IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

Statistical and Machine Learning Analysis of Factors Affecting Sugar farm Yield: A Farm Level Study

Srushti Vijay Daga
Academic Advisor, Department of Statistics, Willingdon College, Sangli, Maharashtra
dagasrushti@gmail.com

Abstract: *This study represents an integrated analytical framework combining statistical techniques and machine learning models to evaluate agricultural productivity. It explores the key factors influencing sugarcane yield using a primary dataset collected directly from farmers. Data preprocessing techniques were applied to handle missing values and prepared the dataset for analysis. The study investigates the influence of land size, fertilizer cost, expenditure, and income using both inferential statistics and predictive modelling. In addition, machine learning models such as Linear Regression, Random Forest, and XGBoost were implemented to predict yield and capture complex relationships among variables. Among the models, modern predictive models showed better predictive performance compared to traditional methods. This study highlights the effectiveness of machine learning techniques in handling complex agricultural datasets.*

Keywords: Sugarcane Yield, Machine Learning, Regression Analysis, Random Forest, XGBoost.

I. INTRODUCTION

This Agriculture plays a vital role in the economic development, especially in regions where farming is a primary source of income. Among various crops, sugarcane is one of the most important cash crops due to its wide industrial applications and contribution to rural income. However, sugarcane yield is influenced by multiple factors including land size, irrigation practices, fertilizer usage and economic conditions. Understanding these factors is essential for improving productivity and ensuring sustainable farming practices. In recent years, data – driven approaches have become increasingly important in agricultural research. Statistical methods help identify relationships between variables, while machine learning models can capture complex patterns and improve prediction accuracy. The objective is to analyse the key factors affecting sugarcane yield and to develop predictive models using regression and machine learning techniques. The findings aim to support better decision-making and enhance agricultural productivity. The following are the references used in this paper for machine learning and other analysis: Breiman, L. (2001), Chen & Guestrin (2016), Hair (2019).

II. METHEDOLOGY

A. Sampling Techniques: Stratified sampling and systematic sampling were used for sample selection.

B. Statistical Techniques: 1. Chi-square test: To examine the association between categorical variables, 2. Shapiro Test: To check normality, 3. Mann-Whitney U Test: To compare two independent groups, 4. Kruskal-Wallis Test: To compare multiple groups, 5. Cramer's V: To measure strength of association.

C. Diagnostic Analysis: To check multicollinearity, relationships among variables and impact of variables we use VIF, Correlation Heatmap and Feature Importance (Graphical).

DOI: 10.48175/IJARSCT-36865

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 7, June 2026 Impact Factor: 8.2

D. Machine Learning Models and Evaluation Metrics: The methods were implemented such as Linear Regression, Random Forest and XGBoost. The Model peformance was evaluated with the help of Score, R^2 Score, RMSE and Cross-validation Score for checking goodness of fit, error prediction and model stability respectively.

E. Model Diagnostics: To check model assumptions we use residual plot.

III. RESULT AND DISCUSSION

The data is prepared and it can be made available on request.
https://docs.google.com/spreadsheets/d/1ynFdyUJ_HSH9vjgFd_t1z01SEMe6-VB4/edit?usp=drive_link&ouid=109370519121685317013&rtpof=true&sd=true

A. Sampling Techniques

The study uses stratified sampling to divide the population into riverside villages (Total population= 7600) and non-riverside villages (total population = 18,390).
From each group, Samples were selected based on population size. Within each village, systematic sampling was applied to select respondents. The total sample size for the study is 208 respondents.

B. Statistical Analysis

B.1 The Chi-square test was applied to examine the relationship between soil type and sugarcane productivity and the results indicate a statistically significant association ($p < 0.001$). This shows that variations in soil type directly influence productivity levels. In contrast, the relationship between riverside location and farming problems was not statistically significant ($p > 0.05$), suggesting that the river does not have a major impact on the problems faced by farmers.

B.2 The Shapiro-Wilk test was used to check the normality of income and the results indicate that the data is not normally distributed ($p < 0.05$),This suggests that non-parametric statistical methods are more suitable for further analysis.

B.3 The comparison of income between riverside and non-riverside farmers using the Mann-Whitney U test showed no significant difference ($p > 0.05$) suggesting that location does not significantly affect income levels.

B.4 The Kruskal-Wallis test revealed a significant difference in income across different land size groups ($p < 0.001$) indicating that land size has a strong influence on income and productivity.

