RANDOM FOREST AND REGRESSION ANALYSIS FOR ATTITUDE OF RESEARCH SCHOLARS TOWARDS INTERNET EXPOSURE

R. S. Parmar[1], N. M. Vegad[2], and M. P. Raj[3]

1 Professor, College of Agricultural Information Technology, AAU, Anand - 388110 India
2&3 Assistant Professor, College of Agricultural Information Technology, AAU, Anand - 388110 India
Email: rsparmar@aau.in

ABSTRACT

Internet has added a new aspect to our existence by placing within easy reach huge variety of information. Internet gives each of us the choice to be a publisher of our data, information and views. The Internet offers prosperity of agricultural business opportunities. Besides, to achieve another main objective to study an attitude of research scholars towards internet exposure using random forest and linear regression models. Random forest model was better with prediction accuracy of 87 % (R^2), lowest MAE of 1.71 and RMSE of 2.29 as compared with linear regression model. It was further noticed that, the actual and predicted attitude towards internet exposure are closed to each other. The regression analysis suggested that input features namely Academic Performance, Native of the Scholar, Family Occupation, Possession of e-extension Tools, Professional Zeal, Attitude towards e-extension Employability and Attitude towards Application of Mobile Technology has significant impact on attitude towards internet exposure.

Keywords: linear regression, random forest

INTRODUCTION

The internet in this age of development has emphasized new modes of knowledge transformation and communication patterns (Darji *et al.,* 2017; Jagadeeswari *et al.,* 2019). Internet has opened up uncommon opportunities for developing countries in terms of providing low cost access to information (Parmar *et al.,* 2018). Internet is the fastest growing tool of communication ever with a rapid increase in users (Brown, 2002). India has 60% of its population, which is dependent on Agriculture for its livelihood. Considering this, use of internet in Agriculture is of strategic importance in a country like India. Regression analysis has become one of the most widely used tools for analyzing functional relationships among the variables which is expressed in the form of an equation connecting the target variable Y and one or more features $X_1, X_2,…, X_n$ (Montgomery *et al.,* 2003). Random forest, linear regression and feature selection algorithm in regression analysis are widely used to discover useful prediction knowledge from the experimental database. Crop yield estimation with regression analysis is a newest topic in literature, and was considered for different crops like a rice (Baby *et al*., 2021), groundnut (Shah and Shah, 2018) and wheat (Hunt *et al.,* 2019). In this study, two regression models namely random forest and linear regression were upto build the most accurate and effective models since the learning information occurs with required outputs and also the objective of the study was to determine a common

rule of showing input to output. Moreover, regression models take a data-driven technique to learn useful models and relationships from input data (Willcock *et al.,* 2018) and provide a best way for improving yield predictions. In addition, regression models have some individual benefits like; they can model non-linear relationships between multiple data sources (Chlingaryan *et al.,* 2018). The present research will be useful for researchers to know an attitude of research scholars towards internet exposure.

OBJECTIVE

To know the attitude of research scholars towards internet exposure using random forest and linear regression models

METHODOLOGY

The present investigate was performed in SAUs of Gujarat. The ex-post facto research design was used for the research. Data were selected from a random sample of 120 researchers looking for post graduation in agricultural extension in SAUs of Gujarat. The Dataset having 27 features namely Academic Performance, Medium of Education at School Level, Native of the Scholar, Father's Education, Mother's Education, Skill of E-Extension Related Components, Annual Family Income, Family Occupation, Possession of E-Extension Tools, Involvement in Extracurricular Activities, Library Exposure, Information

Collection Behaviour, Job Preference, Attitude towards Extension Work, Achievement Motivation, Self Confidence, Interpersonal Communication, Innovativeness, Professional Zeal, Willingness to Work in Rural Area, Knowledge about E-Extension Related Components, Attitude towards E-Extension Employability, Attitude towards Mobile in TOT, Total Experience of E-Extension Related Components, Frequency of Use of E-Extension Components,, Training Taken on E-Extension Enabled Components and Purposes of Use of E-Extension Components. Attitude towards internet exposure is target variable. The experimental dataset was formed in excel sheet with .CSV extension for regression analysis by weka. Normalized model was used to normalize the dataset. Selected input features by attribute evaluator namely **"cfsSubsetEval"** and search method namely "**BestFirst**" are Academic Performance, Native of the Scholar, Family Occupation, Possession of e-extension Tools, Professional Zeal, Attitude towards e-extension Employability and Attitude towards Application of Mobile Technology, The two regression models namely random

forest and linear regression were used over the data set. Goodness of fit like coefficient of determination (R^2), MAE, RMSE, Relative Absolute Error and Root Relative Squared Error were taken into consideration for comparison each regression model.

