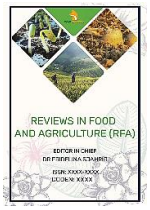
ZIBELINE INTERNATIONAL™ PUBLISHING

Reviews In Food And Agriculture (RFNA)

DOI: http://doi.org/10.26480/rfna.02.2024.86.90

ISSN: 2735-0312 (Online)
CODEN: RFAEAW



CrossMark

RESEARCH ARTICLE

A DATA-DRIVEN APPROACH TO INVESTIGATE THE YIELD PATTERNS OF BARI RELEASED POTATO VARIETIES

Istiak Ahmed[a*], Al Emran[b], Mohammad Rasel[a], Jamila Khatun Prioty[a], Md. Shakil Hossain[a]

[a]Agricultural Statistics & ICT Division, Bangladesh Agricultural Research Institute (BARI), Gazipur, Dhaka, Bangladesh.
[b]Department of Statistics, Jahanginagar University, Savar, Bangladesh
Corresponding author: istiak@bari.gov.bd

ARTICLE DETAILS

Article History:

Received 23 July 2024
Revised 18 August 2024
Accepted 13 September 2024
Available online 17 September 2024

ABSTRACT

Data Science and Analytic systems in agriculture offer policymakers a wealth of data and information, aiding in informed decision-making and potentially increasing crop yields through analysis of environmental conditions. However, this study faced limitations due to insufficient data for clustering analysis. It focused on a potato variety developed by the Bangladesh Agricultural Research Institute (BARI), revealing that most varieties were released after 2011, peaking in 2014. Yield-wise, varieties released in the same year performed similarly. Years with only one variety were excluded. Pre-2012 varieties had low yields, while high-yielding ones began in 2012 but were released irregularly. BARI Alu-74, released in 2017, has been among the low-yielding varieties since 2011. To enhance crop yields, a reliable system leveraging historical data for analysis and delivering more precise outcomes must be established. Such a system could compare and analyze data and parameters like seed quantity, watering methods, and seed type through clustering.

KEYWORDS

Data science, statistics, EDA, k-means

1. INTRODUCTION

Data science is the process of gaining knowledge from data. It enables the use of real-time and historical data to generate meaningful insights regarding consumer behavior, customer credit behavior, product testing, and cropping patterns, among others. Already, data science plays a significant role in the banking and healthcare industries. The principles of data science could also be applied to the agriculture industry, from farms to retailers, along the entire value chain. Farmers confront a lack of access to best practices from sowing to harvesting and storage as one of the obstacles. Farmers must tailor their crop selection and crop care practices to the soil and climatic conditions of their land in order to maximize yield. The automation of agricultural practices increases food production (Sreekantha, 2016).

Data science can inform farmers' land and ownership patterns about what to grow, when to plant, and what agricultural practices to use. Well-informed judgments can boost a farmer's profits. Researchers showed how data mining can access vast agricultural data (Milovic and Radojevic, 2015). Real-time farm monitoring helps farmers take immediate action. Early crop disease identification and soil moisture monitoring to maximize production. Financial institutions lack crop yield and farmer income data. A study explained how cloud computing might benefit the agriculture industry by allowing farmers to use various cloud computing services and ends customers to receive services without knowing anything about them (Patel et al., 2013). Data science solutions on farmland and agricultural trends can help financial services organizations underwrite agri-loans and insurance. Most Bangladeshi farmers possess small plots of land, making financing and insuring them unprofitable. Research has listed four agricultural chores robots can do alone (Katariya et al., 2015). Over the past decade, mobile data and IoT devices, controllers, and chipsets have become cheaper. Smartphone and mobile device use would climb. Data will flood India as it digitizes. Data-driven companies are growing. The agriculture sector's digital transformation has begun, and many inventive enterprises will find their niche while addressing its most pressing requirements. Artificial intelligence in agriculture may solve agricultural issues (Brad, 2011). Sensor-based water pumping that pumps water to the field based on soil moisture was described by (Borkar, 2015). A neural network controller-based irrigation method was proposed by (Muhammad et al., 2010). In addition, researchers described a wireless remote monitoring system utilized by researchers to improve agricultural productivity (Jain et al., 2014).

