics applications. Machine learning techniques are used to fulfill this challenging gape of prediction. Prediction on High-Dimensional data such as historical and contextual data is observed through Advance Machine Learning Methodologies (Al-Helal, 2019). The complex and complicated impartial functions cause difficulty to solve optimization model(s), even though one is prepared with efficient and strong predictive model. Thus a predict-then-optimized model is widely used in analytics practice. By using Python or R-Programing languages, for the betterment in results, various optimization algorithms such as Stochastic Optimization Algorithm, Stochastic Optimization Algorithm, Convergent Parallel Algorithm and Parallel Selective Algorithms will be implemented.

9. CONCLUSION

Most of the rice farmers of Larkana are unskilled to predict the upcoming crop yield. The issue of rice yield analysis and prediction is chosen in this study for experiments using the RA technique with Python programming language. The rice data is collected from the web site of Agriculture Statistics Department, Islamabad with three factors Area, Production and Yield. RA is used to examine the factor of AUC (Independent) and its infliction on Y (Dependent) of rice crop. The received value of R^2 showed that there is a positive, direct and moderate relationship between dependent and independent variables. This analytical study can be extended with some more effective factors as described in discussion section that usually effects on rice yield.

REFERENCES:

Al Helal, M., A. I. Chowdhury, A. Islam, E. Ahmed, S. Mahmud, and S. Hossain, (2019). An Optimization Approach to Improve Classification Performance in Cancer and Diabetes Prediction. International Conference on Electrical, Computer and Communication Engineering, 1-5.

Altaf, S. A., D. Jan, and B. R. Inbal, (2015). Rice Yield Estimation using Landsat ETM+Data. Journal of Applied Remote Sensing, 9, 1-16.

Chlingaryan, A., S. Sukkarieh, and B. Whelan, (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.

Ghadge, R., J. Kulkarni, P. More, S. Nene, and R. L. Priya, (2018). Prediction of Crop Yield using Machine Learning. International Research Journal of Engineering and Technology, 5(2), 2237-2239.

Ghosh, K., A. Singh, U.C., Mohanty, N. Acharya, R. K. Pal, K. K. Singh, and S. Pasupalak, (2015). Development of a Rice Yield Prediction System Over Bhubaneswar, India: Combination of Extended Range Forecast and CERES Rice Model. Meteorological Applications, 22(3), 525-533.

Hair, J. F., W. C. Black, B. J. Babin, R. E. Anderson, and R. L. Tatham, (2006). Multivariate Data Analysis. 6thed., Pearson Education Inc.

Iizumi, T., Y. Shin, W., Kim, M. Kim, and J. Choi, (2018). Global Crop Yield Forecasting using Seasonal Climate Information from Multi-Model Ensemble Climate Services, 11, 13-23.

Lee, H., and A. Moon, (2014). Development of Yield Prediction System Based on Real-Time Agricultural Meteorological Information. International Conference on Advanced Communication Technology, 1292-1295.

Narayanan Nadar, E. (2011). Statistics. PHI Learning Private Limited, 242-267.

Noonari, S., I. N. Memon, A. A. Jatoi, A. Memon, S. A. Wagan, and A. A. Sethar, (2016). Analysis of Rice Profitability and Marketing Chain: The Case Study of Taluka Pano Akil District Sukkur Sindh Pakistan. Global Journal of Agricultural Research, 4(3), 29-37.

Sellam, V. and E. Poovammal, (2016). Prediction of Crop Yield using Regression Analysis.Indian Journal of Science and Technology,9(38), 1-5.

Usman, M. (2016). Contribution of Agriculture Sector in the GDP Growth Rate of Pakistan. Journal of Global Economics, 4(2), 1-3.

Vivekananth, P., L. John, and A. Baptist, (2015). An Analysis of Big Data Analytics Techniques. International Journal of Engineering and Management Research, 5(5), 17-19.

Zambon, I., M. Cecchini, G. Egidi, and M. G. Saporito,. (2019). Revolution 4.0: Industry vs. Agriculture in a Future Development for SMEs. Process, 7(36), 1-16.