RESULTS AND DISCUSSION

Regression analysis is a consistent method of identifying which features have impact on a topic of target. The process of performing a regression allows you to with assurance determine which features matter most, which features can be ignored, and how these features influence each other. Regression analyses (Draper and Smith (1966), Snedecor and Cochran (1967)) were performed for various approaches. The random forest and linear regression models used to examine the attitude of research scholars towards internet exposure. The result of each regression model is checked in terms of R^2, MAE, RMSE, RAE and RRSE**.** The Fig. 1 shows that the selected attributes have differed
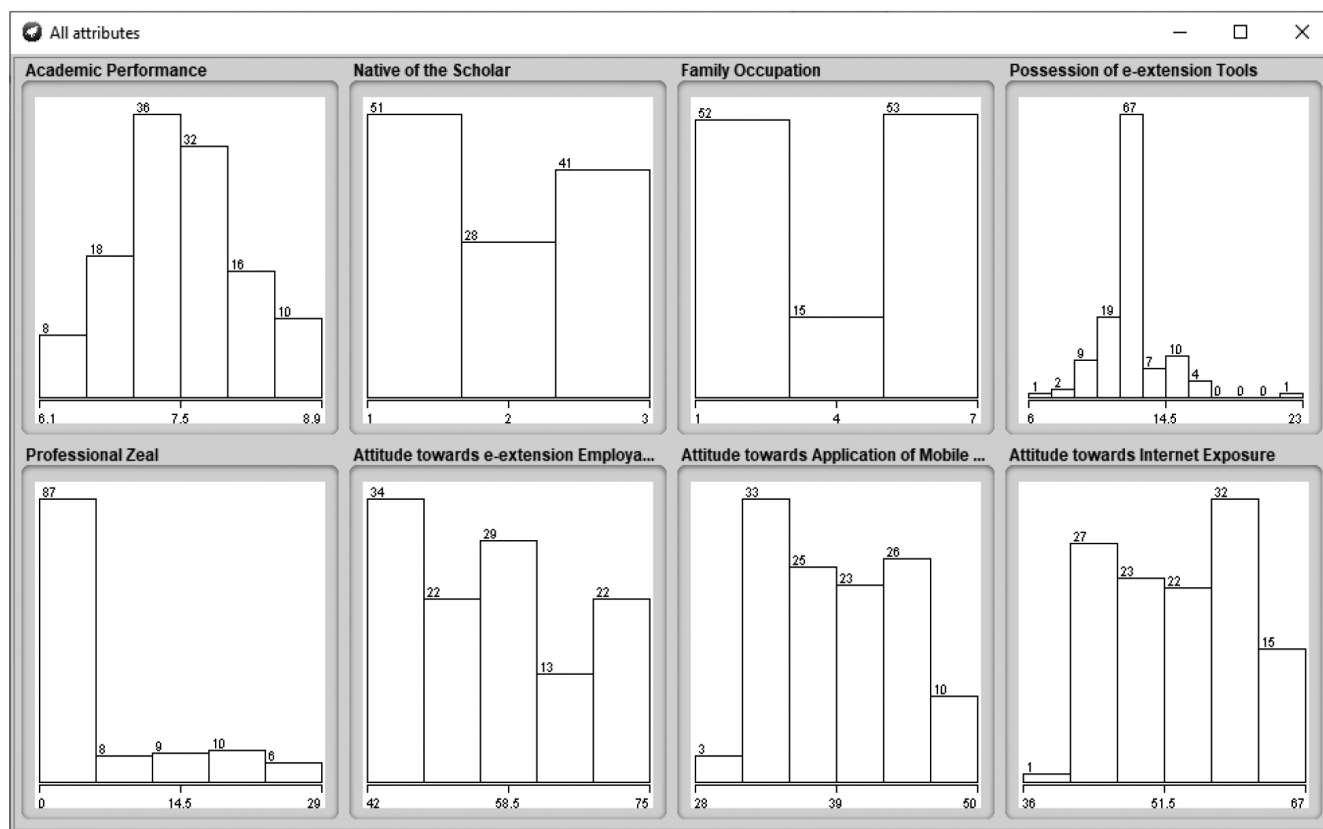
distribution range.

Fig. 1: Selected Attributes Distributions

The characteristics of fitted random forest and linear regression models in Table 1 explains that the random forest model have better performance than the linear regression model, further underlining that the random forest model provides better prediction accuracy than linear regression model.