B.5 The strength of association between categorical variables was measured using Cramer's V, which indicated a moderate relationship. This suggests that while associations exist, they are not extremely strong and may be influenced by multiple interacting factors.

C. Diagnostic Analysis

C.1 Variance Inflation Factor (VIF)

The Variance Inflation Factor (VIF) was used to show multicollinearity among the various variables. The result show high VIF values for fertilizer cost (26.85) and expenditure (25.60) indicating strong multicollinearity. The variables income (10.95) and land size (9.83) also show relatively high values. Since VIF values greater than 10 indicate serious multicollinearity, these results suggest that fertilizer cost, expenditure and income are highly interrelated which can affect the performance of Linear Regression models.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

C.2 Correlation Heatmap Graph

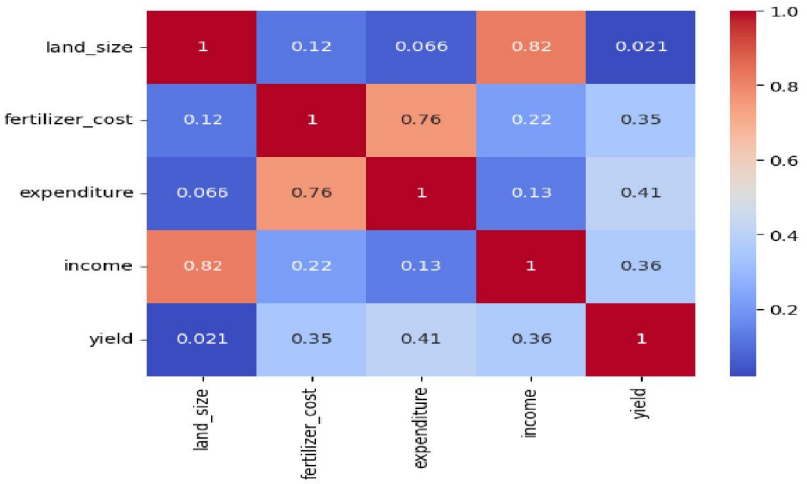

Fig. 1 Correlation Heatmap

The correlation analysis (Fig. 1) shows a strong positive relationship between income and expenditure, indicating that higher income is associated with higher spendings. A moderate relationship is observed between land size and income, suggesting that larger landholdings contribute to higher earnings.

The heatmap revealed strong correlations among income, expenditure and fertilizer cost indicating the presence of multicollinearity. This finding supports the results obtained from VIF analysis.

C.3 Feature Importance

TABLE I

	Feature	Importance
3	income	0.427295
2	expenditure	0.280372
0	land_size	0.164000
1	fertilizer_cost	0.128333

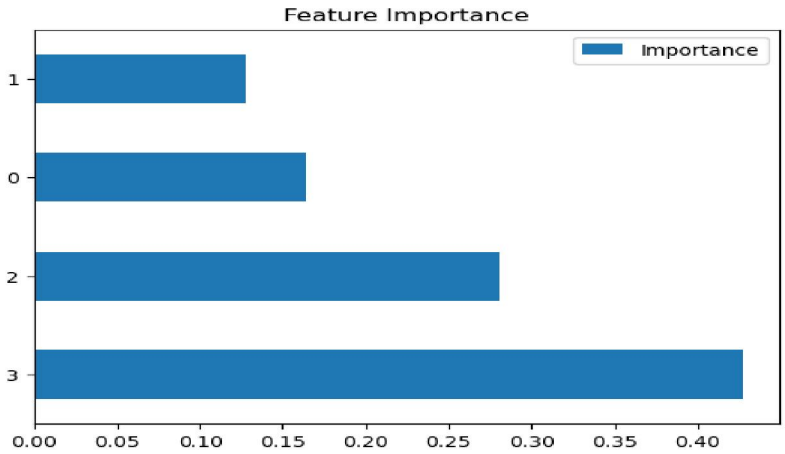

Fig. 2 Feature Importance (Based on TABLE I)

DOI: 10.48175/IJARSCT-36865

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

The feature importance analysis (Fig. 2) shows the contribution of each variable in predicting sugarcane farm productivity. The results indicate that income is the most influential variable, followed by expenditure and land size. While, fertilizer cost was found to have a comparatively less influential variable, indicating a weaker influence on the outcome.

