INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Extra-tree learning based Socio-economic factor analysis and multi-class adaptive boosting meta-estimator for prediction of agricultural productivity

Bighnaraj Naik[1]*, Pandit Byomakesha Dash[1], Janmenjoy Nayak[2]

1 Veer Surendra Sai University of Technology (Formerly UCE Burla), Burla, Sambalpur, 768018, Odisha, India. Tel.: +91-7978568017
2 Aditya Institute of Technology and Management, K. Kotturu, Tekkali, Srikakulam, Andhra Pradesh, India. Tel.: 532201

OPEN ACCESS

Received: 07-06-2020
Accepted: 05-07-2020
Published: 11-08-2020

Editor: Dr. Natarajan Gajendran

Citation: Naik B, Dash PB, Nayak J (2020) Extra-tree learning based Socio-economic factor analysis and multi-class adaptive boosting meta-estimator for prediction of agricultural productivity. Indian Journal of Science and Technology 13(29): 2981-3001. https://doi.org/10.17485/IJST/v13i29.839

***Corresponding author**.

Tel: +91-7978568017
bnaik_mca@vssut.ac.in

Funding: Science and Engineering Research Board (SERB), Department of Science and Technology (DST), New Delhi, Govt. of India, under the research project Grant No. EEQ/2017/000355

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Background/Objectives: In socio-economic factor analysis, the observed data are essential in the random distribution for the adequate representation of the random components associated with various factors and lead to poor prediction in the case of the Logit and Probit model. The objective of this work is to have machine learning based model for socio-economic factors analysis and ensemble learning based model for efficient prediction of agricultural productivity. **Methods:** In this work, extra-tree classifier machine learning model based socio-economic factors selection has been used and found capable to evaluate the socio-economic factors that contain relevant information to the target variable agricultural productivity. In addition to this, the multi-class adaptive boosting ensemble learning approach is used for the prediction of agricultural productivity of respondents (farmers) from their socio-economic profiles. This proposed research has been evaluated by using the test case of analyzing the socio-economic factors of the farmers affecting agricultural productivity in Sambalpur District, of Odisha State, India. The farmers' socio-economic data are collected by using structured interviews through questionnaires that are in line with standard Participatory Rural Appraisal. **Findings:** It is found that the proposed approach of socio-economic factor identification is efficient for computing the relationships between socio-economic factors and agricultural productivity. **Novelty:** In this application domain of socio-economic factor analysis, the proposed method employs extra-tree classifier and boosting ensemble learning for socio-economic factor analysis towards agricultural productivity which is found efficient than other existing approaches such as Logit, Probit, Linear Regression, Linear Discriminant Analysis, Naïve Baise, and other counterparts.

Keywords: Socio-economic factor analysis; multiclass adaptive boosting; ensemble learning; extra-tree classifier; Probit; Logit

1 Introduction

Un-doubtfully, agriculture is the most important gift of environmental services including water, forest, pastures, and soil nutrients. However, socio-economic factors of farmers such as Marital Status, Household Size, Total Annual Income, Educational Level, Farm Size, Membership of farmers cooperative society, Years of residence, Available amenities (such as Electricity, Pipe borne water, Tarred roads, etc.), Farming experience, Quantity and Type of fertilizer used, Access to Government Schemes, etc., also plays important roles in sustainable agricultural productivity. In comparison with many other developed countries like the USA, Brazil, China, etc., India is becoming the largest hub of outsourcing of various agricultural products such as banana, guava, papaya, sugar cane, mango, etc. to other countries. In the 20th Century, India has been evolved as one of the leading farming countries for the production of several agricultural products. However, some national and international reports [(1)] indicated that the country needs to produce more major agricultural products such as rice and wheat for the increasing number of population. Many researchers investigated the reasons such as poorly maintained infrastructure, improper irrigation systems, inconsistent Govt. policies, etc. are the shortfall and low growth in agriculture. Moreover, some other major factors related to environmental, social, technological, and policy-oriented need to be taken care of at the utmost level. Technology can help in evaluating and reducing crop losses, upgrade infrastructure, and restore traditional methods of cultivation, all of which dispense towards the larger goal of enlarged productivity. Therefore, the distribution of research and modernization has the future to unlock enormous benefits in the Indian agricultural sector in which a major part of the country's population is directly and indirectly associated.

The socio-economic factors of farmers such as marital status, household size, total annual income, educational level, farm size, membership of farmers cooperative society, years of residence, available amenities (such as electricity, pipe-borne water, tarred roads, etc.), farming experience, quantity and type of fertilizer used, access to government schemes, etc., also play important roles in sustainable agricultural productivity. Coming to technological issues, 'availability of advance technological and financial requirements such as banks, cellular phones with app accessing the facility, radio signals effect, awareness about the quantity of fertilizers and pesticides to be used' are some of the major reasons for the decrement in the agricultural production. However, some of the solutions are discussed by the researchers to improve and assess productivity. In his study, Fusi et. al [(2)] mentioned that rice fertilized with urban sewage sludge and possible mitigation strategies: an environmental assessment. Their results suggested that the main correspondence to the environmental effect of rice is nitrogen emissions related to the application of diesel recycled for fieldwork and fertilizers, methane emissions reacted with the depravity of organic matter at the time of flooding period. If rice fertilized with urban sewage sludge, replacement of urban sewage sludge with organic fertilizer decides a development in categories of toxicity related impacts and applications of additional aeration period in the time of cultivation is profitable for climate change are the two productive possibilities to decrease the environmental stress. Pingali and Roger [(3)] discussed the impact of pesticides on farmer health and rice environment and analyzed that, Asia's exclusive are raising apprehension about unfavorable effects of pesticides on human health and the environment. In Asia, usage of pesticides is small, whereas usage of chemicals humiliates very quickly in tropical flooded conditions. These chemicals are very dangerous to humans, which affects their health problems. So many pesticides in Asia are intensely hazardous and effects on farmers. The primary result is the contradictory impact of pesticides on human health is more and ruin the force on the environment and as well as the paddy ecosystem. Tuong and Bouman [(4)] emphasized on rice production in water-scarce environments. Water-saving inundation, such as dying and alternate wetting and saturated soil culture can fail the non-productive water outflows and raises water productivity. It will switch aside from being often anaerobic to complete aerobic through the acceptance of water-saving irrigation technologies. The shifts will have influential and higher unknown effects on the lowland rice ecosystem. Wassmann et.al [(5)] proposed the regional vulnerability of climate change impacts on Asian rice production and scope adaptation. Any degeneration of rice productions through climate change would dangerously harm food security in Asia. The rice economies are experiencing particular climate change affected due to the rise of sea level. Powerful developments of rice production systems, i.e., larger elasticity to salinity and flooding are critical for cultivating or raising yields in these fertile deltaic regions. Indo Gangetic plans afflicted by the melting of Himalayan glaciers have a high climate change threat in Asia. Masutomi et.al [(6)] introduced impact assessment of climate change on rice production in Asia in comprehensive consideration process/parameter uncertainty in general circulation models to focus on process either in general circulation model in the evolution of the effects of climate change on rice production by using any number of climate predictions. There are three considerations on a special report on emission scenarios (SRES). The starting condition is not taken into application due to the non-availability of data. Other research related to the analysis of factors affecting the choice of crop [(7)], adoption of seed and fertilizer [(8)], changes in farmland prices [(9)] and loyalty of members in marketing co-operatives [(10)] have been reported in the field of agricultural study. Further, a study on input use in agriculture through multi-criteria analysis [(11)] has been reported and found successful for sustainable agriculture. Hamade et.al, [(12)] have analyzed qualitative and quantitative approaches to rural development through identifying impacts of technological innovation used in farming on rural farmers' households.