Table 1: Characteristics of fitted random forest and linear regression models

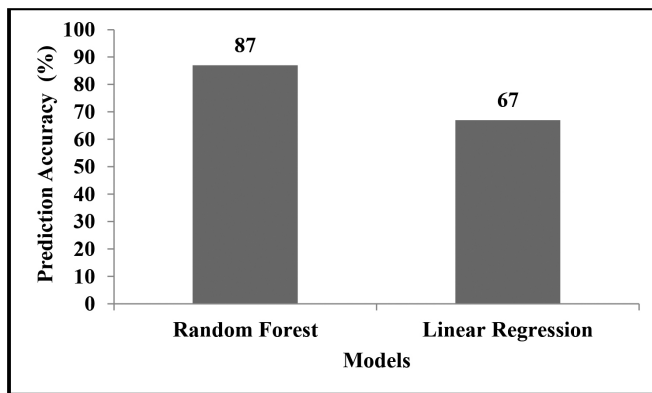| Models	Parameters				
	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)	Relative Absolute Error (RAE)	Root Relative Squared Error (RRSE)	Coefficient of Determination (R²)
Random forest	1.71	2.29	33.73 %	37.53 %	87 %
Linear regression	2.65	3.55	52.08 %	58.17 %	67 %

Fig. 2: Predication accuracy of fitted random forest and linear regression models

The Fig. 2 demonstrates the predication accuracy of fitted random forest and linear regression models. Random forest model (R^2=87%) was better in prediction accuracy as compared to linear regression model (R^2=67%).
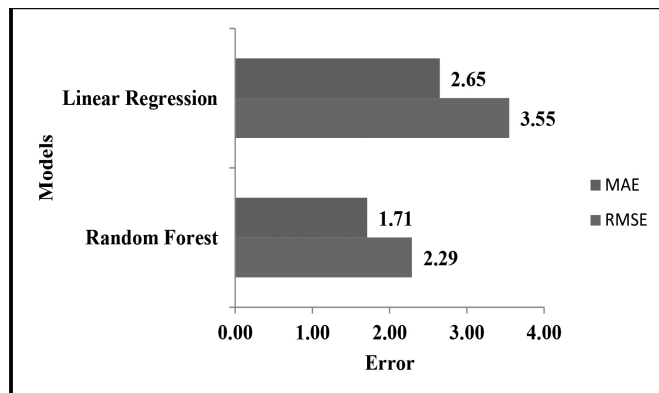

Fig. 3: Error results of fitted random forest and linear regression models

The Fig. 3 shows the error results of fitted random forest and linear regression models. Random forest demonstrated lowest MAE of 1.71 and RMSE of 2.29. This showing minimal error expected during the prediction processes. In contrast, linear regression model had resulted in highest error with 2.65 and 3.55 of MAE and RMSE,

respectively.

The actual and predicted attitude towards internet exposure using random forest model based on the testing data set is demonstrated in Table 2.

Table 2: Actual and predicted attitude towards internet exposure using random forest model

Sr. No.	Attitude towards internet exposure		
	Actual	Predicted	Residual
1	66	63.24	-2.76
2	51	52.28	1.28
3	50	50.62	0.62
4	57	56.36	-0.64
5	54	56.22	2.22
6	55	52.86	-2.14
7	60	59.69	-0.31
8	50	51.79	1.79
9	53	51.2	-1.8
10	64	57.44	-6.56
11	58	58.27	0.27
12	55	55.35	0.35
13	60	59.91	-0.09
14	58	60.43	2.43
15	59	60.06	1.06
16	51	50.49	-0.51
17	53	54.68	1.68
18	57	52.83	-4.17
19	44	48.08	4.08
20	57	54.23	-2.77
21	44	45.85	1.85
22	45	45.82	0.82
23	45	45.81	0.81
24	46	45.9	-0.10

The same is presented in Fig.4. It is observed that, the actual and predicted attitude towards internet exposure

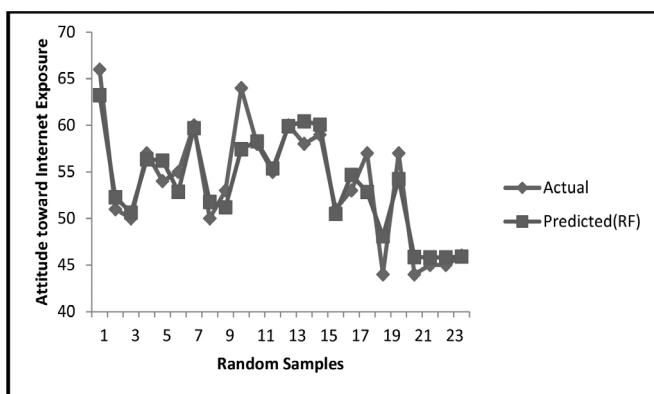are closed to each other. The residual ranged from -6.56 to 4.08.