Agriculture is vital to the survival of humanity, but a dearth of knowledge and decision support will lead to a decrease in our country's production. The techniques for forecasting harvests have become increasingly sophisticated. In agriculture, highly refined statistical techniques are currently used to extract information from historical data and to project future values for economic variables. These advances in the science of harvest prediction have been made possible in large part by the development of information technology. But solitary statistical techniques do not provide an ideal future situation. Therefore, it is necessary to analyze the correlation between monitoring crop environments and harvest statistics. This information on the statistical pattern of the crop is anticipated to be obtainable from a Data Science and analytics-based decision support system. Using data science techniques, this study aims to acquire insight and knowledge about BARI-released potato varieties.

2. MATERIALS AND METHODS

Exploratory data analysis summarizes data sets using statistical graphics and other visualization approaches in statistics. Exploratory Data Analysis (EDA) can use a statistical model, but its main purpose is to learn more

Quick Response Code	Access this article online	
	Website: www.rfna.com.my	**DOI:** 10.26480/rfna.02.2024.86.90

about the data. John Tukey developed EDA to encourage statisticians to explore data and form hypotheses that could lead to new data gathering and experiments. Tukey's statistical innovations supported the analytic theory of statistical hypothesis testing. Tukey believed that statistical hypothesis testing (confirmatory data analysis) was overemphasized, and that data should be used to develop hypotheses. Unlike conventional design methods, EDA does not adhere to a rigid protocol. EDA is essentially a way of thinking. In the early stages of EDA, one is encouraged to explore any and all ideas that come to mind. Not all of these concepts will be viable in practice. Even though the questions are predetermined, EDA is still necessary because it is always necessary to investigate the quality of data. One use of EDA is in determining whether or not cleaned data is satisfactory. When cleaning data, it's necessary to use every EDA tool at your disposal, including visualization, transformation, and modeling. The purpose of EDA is to learn how to make sense of information. Using questions to direct the inquiry is the simplest method. Asking a question can assist narrow down the dataset and guide decisions about which models, graphs, and transformations to apply.

Partitioning a sample into homogenous groups in order to arrive at an operational classification is a common statistical technique. Informally measuring dimensionality, finding outliers, and proposing intriguing association theories are all possible through grouping. Classification issues can be tackled with the use of cluster analysis. The goal is to organize data by forming clusters. As a result, there is a high level of correlation within the same cluster, but only a weak level of association between clusters. The Euclidean distance is the standard for comparing the closeness of two things. The distance between two points is essentially the length of a straight line drawn between them. Euclidean distance between two p-dimensional observations (item) is given by

$$d(x,y) = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\ldots+(x_p-y_p)^2} = \sqrt{(X-Y)'(X-Y)}$$

In general, $d(x,y) = \left\{\sum_{k=1}^{p} w_k (x_{rk}-x_{sk})^2\right\}^{\frac{1}{2}}$

Un-standardized, $w_k = 1$

Standardized by S.D, $w_k = \frac{1}{S_k^2}$ (Karl Pearson distance)

Standardized by range, $w_k = \frac{1}{R_k^2}$

Generally, the distance between two points is taken as a common metric to assess this similarity among the components of a population. The most commonly used distance measure is the Euclidean metric which defines the distance between two points

$$p=(p_1,p_2,\ldots,p_k) \text{ and } q=(q_1,q_2,\ldots,q_k) \text{ as } d=\sqrt{\sum_{i=1}^{k}(p_i-q_i)^2}$$

A clustering algorithm attempts to find natural groups of components or data based on some similarity. The clustering algorithm also finds the centroid of a group of data sets. The centroid of a cluster is a point whose parameter values are the mean of the parameter values of all the points in the clusters.

Most of the commonly used clustering algorithms can be classified in two general categories namely; Hierarchical clustering and Non-Hierarchical clustering. Non-Hierarchical Clustering algorithms produce disjoint clusters and thus work well when a given set is composed of a number of distinct classes or when the data description is flat. K −means is one of the Non-Hierarchical Clustering algorithms.

***K* -means:** Suppose we have n feature vectors $X_1, X_2, \cdots, X_n$ all belonging to the same class C and we know that they belong to Kclusters such that $K < n$. If clusters are well separated, we can use a minimum distance classifier to separate them. We first initialize the means $\mu_1, \mu_2, \cdots, \mu_K$ of K clusters.

One of the ways to do this is just to assign random number to them. We then determine the membership of each X by taking the$||X-\mu_i||$. The minimum distance determines$X$'s membership in a respective cluster. This is done for all n feature vectors.

3. Results and Discussion

This study used the maximum and minimum yield data of BARI released 91 potato varieties from 1990 to 2019. BARI declared the maximum and minimum yield of a variety developed by them at the time of its release along with some characteristics. This official maximum and minimum yield data of each variety was used in this study. At first, the frequency of variety released by year is explored and showed in Figure 1 .