From the literature study in this domain of socio-economic factor analysis for agricultural productivity, it is found that few statistical and other mathematical modeling based approaches are developed for identifying the socio-economic factors affecting agricultural productivity. Yugada et.al[13] conducted a study on socio-economic factors and constraints influencing productivity among cassava farmers. Cassava is one of the important food crops grown in Africa. It is a drought-resistant and high acquiescent with enhanced pest management practices. Their study concluded that many characteristics on socio-economics of farmers in the study field such as farming experience, education, farm size along with others affect the production of cassava while harms such as; low accessibility of labor, insufficient funds, and adverse prices were with the main troubles faced by farmers are explained in their study field. Cassava farmers have socio-economic characteristics information such as age, marital status, gender, occupation, and experience on farming. Their study discloses that there was more male in the production of cassava when compared to female and the majority of cassava farmers are married. The farming experience is more than half a percent. Finally, the author's study showed that the majority of cassava farmers have knowledge of cassava farming and engaged in small scale production. Nigeria is one of the advancing countries facing scarcity of cereal crops like maize. Depend on this, Ajah and Nmadu[14] made a research on social-economic factors influencing the output of small-scale maize farmer's outcomes was held in Abuja. A multistage trail models and semi-structured inquiry were recycled for data collection. Their results showed that the land rent, the land area cultivated, years of farming experience, and the quantity of fertilizer applied were the important socio-economic aspects that significantly influenced maize outcomes. This supported the presumption that socio-economic factors impact maize output. Based on the discovery, their paper was endorsed that farmers in the study field should be intimated through augmentation services of socio-economic factors that impact on maize outcome so that farmers will consider them in the result decision-making process. Vegetables are profitable for their endowment to the share of cultivation in the Swaziland economy. At present local production of vegetables are lower than local demand, hence space is loaded by imports from South Africa. Xaba and Masuku[15] study intended to recognize the factors affecting the productivity and profitability of vegetable production. Their results showed that the factors that extensively exaggerated productivity of vegetable farmers were admittance to the gender of the farmer, fertilizer quantity, selling price, distance to market and credit were important and certainly related to the yield of the vegetable farmers whereas the distance to market was miserably related to productivity. Sorghum is the third most important cereal crop grown in the world. It is a scratchy standard rising grass used as livestock feeds, fencing houses, and food. Sorghum has been used various food items such as cake, malted beverage, bread and ethanol, and some other in major parts of the world. Zakuwai[16] conducted a study on socio-economic factors that affect sorghum production in Adamawa state, Nigeria. Socio-economic factors like age, education, marital status, and so on are the major factors affecting the level of productivity in Nigeria. Therefore their results helped makers in the country to create more knowledgeable decisions in civilizing livelihood and production of the farmers. Data were collected from 240 farmers with the help of the ordered list, using a purposive and arbitrary case. Their results disclose that mostly married with small family size, male farmers take over the venture, with small farm size, The coefficient of gender, education, credit variables, and age were expected to be unenthusiastic and statistically significant. Usman and Dodo[17] performed a study on socio-economic factors influencing agricultural insurance in rice production in Kano state, Nigeria. Agricultural insurance is necessary in urbanized countries and its profits are appreciated in the whole world. The main objectives of their study were to recognize socioeconomic factors controlling agriculturalist's compliance continue to insure their production of rice and the authors tested this assumption. Their main data were composed of a survey field using a questionnaire controlled to 120 rice farmers in the scheme of agricultural insurance. Finally, their results were concluded that farm size and formal schooling are the socioeconomic factors that manipulate farmer's compliance to continue taking rice insurance. Agricultural productivity refers to produced output by a given input in the farming sector. It can be described as the ratio of output to the inputs in farm production. Sustainable agriculture means cultivating in sustainable ways depends on the understanding of the echo system and a brief study on the association between an organism and their environment. EGWU and William[18] conducted a study on factors affecting sustainable agricultural productivity in Ebonyi state, Nigeria. Their results showed that males are the majority of respondents and they further revealed that constraints restraining sustainable farming productivity were environment, land ownership system, and funds.

Production of food in Nigeria is no longer maintained with population growth. To examine the recognized problem of apparently rejected food production in Nigeria, Anibogu et.al[19] performed a study of socioeconomic factors influencing agricultural production among cooperative farmers in Anambra State, Nigeria. Their results are vigorous with varying insightful consequences. Gender has consequence and converse relationships with farming production which indicates that a rise in more males than females in farming production activities will carry out a decrease in output of farmers. Their study explained that marital status has an optimistic relationship with farmer's outcome levels and many other educational qualifications, farming experience, type of technology employed, crop type, seeding obtained, and fertilizer acquired have a bright and important association with the output of farmer. Women cultivate an extensive amount of food eaten by entire families, but they still have

no idea or less admittance to technology, land, credit, and knowledge than men. The main objectives of the Jiriko's study [20] are to recognize the socio-economic individuality of women farmers and to resolve the association between food production and socioeconomic distinctiveness. The author's results showed that women have a low level of education and still active. So, they further cannot be engaged in the formal sector. In their study, six villages were selected and in these six villages eighty percent of women were arbitrarily chosen, two hundred women were managed with a structured questionnaire. The author found that the respondent's farm size is small, has low socio-economic distinctiveness and as a result, income produced is poor and low. The regressive analysis revealed that income, training, farm size, wealth, and inputs are the socio-economic characteristics that contributed drastically to food productions.

Socio-economic factor analysis for agricultural productivity has attracted many researchers in this field and other allied fields of science and engineering due to the social impact of this study. It is evident from literature survey that, various statistical model and few machine learning model have been applied in the field such as Descriptive statistics [13,16] , multiple regression analysis and descriptive statistics [14], Descriptive and inferential statistics [15,18] , logit model (regression model) [17,19] and Probit [21] . This study offers an advanced machine learning based model for socio-economic factor analysis and efficient ensemble learning based model for prediction of agricultural productivity. The objective of this study is to mining these socio-economic factors and designing an automated system identification model for i) Quantification of the socio-economic factors of farmers in the study area, towards agricultural productivity, iii) Extraction of other unidentified socio-economic factors through data acquisition methods and evaluation of its degree of influence of these factors on agricultural productivity through feature selection and evaluation techniques, iii) Designing of system identification model for a data-driven automated operational system for the evaluation of socio-economic factors affecting sustainable agricultural productivity, and iv) Identification of the issues of low productivity and suggestions way out.

The main contribution of this research can be summarized into two parts:

(i). Extra-tree learning based Socio-economic factors identification affecting the agricultural productivity. The major steps implemented for this approach are as follows: a) Drawing of the predefined number of sample of socio-economic profiles based on the chosen unique set of socio-economic factors; b) Designing of pool of Decision tree from the derived samples; and c) Finding of socio-economic factors from the aggregates of the results of multiple Decision trees.

(ii). Designing multi-class adaptive boosting ensemble learning-based model for prediction of agricultural productivity from selected social-economic factors from Extra-tree learning model. This consists of two major steps: i) Initialization of weight vector for each socio-economic profile, iv) Obtain the vector of weighted AO prediction error and weight parameter and v) Update the weight vector and repeat until the error reaches a threshold.

The paper is organized as follows: Section 2 describes some important early developed methods and their approach to solve the problem; Section 3 comprises of Data Collection and Preprocessing; Section 4 includes Proposed Model for Socio-economic Factor Analysis and Proposed model for prediction of agricultural productivity; Simulation Results and Analysis is presented in Section 5 followed by Conclusion in Section 6.

2 Data collection and preprocessing

This study has been planned to evaluate socio-economic factors affecting agricultural productivity based on intelligent machine learning approaches and the results of the proposed methods have been considered as a case study for the Sambalpur District, Odisha State, India. Data based on a survey in 2008 by Dept. of Agriculture and Farmer's Empowerment, Govt. of Odisha [22], out of 15.582 million hectares area, the State has cultivated area of 61.80 lakh hectares (39.7% of total land). Further, these cultivated areas is consist of three types of land, such as high land, medium land, and low land, and their distribution is 48% (29.14 lakh hectares), 28% (17.55 lakh hectares) and 24% (15.11 lakh hectares) respectively. According to the Census of India [23], farming is the main livelihood for peoples of Odisha, where 61.8% of the working population are engaged in agricultural activities. Sambalpur district comprises of 9 blocks: Bamra, Jamankira, Jujomora, Kuchinda, Maneswar, Naktideul, Rairakhol, Rengali, Sambalpur and 3 Sub-Divisions: Kuchinda, Rairakhol and Sambalpur.

Un-doubtfully, agriculture is the most important gift of environmental services including water, forest, pastures, and soil nutrients. However, socio-economic factors of farmers such as Marital Status, Household Size, Total Annual Income, Educational Level, Farm Size, Membership of farmers cooperative society, Years of residence, Available amenities (such as Electricity, Pipe borne water, Tarred roads, Television service, Radio signals, GSM networks, Banks and Markets, etc.), Farming experience, Quality of seeds used, Quantity and Type of fertilizer used, Sources of labour, Sources of seeds, Pesticides Usage and Access to Government Schemes, etc., also play important roles in sustainable agricultural productivity. As per the Census of India 2011 [23], the district has a population of 10.4 Lacs, out of which 70K are the cultivators. The Rice, Groundnut, Gram, Mustard, Arhar, Castor, Linseed, and Sugarcane have mostly cultivated crops in Sambalpur. The Sambalpur sub-division has 5381 no. of cultivators out of which 4896 are male and 485 are female. Out of the total population, we have collected the sample

of farmers by using Eq.1 .

$$N = \frac{z^2 \times p \times (1-p)}{c^2} \quad (1)$$

In Eq.1 , z is the z-score value, p is the probability to be added in the sample (confidence level) and z is the confidence interval. Therefore, the Sambalpur sub-division with 5381 no. of farmers, confidence level 95%, and confidence interval 0.05, the sample size became 5373. Here a structured interview with the quaternary method has been used to collect socio-economic data from respondents (farmers). The questionnaires have been prepared in line with the Participatory Rural Appraisal standard [24–27] to collect data on various socio-economic factors as listed in Table 1.