Fig .4 Actual and predicted attitude towards internet exposure using random forest model

Table 3: Actual and predicted attitude towards internet exposure using linear regression model

Sr. No.	Attitude towards internet exposure		
	Actual	**Predicted**	**Residual**
1	66	62.021	-3.979
2	51	51.145	0.145
3	50	49.708	-0.292
4	57	56.658	-0.342
5	54	54.96	0.96
6	55	52.147	-2.853
7	60	59.329	-0.671
8	50	52.286	2.286
9	53	47.489	-5.511
10	64	55.272	-8.728
11	58	59.122	1.122
12	55	54.569	-0.431
13	60	60.304	0.304
14	58	59.239	1.239
15	59	61.084	2.084
16	51	46.081	-4.919
17	53	55.719	2.719
18	57	52.036	-4.964
19	44	50.859	6.859
20	57	51.071	-5.929
21	44	45.668	1.668
22	45	48.691	3.691
23	45	46.707	1.707
24	46	46.077	0.077

The actual and predicted attitude towards internet

exposure using linear regression model based on the testing data set is demonstrated in Table 3. The same is presented in Fig.5. As compared to random forest model, the actual and predicted attitudes towards internet exposure are less closed to each other. The residual ranged from -8.73 to 6.86.
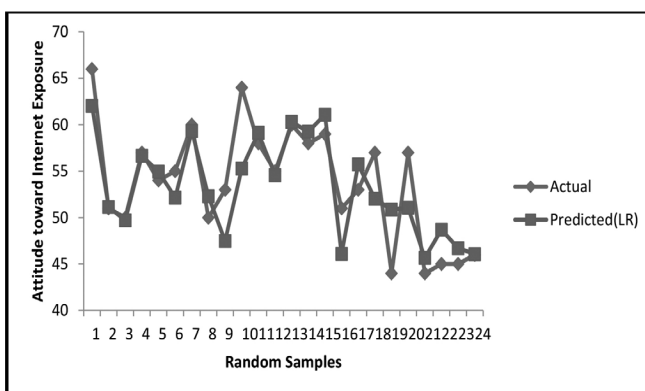

Fig. 5 Actual and predicted attitude towards internet exposure using linear regression model

CONCLUSION

The regression analysis, using random forest and linear regression models suggested that input features namely Academic Performance, Native of the Scholar, Family Occupation, Possession of e-extension Tools, Professional Zeal, Attitude towards e-extension Employability and Attitude towards Application of Mobile Technology has significant impact on Attitude towards Internet Exposure. Comparison of random forest and linear regression models with respect to the coefficient of determination (R^2), MAE and RMSE revealed that the model, which considered random forest, exhibited the highest R^2 (87 %); and lowest MAE(1.71) and RMSE(2.29).

REFERENCES

Baby, A., Parmar, R. S., Raj, M. P. and Reddy, K. I. (2021). Prediction for rice yield using data mining approach in Ranga Reddy district of Telangana, India. *J. Agrometeorology.,* 23 (2): 242-248.

Brown, M. M. (2002). 'Can ICTs address the needs of the poor?' *Choices.* 10 (2):4.

Chlingaryan, A., Sukkarieh, S. and Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. *Comput. Electron. Agric*. 151, 61–69.

Darji, A. R.. Panchasara B. R. and Vinaya Kumar H.M. (2017). Determinates of attitude of postgraduate

research scholars towards the use of computer for their empowerment. *Guj. J. Ext. Edu.,* 28 (2): 277-281.

Draper, N.R. and Smith, H. (1966). Applied Regression Analysis, John Wiley and Sons, New York.

Hunt, M.L., Blackburn, G.A., Carrasco, L., Redhead, J.W. and Rowland, C.S. (2019). High resolution wheat yield mapping using Sentinel-2. Remote Sens. *Environ*. 233, 111410.

Jagadeeswari, B., Vinaya Kumar H. M., & Patel, J. B. (2019). Attitude of postgraduate students towards research, *Guj. J. Ext. Edu.,* 30(1), 87-89.

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2003). Introduction to Linear Regression Analysis. John Wiley & sons, Inc.

Parmar, R. S., Kathiriya, D.R. and Kamani, G. J. (2015). New ICT in Agriculture: Opportunities and Challenges. *Guj.J.Ext.Edu*. Vol.26: Issue 1: 14-17.

Shah, V. and Shah, P. (2018). Groundnut crop yield prediction using machine learning techniques. *Int. J. Scient. Res. Comput. Sci. Eng. Inform. Technol.* 3 (5), 1093–1097.

Snedecor, C.W. and Cochran, W.G. (1967). 'Statistical Methods', Oxford and IBH Publishing Co., New Delhi (Sixth Edition).

Willcock, S., Hooftman, D.A., Bagstad, K.J., Balbi, S., Marzo, A., Prato, C., Sciandrello, S., Signorello, G. and Voigt, B. (2018). Machine learning for ecosystem services. *Ecosystem services* 33, 165–174.