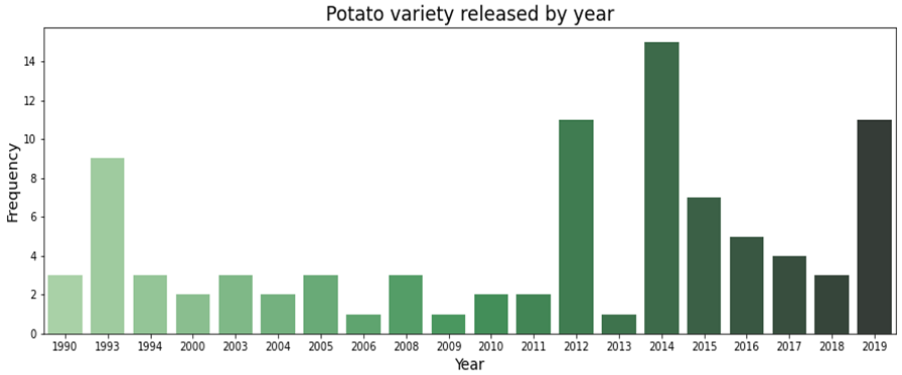


Figure 1: Potato variety released by BARI

So, most of the potato varieties are released after 2011 and in 2014. Now a box plot analysis of maximum and minimum yield is done yearly to see if the distribution is normal for yearly released variety or any of the yield is an odd ball observation.

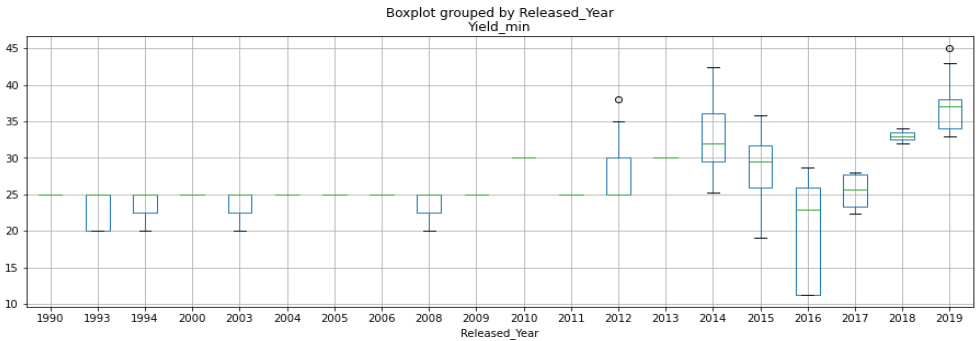


Figure 2: Year wise box plot representing the minimum yield producing potato varieties

Figure 2 reveals that most of the varieties released in the same year are very normal. The years 2012 and 2019 have one extreme observation each. These two observations should be re-checked to confirm whether they are correctly noted or not. Note that there are some years where only one variety is released. This study ignored those years. Figure 4 shows the year wise box plot representation of maximum yield potato variety.

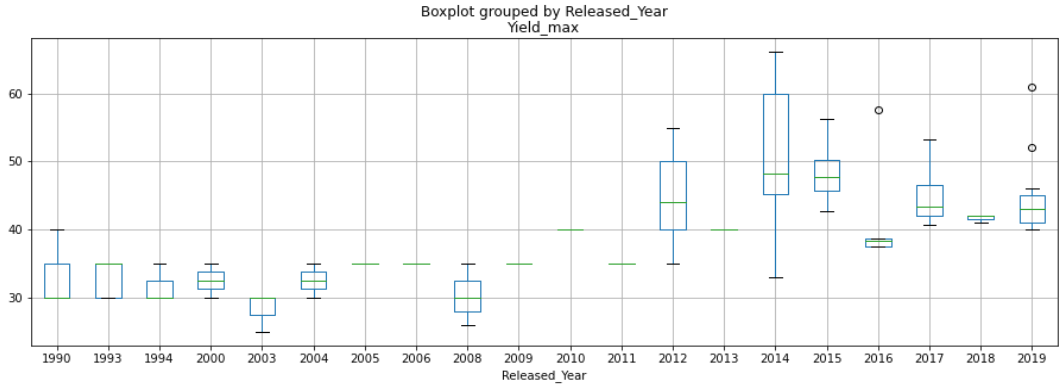


Figure 3: Year wise box plot representation of maximum yield potato variety

Here also most of the variety released in the same year are very normal. Year 2016 has one and 2019 has two extreme observations. These observations should be re-checked to confirm whether they are correctly noted or not.