D. Machine Learning Models and Evaluation Metrices

TABLE II

Model Comparison			
Model	R^2 Score	RMSE	CV Score
Linear Regression	0.429	6.67	0.22
Random Forest	0.600	5.57	0.48
XGBoost	0.536	6.01	0.43

The Linear Regression model has lower predictive capability (R^2 = 0.429, RMSE = 6.67) showing less accuracy due to its assumptions. Whereas, The Random Forest model has the highest accuracy (R^2 = 0.600, RMSE = 5.57) indicating better prediction capability and lower error. The XGBoost model has moderate performance (R^2 = 0.536, RMSE = 6.01) performed better than Linear Regression but slightly lower than Random Forest. So, it shows that even after boosting techniques Random Forest is still performed best for the analysis (from TABLE II)

E. Model Diagnostics

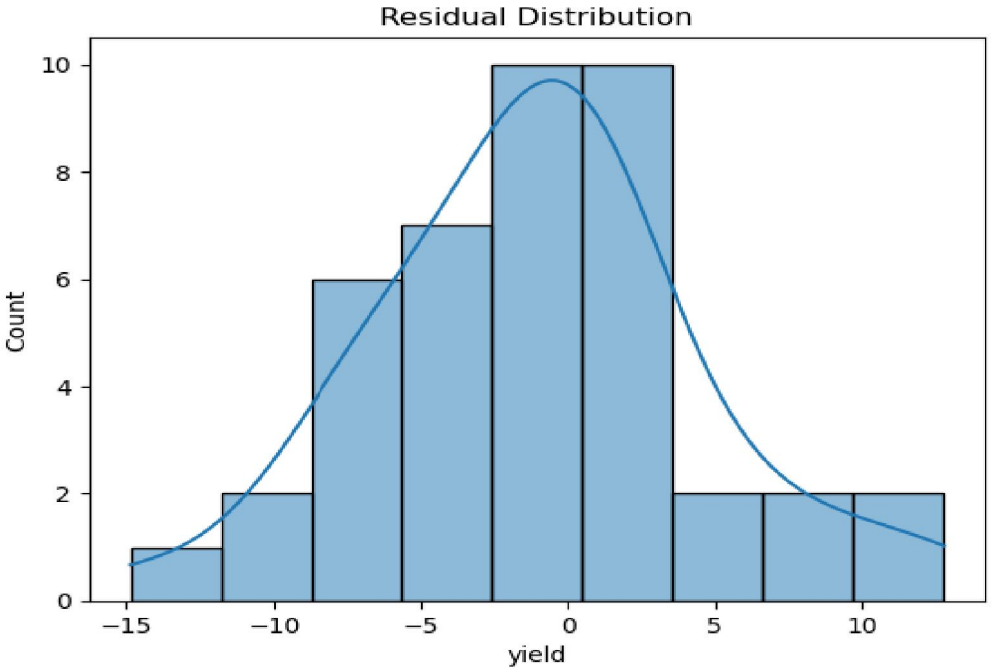

Fig. 3 Residual Plot based on Linear Regression Model

Fig. 3 shows that the residuals are not perfectly randomly distributed, indicating that model does not show all patterns in the data. There are some patterns in residuals shows the presence of non-linear relationships between income, expenditure and land size. This indicates the the limitation of Linear Regression model and suggest the use of modern predictive models such as Random Forest and XGBoost.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

IV. CONCLUSION

The study indicating that soil type and land size have a significant relationship with sugarcane productivity, while farm location does not significantly influence income and farming problems. The income data was not normally distributed, which justified the use of non-parametric tests.The correlation analysis shows a strong relationship between income and expenditure and heatmap indicates strong correlations among income, expenditure and fertilizer cost shows the presence of multicollinearity which was proved by using VIF. This multicollinearity affects the performance of Linear Regression which is supported by residual analysis. The feature importance plot shows that income, expenditure and land size are the most influential variables.

Among the machine learning models, Random Forest shows the best performance as it has less error prediction, higher stability and also it has ability to handle non-linear relationships and multicollinearity. While, Linear Regression model peforms the weakest. So, this indicates that the machine learning models provide better predictive performance than traditional methods.

REFERENCES

1. Breiman, L. (2001). Random Forests. Machine Learning.
2. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System.
3. Hair, J. F. (2019). Multivariate Data Analysis.