Table 1. Socio-economic factors information

SL. No.	Socio-economic Factors
1	Age of Household Head(AH)
2	Educational Qualification of Household Head(EQHHH)
3	Household Size(HS)
4	Household participation in Farming(HPF)
5	Household Participation Qualification (HPQ)
6	Total Educational Qualification of Household(TEQH)
7	Family Type(FT)
8	No. of Dependents(NOD)
9	Agricultural Labor Units(ALU(acre))
10	Real Asset Value Status(RAV)
11	Accessibility to Outside Village(AOV)
12	Access to Electricity(AE)
13	Part Time Occupation(PTO)
14	Shelter Type(ST)
15	Farmer Type(FRT)
16	Land Size(LS)
17	Land Type - Irrigated(LT(I))
18	Land Type - Non-irrigated(LT(NI))
19	Land Type - High(LT(High))
20	Land Type - Low(LT(Low))
21	Crop Frequency(CF)
22	Awareness of Govt. Schemes(AGS)
23	Available Govt. Schemes (AvailGS)
24	Water Resources for Farming(WRF)
25	Farm Tools
26	Source of Seeds and Plants(SSP)
27	Farming Details_Kharif Crops(FD(KC))
28	Farming Details_Rabi Crops(FD(RC))
29	Farming Details_Vegetable(FD(V))
30	Farming Details_Nuts(FD(N))
31	Communication with Broadcasting and Training Program(CBTP)
32	Access to Information(AIT)
33	Credit Accessibility(CA)
34	Land Location(LL)
35	Use of Fertilizer(UF)
36	Use of Pesticide(UP)
37	High Yielding Varieties(HYV)
38	Crop Rotation System(CRS)
39	Inter-Crop System(ICS)
40	Available Extension Service(AES)
41	Farmer Membership(FM)
42	Tropical Livestock Unit(TLU)
43	Pack Animals(PA)
44	Non-farm Training(NFT)

Continued on next page

Table 1 continued

SL. No.	Socio-economic Factors
45	Agricultural Outcome(Per Acre)(AO)

In Table 1, HPQ and TEQH are the derived attributes whose value has been computed by using the Eq.2 and Eq.3, which is as per previous studies [28,29] .

$$HPQ = HPF_{[<14]} \times 0.3 + HPF_{[\geq 14 \; and \; >18]} \times 0.5 + HPF_{[\geq 18 \; and \; <50]} \times 1 \tag{2}$$

In Eq.2 , $HPF_{[<14]}$ represents no. of Households Participation in Farming (HPF) having age below 14 years. Similarly $HPF_{[\geq 14, <18]}$ is the no. of HPF having age between 14 to 18 and $HPF_{[\geq 18, <50]}$ is the no. of HPF having age between 18 to 50.

$$EQHQ = EQH_{NS} \times 0.1 + EQH_{PS} \times 0.2 + EQH_{UPS} \times 0.5 + EQH_{HS} \times 0.75 + EQH_{C} \times 1 \tag{3}$$

In Eq.3, EQH_{NS} represents no. of household with Education Qualification of Household (EQH) having no schooling (NS). Similarly EQH_{PS}, EQH_{UPS}, EQH_{HS} and EQH_{C} is for Primary Schooling (PS), Upper Primary Schooling (UPS), Higher Schooling (HS), and College Level (C). The data distribution of the collected socio-economic data from respondents has been presented in the form of a boxplot in Figure 1.

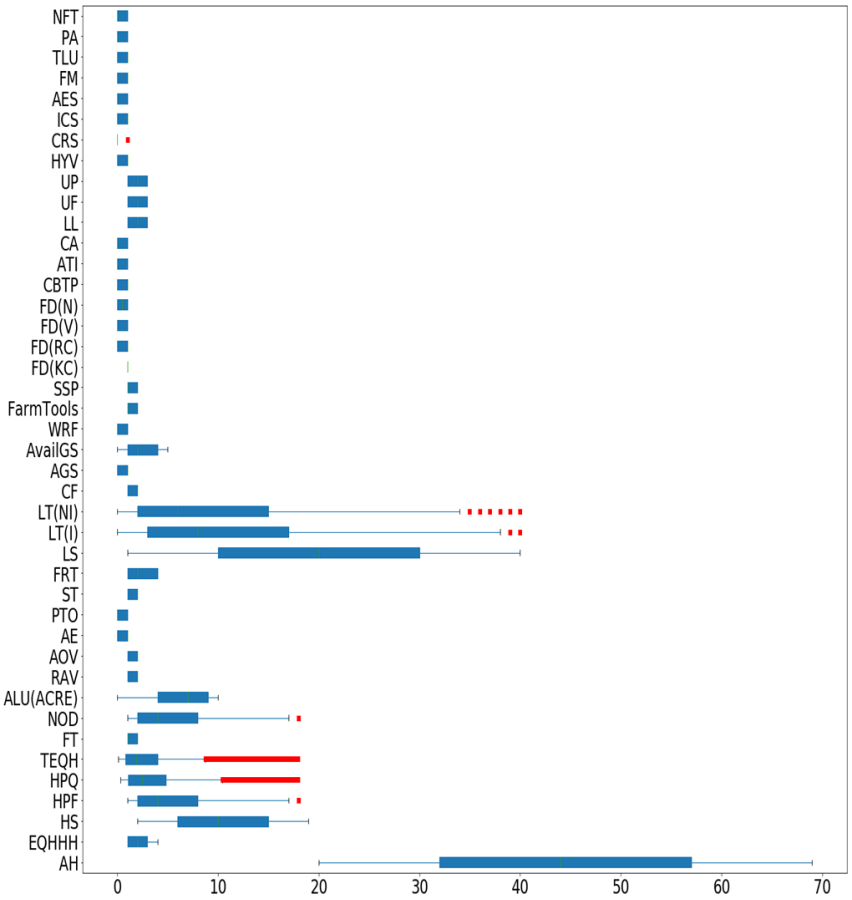

Fig 1. Data distribution of collected socio-economic profiles

In the data preprocessing phase, the data collected against the identified socio-economic factors affecting the agricultural productivity are converted to numerical values by using label encoding. For example, the respondent may provide the data for the socio-economic attribute Agricultural Outcome (AO) as 'Very Good', 'Good', 'Average', and 'Poor' which indicates the

status of their agricultural productivity. While employing label encoder, all these values 'Very Good', 'Good', 'Average' and 'Poor' are converted into '1', '2', '3' and '4' respectively. Similarly, the obtained 'Yes' and 'No' values against AvailGS are replaced with '1' and '0' respectively. During socio-economic study related to agricultural productivity in line with PRA [24,30] , we have identified 44 no. of socio-economic factors (Table 1) affecting agricultural productivity. The list of socio-economic factors are Age of Household Head(AH), Educational Qualification of Household Head(EQHHH), Household Size(HS), Household participation in Farming(HPF), Household Participation Qualification (HPQ), Total Educational Qualification of Household(TEQH), Family Type(FT), No. of Dependents(NOD), Agricultural Labor Units(ALU(acre)) , Real Asset Value Status(RAV), Accessibility to Outside Village(AOV), Access to Electricity(AE), Part-Time Occupation(PTO), Shelter Type(ST), Farmer Type(FRT), Land Size(LS), Land Type - Irrigated(LT(I)), Land Type - Non-irrigated(LT(NI)), Land Type - High(LT(High)), Land Type - Low(LT(Low)), Crop Frequency(CF), Awareness of Govt. Schemes(AGS), Available Govt. Schemes (AvailGS), Water Resources for Farming(WRF), Farm Tools, Source of Seeds and Plants(SSP), Farming Details_Kharif Crops(FD(KC)), Farming Details_Rabi Crops(FD(RC)), Farming Details_Vegetable(FD(V)), Farming Details_Nuts(FD(N)), Communication with Broadcasting and Training Program(CBTP), Access to Information(AIT), Credit Accessibility(CA), Land Location(LL), Use of Fertilizer(UF), Use of Pesticide(UP), High Yielding Varieties(HYV), Crop Rotation System(CRS), Inter-Crop System(ICS), Available Extension Service(AES), Farmer Membership(FM), Tropical Livestock Unit(TLU), Pack Animals(PA) and Non-farm Training(NFT). Further, we have considered all the possible socio-economic factors by referring to related research in this field [14,21] .

3 Proposed model

This section includes the proposed methods for (i) Socio-economic factors identification affecting the agricultural productivity (Sect. 4.1), and (ii) Designing multi-class adaptive boosting ensemble learning-based model for prediction of agricultural productivity from optimal social-economic factors (Sect. 4.2).