Table 1: Top 5 variety in terms of minimum yield

Sl	Variety	Year	Min_yield
01	BARI Alu-72	2016	11.32
02	BARI Alu-73	2016	11.32
03	BARI Alu-68	2015	19.15
04	BARI Alu-13	1994	20.00
05	BARI Alu-29	2008	20.00

Table 2: Top 5 variety in terms of maximum yield

Sl	Variety	Year	Max_yield
01	BARI Alu-49	2014	66.11
02	BARI Alu-47	2014	63.06
03	BARI Alu-50	2014	62.87
04	BARI Alu-48	2014	62.41
05	BARI Alu-87	2019	61.00

Table 1 and 2 show the top 5 potato varieties with the lowest and highest yields respectively. BARI Alu-72 has the minimum yield with 11.32 t/ha and BARI Alu-49 has the maximum yield with 66.11 t/ha. This study intends to classify these potato varieties in two classes- high yielding and low yielding.

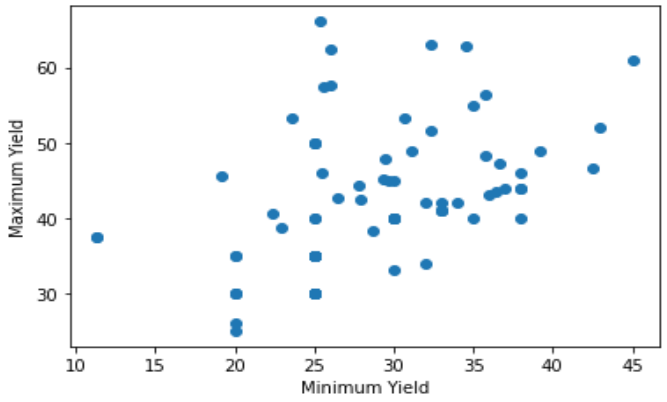


Figure 4: Scatter plot of minimum and maximum yields of different potato varieties

Figure 4 shows the scatter plot of minimum and maximum yields of the potato varieties. K-means algorithm was used to classify the varieties based on minimum and maximum yield. According to the algorithm as the varieties are being classified into two groups, the data point will have two centers. The XY coordinate of the centers are:

Center 1: (24.37, 34.77)

Center 2: (32.61, 48.85)

This means the points close to center 1 will fall low yielding variety and the points

close to center 2 will fall high yielding variety.