3.1 Proposed model for socio-economic factor analysis based on extra-tree classifier

This section includes the proposed Extra-tree learning based model for socio-economic factor analysis. The Logit model has a limitation of representing random variation and it is unable to handle the unobserved factors that are correlated over time. Eventually Probit model can handle these issues of temporally correlated errors. However, the limitation of the Probit model is that it requires all the data to be in normal distributions. In many real-life events, the normal distributions of data provide an inappropriate representation of the random components and may lead to poor prediction. Therefore in this work, machine learning based socio-economic factors selection has been used for effective results with better outcomes. Machine learning models are capable to find out the variables that contain relevant information to the target variable. In addition to this, these are competent to prune out the variables which are entitled to the addition of noise to the predictions. Logit and Probit model is designed for inference about the relationships between independent variables and dependent variables. However, the machine learning model is efficient in terms of target prediction. The proposed model employs the Extra trees classifier [31] for the selection of optimal socio-economic factors. The proposed method of socio-economic factors selection using Extra trees classifier has been presented in Algorithm 1 and Figure 3.

Algorithm 1: Extra-Tree learning model for optimal socio-economic factor

1. Let $X_i = \{X_{i,1}, X_{i,2} \ldots X_{i,N}, ao_i\}$ be the i^{th} socio-economic profiles of respondents (farmer). Here, N is the number of socio-economic factors and $ao_i \in ao$, $ao = \{1,2,3,4\}$ ={'VeryGood', 'Good', 'Average', 'Poor'} represent the productivity status. The $X = \{X_i = \{X_{i,1}, X_{i,2} \ldots X_{i,N}, a_0\}\}_{i=1}^{n}$ denotes the complete corpus with n number of socio-economic profiles.

2. Repeat for all unique k random socio-economic factor (S_k) from the total no. of socio-economic factors N $(S_N)(N = 44)$, Here S_k and S_N are set off k and N no. of socio-economic factors.

3. Create a dataset sample $X^k \subseteq X$ of k random socio-economic factors from the factor-set, where $k \subseteq N$.

4. Design a Decision Tree DT^k on the sampled data X^k by selecting suitable factors for splitting (Fig.2) based on information gain (Eq.4) [32] by using the Gini Index (Eq.5) [33] for the best splitting of the data.

$$IG\left(F_{X^k},\, f^j_{X^k}\right) = \text{InfoM}\left(F_{X^k}\right) - \frac{F_{X^k}{}^L}{F_{X^k}} \text{InfoM}\left(F_{X^k}{}^L\right) - \frac{F_{X^k}{}^R}{F_{X^k}} \text{InfoM}\left(F_{X^k}{}^R\right) \tag{4}$$

$$\text{InfoM}_{gini}\left(X^k\left[F^S_{X^k}\right]\right) = 1 - \sum_{ao_i \in ao} P\left(ao_i \mid X^k\right), S \in \{L,R\} \tag{5}$$

Here in Eq.4 and Eq.5, $IG\left(F_{X^k}, f_{X^k}^j\right)$ is information gain obtained after splitting the socio-economic factor set F_{X^k} along selected factor $f_{X^k}^j$, Info $M\left(F_{X^k}\right)$ is the information measure on F_{X^k}, Info $M_{\text{gini}\left(X^k\left[F_{X^k}^S\right]\right)}$ is the Gini information measure on the dataset X^k with selected factor $F_{X^k}^S$ and $P\left(ao_i \mid X^k\right)$ is the conditional probability of ao_i given data distribution X^k.

5. Select an optimal list of socio-economic factors from the aggregates of the results of multiple Decision trees $\left\{DT^k\right\}_{k=1}^{2^N-1}$ (34) and prediction performance.

6. Sort the socio-economic factors in descending order according to the Gini Importance.

7. Select and return the top s (pre-defined) number of socio-economic factors.

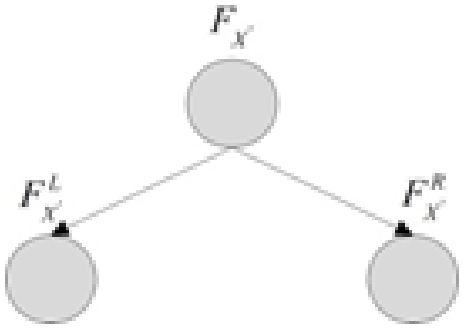

Fig 2. DT Construction through splitting along factors

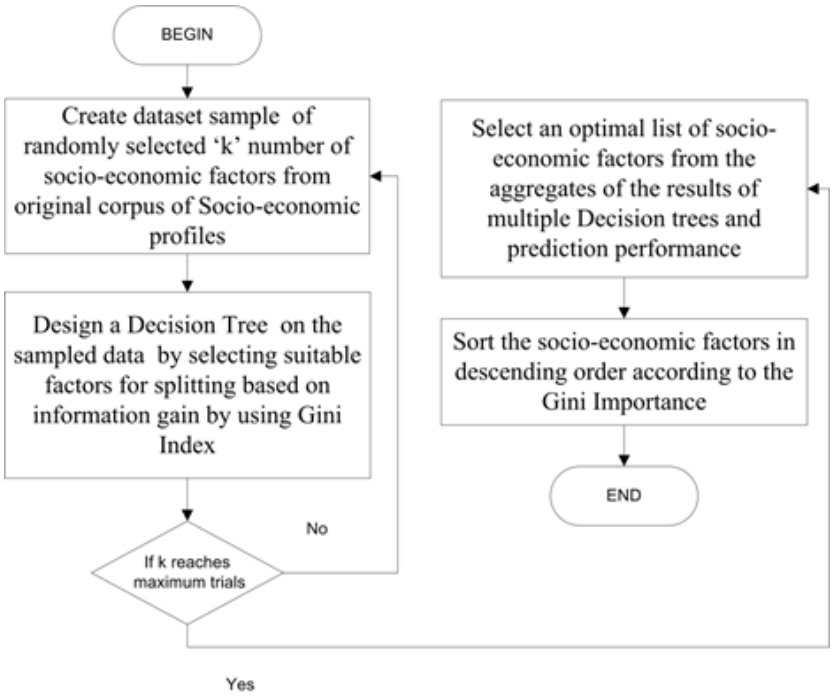

Fig 3. Proposed Approach for Socio-economic Factors Analysis

3.2 Proposed multi-class adaptive boosting ensemble learning-based model for prediction of agricultural productivity

This section presents the proposed meta-estimator based model for the prediction of agricultural productivity from selected socio-economic factors (AH, EQHHH, HS, HPF, HPQ, EQHQ, NOD, ALU(ACRE), FRT, LS, LT(I), LT(NI), AvailGS, LL, and UF) (Figure 6) by using Algorithm 1.

Let $X = (X_i, ao_i)_{i=1}^{n}$ be the recorded 'n' no. of socio-economic profiles of 'n' no. of respondents with instances of various agricultural productivity (AO) label collected from field study through structured interviews with questionnaires. Here X_i (Eq. 6) denotes i^{th} instance of recorded socio-economic profiles and ao_i represents the corresponding AO type. X_i is having 'k' no. of selected optimal socio-economic factors out of 44 no. of considered socio-economic factors (Table 1 in Appendix Section). The target variable AO are of four classes, 'Very Good', 'Good', 'Average', and 'Poor, which represent the level of agricultural productivity of the respondents.

$$X_i = \left\{X_{i,1}, \mathrm{X}_{i,2} \ldots \mathrm{X}_{i,\mathrm{k}}, ao_i\right\} \tag{6}$$

In Eq.6, k is the number of socio-economic factors of socio-economic profile in the dataset and $ao_i \in ao$ (Eq.7) is any one of the activity type.

$$ao = \{1, \quad 2, \quad 3, \quad 4\} = \{\ 'VeryGood', 'Good', 'Average', 'Poor'\ \} \tag{7}$$

$$ao_{x_i}^{j} = \Psi_{C^j}(X_i) \tag{8}$$

In Eq.8, $ao_{X_i}^{j}$ represents predicted AO type of j^{th} classifier, X_i denotes i^{th}instance of past socio-economic profile without AO type information and $\Psi_{C^j}(X_i)$ is the prediction of j^{th} classifier on X_i .

In this work, we have used Multi-class Adaptive Boosting as the model for prediction of AO type. The proposed Multi-class Adaptive Boosting[35] model makes use of the Decision tree (DT) as base classifier for the prediction of AO type ao_i. In this present work, multiclass AdaBoost has been used for boosting the performance of DTs for multiclass classification problems. This proposed model is composed of four major steps: i) Initialization of weight vector for each socio-economic profile X_i, ii) Addition of DT sequentially $DT^t(X)$ by using splitting along the features, iii) Predict AO by using each $DT^t(X)$, iv) Obtain the vector of weighted prediction error and weight parameter and v) Update the weight vector and repeat until the error reaches a threshold. The details of step by step computation can be visualized in Algorithm 2. Here the proposed model predicts the AO type from 'N' no. of DTs constructed from weighted instances (socio-economic profiles) from the training data. Sequentially, the DTs are added and trained from weighted instances in training data. The prediction error is obtained by this process and it is continued until the stopping criteria are met. Here, two stopping criteria are considered such as i) no substantial improvement in prediction performance or, ii) the required/predefined no. of DT (i.e. N) has been created. Here the aggregate of weighted average of the resultant pool of DTs' prediction give rise to final AO prediction. Algorithm 2 presents the step by step working scheme of the proposed model.

Algorithm 2: Multi-class Adaptive Boosting Ensemble Learning based Model AO Prediction

1. Initialize the weights (Eq.9) of each $X_i \in X$.

$$W_i^t = 1/n \tag{9}$$

2. For t=0 to N

i) Add DT sequentially $DT^t(X)$ by using splitting along features by using information gain computation (Eq.4) using Gini index (Eq.5).

ii) Predict the AOs (Eq.10) from trained model $DT^t(X)$.

$$ao^{'} = DT^t(\mathrm{X}) \tag{10}$$

In Eq.10, $ao^{'}$ is the vector of AO prediction and $DT^t(X)$ is the i^{th} Decision Tree applied on X.

iii) Select the model $DT^t(X)$ with smallest amount of weighted prediction error (Eq.11):

$$e^t = \mathrm{Error}\left(W^t \left[\begin{array}{c} 1 \\ ao_i^{'} \neq ao_i \end{array}\right]_{i=1}^{n}\right) \tag{11}$$

In Eq.6, e^t is the vector representing weighted AO prediction error and W^t is the t^{th} weight vector.

iv) Calculate the weight parameter (Eq.12) of t^{th} model:

$$\delta^t = \frac{1}{2} \times \ln\left(\frac{1-e^t}{e^t}\right) \tag{12}$$

In Eq.12, δ^t is the t^{th} model's weight parameter.

v) Apply Re-weighting and Update the weight of each socio-economic profile X_i (Eq.13):

$$W_{I_i}^{t+1} = \frac{W^t\left(I_{i,1}, I_{i,2} \ldots I_{i,m}, \mathbf{a_0}_i\right) e^{\left(-\delta^t \times ao_i \times DT^t\left(\mathbf{X}_i\right)\right)}}{\theta} \tag{13}$$

In Eq.13, $W_{X_i}^{t+1}$ is the $(t+1)^{th}$ weight X_i and θ is the normalization factor such that $\sum_{i=1}^{n} W_i^t = 1$.

vi) If $(e^t - e^{t+1} < \lambda$, λ is the threshold) then, Break;

Else, Continue;

End_For

3. Return the final prediction (Eq.14):

$$\Psi_{\text{AdaBoost}}(X) = \sigma\left(\sum_{i=1}^{N} \delta^t DT^t(X)\right) \tag{14}$$

In Eq.14, $\Psi_{\text{AdaBoost}}(X)$ is the final prediction on X .

End_Algorithm

4 Simulation results and analysis

4.1 Simulation environment, system and parameter setup

The experiments have been conducted in a system with Windows 10 Pro 64-bit OS, Processor Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz (8 CPUs), ~3.4GHz and 4GB RAM. The proposed model has been simulated and tested in Python programming environment. This programming environment and setup includes Pandas and Numpy framework (for data analogy); Matplotlib and Mlxtend framework (for data visualization); sklearn framework (for pre-processing of data and classification model); classification-metrics framework (for performances measurement and analysis); Seaborn (for high-level interface with informative statistical graphs for correlation analysis); statsmodels.api (for the experiment on logit and probit model) and, scipy and Itertools (for scientific computing and efficient looping respectively). All the machine learning model parameters are set as per the baseline model and tested with 70% - 30% Training and Testing Spit.

4.2 Results and analysis

In this section, the results obtained from the proposed model for optimal socio-economic factors selection (Algorithm 1) and prediction of AO type from socio-economic profile (Algorithm 2) are presented. The summary of the result of the Logit and Probit model for socio-economic factor analysis are displayed in Tables 2 and 3 respectively. It may be interpreted from the result of the Logit model that a unit increase in Rabi Crop farming (FD(RC)) results in 33.23% increase in agricultural production. However, it is likely to be increased by 20.76% in the Probit model (Table 3 and Figure 4 (b)). The list of selected socio-economic factors by using Probit, Logit, and proposed Extra-tree learning model are listed in Table 4 in the Appendix.

Table 2. Socio-economic factor analysis using Logit

Logit Model Summary				
No. Observations: 49301; Pseudo R-square: 0.009440; Log-Likelihood: -33687				
Socio-economic factors	Coef	std err	z	P>|z|
AH	-0.0007	0.001	-1.175	0.24
EQHHH	-0.0157	0.008	-1.929	0.054
HS	0.0026	0.003	0.931	0.352
HPF	-0.0028	0.006	-0.465	0.642
HPQ	-0.0048	0.007	-0.712	0.476

Continued on next page

Table 2 continued

TEQH	0.0004	0.005	0.068	0.945
FT	-0.0171	0.018	-0.939	0.348
NOD	-0.0033	0.003	-1.135	0.257
ALU(ACRE)	0.0011	0.004	0.313	0.754
RAV	-0.0038	0.018	-0.206	0.837
AOV	-0.0106	0.018	-0.58	0.562
AE	-0.0226	0.018	-1.24	0.215
PTO	-0.0162	0.018	-0.888	0.375
ST	-0.0131	0.018	-0.72	0.471
FRT	-7.16E-05	0.008	-0.009	0.993
LS	0.0002	3.78E+04	4.23E-09	1
LT(I)	0.0093	3.78E+04	2.45E-07	1
LT(NI)	-0.0091	3.78E+04	-2.40E-07	1
CF	-0.0486	0.018	-2.669	0.008
AGS	-0.0097	0.018	-0.535	0.592
AvailGS	-0.0022	0.005	-0.413	0.68
WRF	-0.0425	0.018	-2.337	0.019
FarmTools	-0.0335	0.018	-1.843	0.065
SSP	0.0131	0.018	0.719	0.472
FD(KC)	0.2397	0.109	2.193	0.028
FD(RC)	0.3323	0.018	18.195	0
FD(V)	-0.0433	0.018	-2.379	0.017
FD(N)	0.0025	0.018	0.136	0.892
CBTP	0.0083	0.018	0.455	0.649
ATI	-0.0146	0.018	-0.8	0.424
CA	-0.0267	0.018	-1.469	0.142
LL	0.0145	0.011	1.303	0.192
UF	0.0126	0.011	1.13	0.258
UP	0.0103	0.011	0.927	0.354
HYV	0.0064	0.02	0.324	0.746
CRS	-0.0267	0.023	-1.171	0.242
ICS	-0.0217	0.02	-1.097	0.273
AES	-0.0047	0.018	-0.257	0.797
FM	-0.013	0.019	-0.698	0.485
TLU	0.0256	0.02	1.289	0.197
PA	0.0034	0.02	0.174	0.862
NFT	0.0145	0.018	0.795	0.426

Table 3. Socio-economic factor analysis using Probit

Probit Model Summary				
No. Observations: 49301; Pseudo R-square: 0.009447; Log-Likelihood: -33687				
Socio-economic factors	coef	std err	z	P>|z|
AH	-0.0005	0	-1.183	0.237
EQHHH	-0.0098	0.005	-1.928	0.054
HS	0.0016	0.002	0.934	0.35
HPF	-0.0018	0.004	-0.471	0.638
HPQ	-0.003	0.004	-0.708	0.479
TEQH	0.0002	0.003	0.075	0.941
FT	-0.0107	0.011	-0.941	0.346
NOD	-0.0021	0.002	-1.133	0.257
ALU(ACRE)	0.0007	0.002	0.314	0.753
RAV	-0.0024	0.011	-0.209	0.834
AOV	-0.0066	0.011	-0.578	0.563
AE	-0.0141	0.011	-1.245	0.213
PTO	-0.01	0.011	-0.877	0.381

Continued on next page

Table 3 continued				
ST	-0.0082	0.011	-0.725	0.469
FRT	-1.14E-05	0.005	-0.002	0.998
LS	9.85E-05	6432.096	1.53E-08	1
LT(I)	0.0058	6432.096	9.01E-07	1
LT(NI)	-0.0057	6432.096	-8.82E-07	1
CF	-0.0303	0.011	-2.672	0.008
AGS	-0.0061	0.011	-0.534	0.594
AvailGS	-0.0014	0.003	-0.415	0.678
WRF	-0.0266	0.011	-2.342	0.019
FarmTools	-0.0208	0.011	-1.836	0.066
SSP	0.0082	0.011	0.72	0.471
FD(KC)	0.1499	0.068	2.199	0.028
FD(RC)	0.2076	0.011	18.205	0
FD(V)	-0.027	0.011	-2.379	0.017
FD(N)	0.0017	0.011	0.147	0.883
CBTP	0.005	0.011	0.443	0.658
ATI	-0.0091	0.011	-0.802	0.423
CA	-0.0167	0.011	-1.475	0.14
LL	0.009	0.007	1.299	0.194
UF	0.0079	0.007	1.132	0.258
UP	0.0064	0.007	0.921	0.357
HYV	0.004	0.012	0.324	0.746
CRS	-0.0166	0.014	-1.169	0.242
ICS	-0.0136	0.012	-1.104	0.27
AES	-0.0028	0.011	-0.25	0.803
FM	-0.008	0.012	-0.688	0.491
TLU	0.016	0.012	1.291	0.197
PA	0.0021	0.012	0.172	0.863
NFT	0.009	0.011	0.791	0.429