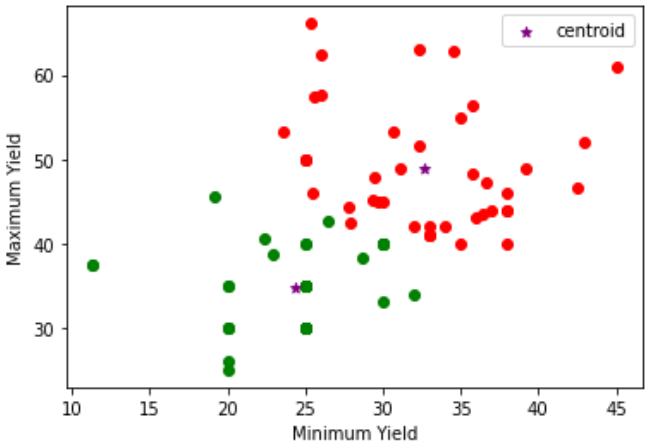


Figure 5: Classified scatter plot of minimum and maximum yield

Table 3: High yielding and low yielding variety according to k-means clustering

Low Yielding	High Yielding
BARI Alu-1, BARI Alu-2, BARI Alu-3, BARI Alu-4, BARI Alu-5, BARI Alu-6, BAI Alu-7, BARI Alu-8, BARI Alu-9, BARI Alu-10, BARI Alu-11, BARI Alu-12, BARI Alu-13, BARI Alu-14, BARI Alu-15, BARI Alu-16, BARI Alu-17, BARI Alu-18, BARI Alu-19, BARI Alu-20, BARI Alu-21, BARI Alu-22, BARI Alu-23, BARI Alu-24, BARI Alu-25, BARI Alu-26, BARI Alu-27, BARI Alu-28, BARI Alu-29, BARI Alu-30, BARI Alu-31, BARI Alu-32, BARI Alu-33, BARI Alu-34, BARI Alu-36, BARI Alu-37, BARI Alu-38, BARI Alu-39, BARI Alu-42, BARI Alu-46, BARI Alu-53, BARI Alu-55, BARI Alu-67, BARI Alu-68, BARI Alu-70, BARI Alu-71, BARI Alu-72, BARI Alu-73, BARI Alu-74	BARI Alu-35, BARI Alu-40, BARI Alu-41, BARI Alu-43, BARI Alu-44, BARI Alu-45, BARI Alu-47, BARI Alu-48, BARI Alu-49, BARI Alu-50, BARI Alu-51, BARI Alu-52, BARI Alu-54, BARI Alu-56, BARI Alu-57, BARI Alu-58, BARI Alu-59, BARI Alu-60, BARI Alu-61, BARI Alu-62, BARI Alu-63, BARI Alu-64, BARI Alu-65, BARI Alu-66, BARI Alu-69, BARI Alu-75, BARI Alu-76, BARI Alu-77, BARI Alu-78, BARI Alu-79, BARI Alu-80, BARI Alu-81, BARI Alu-82, BARI Alu-83, BARI Alu-84, BARI Alu-85, BARI Alu-86, BARI Alu-87, BARI Alu-88, BARI Alu-89, BARI Alu-90, BARI Alu-91

So, most of the varieties released before 2012 are low-yielding and releasing from 2012 are high-yielding varieties. But the release of the high-yielding varieties is not consistent. About 15 potato varieties released

after the year 2011 are low-yielding even BARI Alu-74 released in 2017 is also low-yielding.

4. Conclusions

Using this data science and analytic system for agriculture, researchers and policymakers have gained valuable insight into a wide range of factors. Assisting farmers in raising crop yields is feasible via the monitoring of environmental conditions (parameters) and the subsequent provision of this information to customers. However, there are several important qualifications that should be made about this research. The clustering procedure itself does not need a large number of parameters, to begin with. The lack of data prevents us from drawing any firmer conclusions. The study's results were meant to be useful to BARI in their decision to release the new potato variety. Preliminary studies suggest that the bulk of the variants were introduced after 2011, with 2014 seeing the biggest influx of new types. Variety minimum and maximum yields are often relatively similar among several of the same-year releases. Cluster study shows that most older varieties have low yields compared to newer ones that have been released after 2012. However, the high-yielding variety has not yet been released, and its release date is not known. The BARI Alu-74 variety, produced in 2017, has a poor yield, as do around 15 other varieties published after 2011. Building a reliable system that takes into account past data is important for increasing agricultural yields. Among the many analyses and groupings that may be run by this system is a check of the data's veracity through a comparison of eyewitness accounts. A more robust decision-support framework is the end result of factoring in seed amount, watering strategy, and seed variety.

References

Borkar, P. S. 2015. Sensor Based Water Pumping for Agriculture Field, International Journal of Emerging Trends in Engineering and Basic Sciences, 2(2), Pp. 18-22.

Brad, I. I. D.L. 2011. Adoption of artificial intelligence in agriculture cosmin POPA. Bulletin UASVM Agriculture, 68(1), Pp. 284-293.

Jain, A., S. Kudre and M. Giri. 2014. A review on smart sensors-based monitoring system for agriculture, International Journal of Engineering & Science Research, 4(5), Pp. 352-355.

Katariya, S. S., S.S. Gundal, M.T. Kanawade, Khan and Mazhar. 2015. Automation in agriculture. International Journal of Recent Scientific Research**,** 6(6), Pp. 4453-4456.

Milovic and V. Radojevic. 2015. Application of data mining in agriculture. Bulgarian Journal of Agricultural Science, 21(1), Pp. 26-34.

Patel, R. and M. Patel. 2013. Application of cloud computing in agricultural development of rural India. International Journal of Computer Science and Information Technologies, 4(6), Pp. 922-926.

Raj, M. P., P.R. Swaminarayan, J.R. Saini and D.K. Parmar. 2015. Applications of pattern recognition algorithms in agriculture: a review. International Journal of Advanced Networking and Applications, 6(5), Pp. 2495-2502.

Sreekantha, D. K. 2016. Automation in agriculture: a study. International Journal of Engineering Science Invention Research & Development, 2(12), Pp. 823-833.

Umair, S. M., and R. Usman. 2010. Automation of irrigation system using ANN based controller. International Journal of Electrical & Computer Sciences IJECS-IJENS, 10(02), Pp. 41-47.

Vibhute, A., and S.K. Bodhe. 2012. Applications of image processing in agriculture: a survey. International Journal of Computer Applications , 52(2), Pp. 34-40.