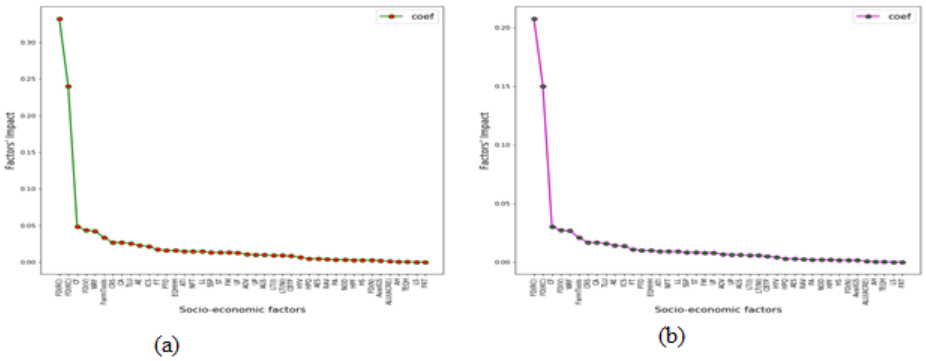

Fig 4. Socio-economic factors with ranked coefficient a)Logit Model, b) Probit Model

The list of selected socio-economic factors (Table 4) from Algorithm 1 and their data distribution has been presented in Figures 5 and 6 presents the correlation matrix of the selected socio-economic factors.

Table 4. Selected Socio-economic factors using Logit, Probitand Proposed Extra-tree Learning

Technique Used	Selected Socio-economic Factors
Logit [17,19]	FD(RC), FD(KC), CF, FD(V), WRF, FarmTools, CRS, CA, TLU, AE, ICS, FT, PTO, EQHHH, and ATI
Probit [21]	FD(RC), FD(KC), CF, FD(V), WRF, FarmTools, CA, CRS, TLU, AE, ICS, FT, PTO, EQHHH, and ATI
Proposed Extra-tree Learning	AH, EQHHH,HS, HPF, HPQ, EQHQ, NOD, ALU(ACRE), FRT, LS, LT(I), LT(NI), AvailGS, LL, and UF

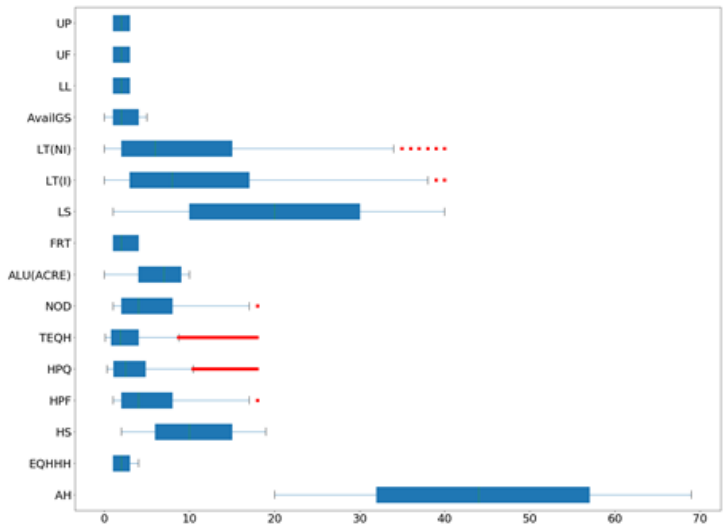

Fig 5. Boxplot Data Distribution of SelectedSocio-economic Factors

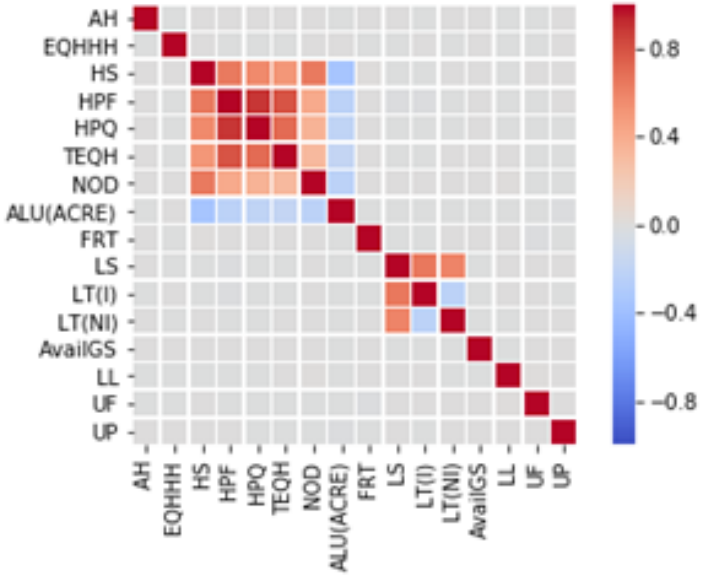

Fig 6. Correlation matrix of selected socio-economic factors

The AO prediction performance of the proposed prediction model (Algorithm 2) has been proposed and its performance has been compared with eleven standard machine learning based models: DT, K-Nearest Neigbhor[36], Naïve Bayes[37], Random Forest[38], Multi-Layer Perceptron[39], Linear Discriminant Analysis (LDA)[40], Linear Regression (LR)[41], Quadratic Discriminant Analysis (QDA)[42] and Stochastic Gradient Descent (SGD)[43]. Various performance metrics such as precision, F1-score, ROC-AUC, and recall are considered to compare all the models. The prediction of agricultural productivity by using various models such as DT, KNN, MLP, RF, NB, LDA, LR, QDA, SGD, and the proposed ensemble based model can be found in Figure 7 (a)- Figure 7(j). These figures represent the prediction of agricultural productivity in terms of four labels such as 'poor', 'average', 'good' and 'very good', where the actual and predicted values are presented with green and red colour marker respectively. Thereby, higher overlapping of each marker tends to greater prediction ability of the model. Here only 1000 no. of predictions is displayed for clear and distinguish presentation.

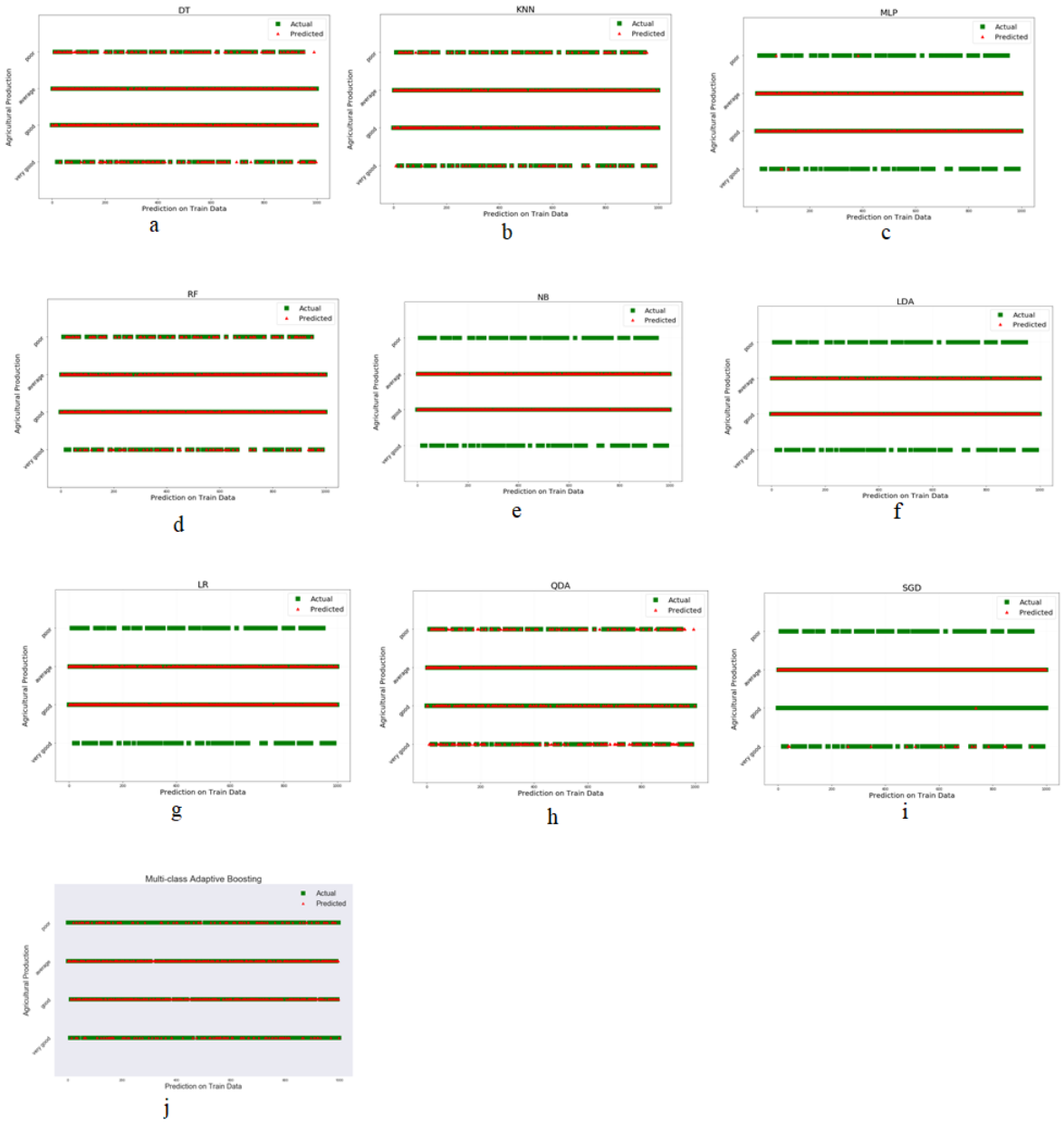

Fig 7. Prediction of productivity by using models: (a) DT,(b) KNN, (c) MLP, (d) RF, (e) NB, (f) LDA, (g) LR, (h) QDA, (i) SGD (j)Proposed method

Further, ROC analysis on the prediction of models DT, KNN, MLP, RF, SGD, NB, LR, LDA, QDA, and the proposed method are presented in Figure 8 (a)- Figure 8(j) respectively. Here, Figure 8 represents ROC w.r.t. agricultural productivity labels 1 to 4, where class 1 indicates the label 'poor', and class 4 indicates the label 'very good'. Here, It is found that Micro-average and Macro-average ROC curve have covered the area of 0.95 and 0.94 respectively and is higher than other compared models. Moreover, the class-wise coverage of the ROC curve for class 1 is 0.93, class 2 is 0.95, class 3 is 0.95 and class 4 is 0.93 respectively. Hence it is evident that the performance of the proposed method is superior to other models. A detailed comparative analysis among RF, KNN, DT, MLP, LR, and Proposed Ensemble Model with the considered performance metrics such as precision, recall & F1-score (for both class wise & overall prediction) and accuracy has been represented for all the classes in Table 5. Similarly, Table 6 presents a comparative analysis on the prediction of other considered models SGD, NB, LDA, and QDA. In Table 5 and Table 6, it is noticeable that the performance of the proposed model is superior to other models in terms of prediction. The

proposed socio-economic factors selection has been compared with the performance of Logit [17] [19] (Table 2) and Probit [21] (Table 3 in) based selected socio-economic factors (Table 2) are presented in Figure 6.

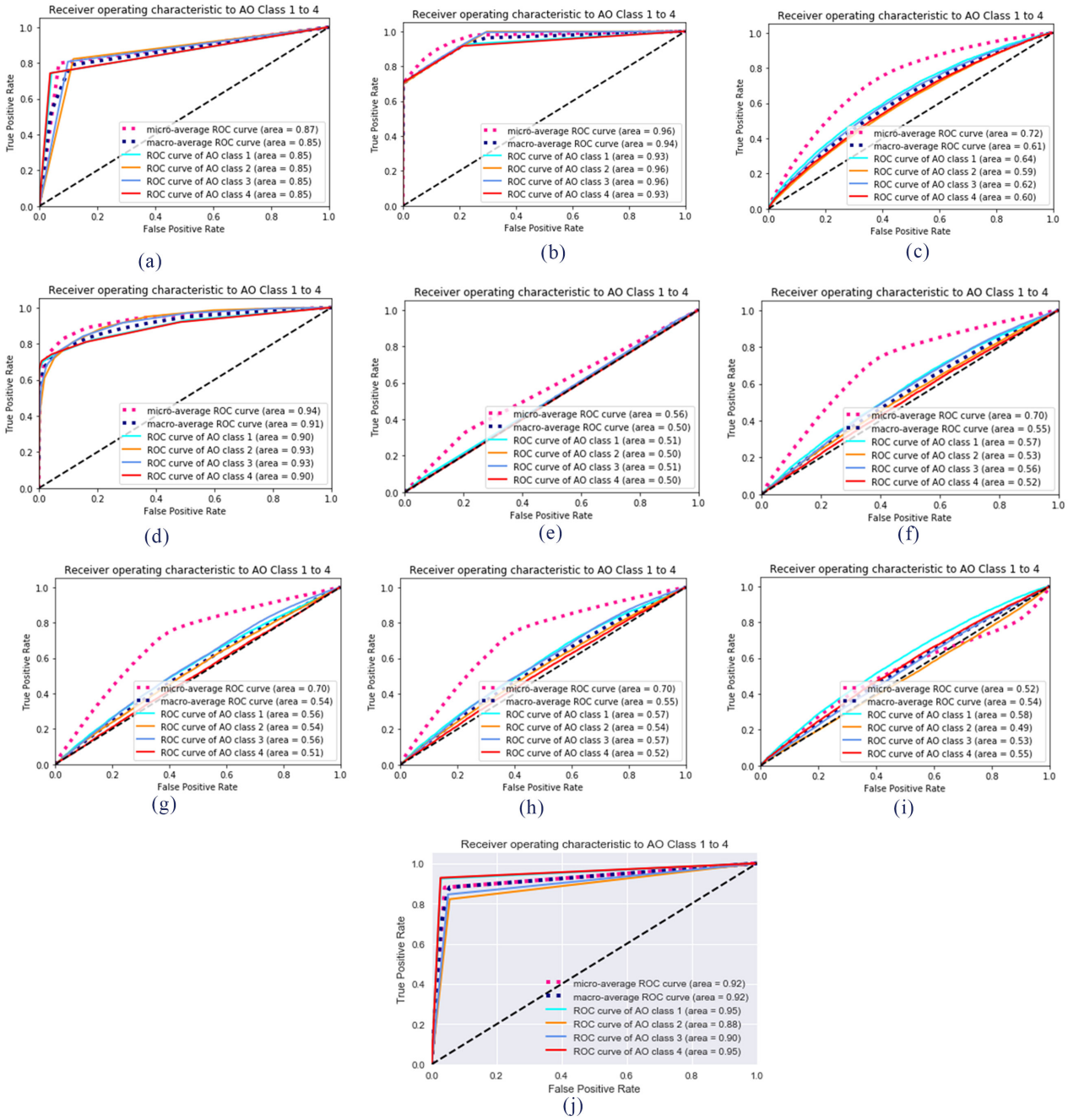

Fig 8. ROC analysis on prediction of models: (a) DT, (b) KNN, (c) MLP, (d) RF, (e)SGD, (f) NB, (g) LR, (h) LDA, (i) QDA, (j) Proposed Method

The proposed Extra-tree classifier based socio-economic factor selection has selected optimal factors such as AH, EQHHH, HS, HPF, HPQ, EQHQ, NOD, ALU(ACRE), FRT, LS, LT(I), LT(NI), AvailGS, LL and UF. While extracting same no. of socio-economic factors by using Logit model, the obtained optimal list of factor is FD(RC), FD(KC), CF, FD(V), WRF, FarmTools, CRS, CA, TLU, AE, ICS, FT, PTO, EQHHH and ATI. Similarly, in the Probit model, the selected optimal socio-economic

factors are FD(RC), FD(KC), CF, FD(V), WRF, FarmTools, CA, CRS, TLU, AE, ICS, FT, PTO, EQHHH and ATI. Here, an equal number of socio-economic factors are considered while the evaluation of models. Table 7 summarizes three major comparisons: i) performance of machine learning models with the optimal list of social-economic factors obtained through Logit model (Table 4); ii) performance of machine learning models with the optimal list of social-economic factors obtained through Probit model (Table 4); and iii) performance of machine learning models with the optimal list of social-economic factors obtained through proposed Extra tree classifier based model (Table 4). Further, the proposed ensemble learning model based prediction of AO is found better than other counterparts. The accuracy of the proposed method is 88%, which is marginally best than the other models. Table 8 represents the performance comparison of the proposed model (Algorithm 1) for prediction of AO type with Extra-tree model based socio-economic factor selection (Algorithm 2) with other machine learning models. Figure 9 represents the performance of the proposed prediction model with number of estimators.

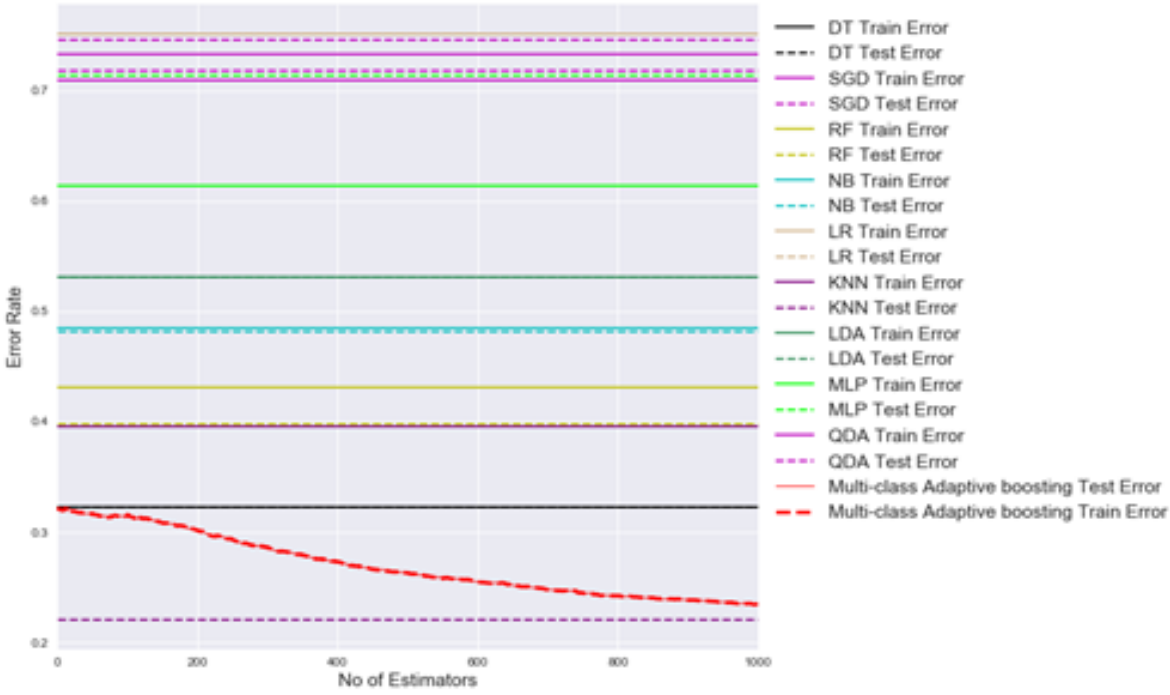

Fig 9. Performance of the model with number of estimators

Table 5. Class-wise performance

Model	Class-wise Metrics		Performance Metrics		
			Precision	Recall	F1-Score
RF	Class-wise Metrics	AO Class 1	0.84	0.71	0.77
		AO Class 2	0.77	0.88	0.82
		AO Class 3	0.81	0.79	0.80
		AO Class 4	0.94	0.68	0.79
	Accuracy	Macro Average	0.84	0.77	0.80
		Weighted Average	0.81	0.81	0.80
KNN	Class-wise Metrics	AO Class 1	0.91	0.71	0.80
		AO Class 2	0.79	0.87	0.83
		AO Class 3	0.80	0.82	0.81
		AO Class 4	0.93	0.71	0.81
	Accuracy	Macro Average	0.86	0.78	0.81
		Weighted Average	0.82	0.82	0.82

Continued on next page

Table 5 continued

DT	Class-wise Metrics	AO Class 1	0.72	0.74	0.73
		AO Class 2	0.83	0.82	0.83
		AO Class 3	0.81	0.81	0.81
		AO Class 4	0.71	0.74	0.73
	Accuracy	Macro Average	0.77	0.78	0.77
		Weighted Average	0.80	0.80	0.80
MLP	Class-wise Metrics	AO Class 1	0.36	0.01	0.02
		AO Class 2	0.46	0.70	0.56
		AO Class 3	0.45	0.48	0.46
		AO Class 4	0.33	0.01	0.03
	Accuracy	Macro Average	0.40	0.30	0.27
		Weighted Average	0.43	0.46	0.40
NB	Class-wise Metrics	AO Class 2	0.46	0.76	0.57
		AO Class 3	0.46	0.41	0.44
		AO Class 4	0.87	0.01	0.02
	Accuracy	Macro Average	0.66	0.30	0.26
		Weighted Average	0.56	0.46	0.39
Proposed Ensemble Model	Class-wise Metrics	AO Class 1	0.00	0.00	0.00
		AO Class 2	0.43	0.70	0.53
		AO Class 3	0.39	0.37	0.38
		AO Class 4	0.00	0.00	0.00
	Accuracy	Macro Average	0.21	0.27	0.23
		Weighted Average	0.31	0.42	0.35
	Class-wise Metrics	AO Class 1	0.91	0.92	0.92
		AO Class 2	0.83	0.82	0.83
		AO Class 3	0.85	0.85	0.85
		AO Class 4	0.92	0.93	0.92
	Accuracy	Macro Average	0.88	0.88	0.88
		Weighted Average	0.88	0.88	0.88

Table 6. Class-wise performance

Model	Class-wise Metrics		Performance Metrics		
			Precision	Recall	F1-Score
SGD	Class-wise Metrics	AO Class 1	0.15	0.03	0.05
		AO Class 2	0.47	0.00	0.00
		AO Class 3	0.35	0.98	0.51
		AO Class 4	0.00	0.00	0.00
	Accuracy	Macro Average	0.24	0.25	0.14
		Weighted Average	0.33	0.34	0.18
LR	Class-wise Metrics	AO Class 1	0.00	0.00	0.00
		AO Class 2	0.43	0.77	0.55
		AO Class 3	0.41	0.30	0.34
		AO Class 4	0.00	0.00	0.00
	Accuracy	Macro Average	0.21	0.27	0.22
		Weighted Average	0.32	0.42	0.35
LDA	Class-wise Metrics	AO Class 1	0.00	0.00	0.00
		AO Class 2	0.43	0.79	0.55
		AO Class 3	0.40	0.28	0.33
		AO Class 4	0.00	0.00	0.00
	Accuracy	Macro Average	0.21	0.27	0.22

Continued on next page

Table 6 continued

		Weighted Average	0.32	0.42	0.34
QDA	Class-wise Metrics	AO Class 1	0.17	0.26	0.21
		AO Class 2	0.42	0.16	0.24
		AO Class 3	0.36	0.60	0.45
		AO Class 4	0.15	0.11	0.12
	Accuracy	Macro Average	0.28	0.28	0.26
		Weighted Average	0.34	0.32	0.30
Proposed Ensemble Model	Class-wise Metrics	AO Class 1	0.91	0.92	0.92
		AO Class 2	0.83	0.82	0.83
		AO Class 3	0.85	0.85	0.85
		AO Class 4	0.92	0.93	0.92
	Accuracy	Macro Average	0.88	0.88	0.88
		Weighted Average	0.88	0.88	0.88

Table 7. Performance comparison of proposed Socio-economicFactor Selection (Algorithm 1) with Logit and Probit Method

Socio-economic Factor Selection Model	Tested Model	Performance Metrics				
		Precision	Recall	F1-Score	ROC-AUC	Accuracy
Logit Model[17,19]	RF	0.560	0.560	0.560	0.654	56.05
	KNN	0.566	0.566	0.566	0.659	56.69
	DT	0.564	0.564	0.564	0.660	56.44
	MLP	0.425	0.425	0.425	0.519	42.51
		0.416	0.416	0.416	0.508	41.63
	SGD	0.253	0.253	0.253	0.504	25.34
	LR	0.416	0.416	0.416	0.508	41.66
	LDA	0.416	0.416	0.416	0.508	41.65
	QDA	0.318	0.318	0.318	0.500	31.81
Probit Model[21]	RF	0.560	0.560	0.560	0.655	56.09
	KNN	0.566	0.566	0.566	0.659	56.68
	DT	0.564	0.564	0.564	0.660	56.43
	MLP	0.427	0.427	0.427	0.513	42.79
		0.416	0.416	0.416	0.508	41.63
	SGD	0.253	0.253	0.253	0.504	25.34
	LR	0.416	0.416	0.416	0.508	41.66
	LDA	0.416	0.416	0.416	0.508	41.65
	QDA	0.166	0.166	0.166	0.499	16.60
Proposed Model for Socio-economic Factors Selection (Algorithm 1)	RF	0.805	0.805	0.805	0.844	80.58
	KNN	0.817	0.817	0.817	0.853	81.71
	DT	0.797	0.797	0.797	0.852	79.77
	MLP	0.456	0.456	0.456	0.538	45.64
	NB	0.419	0.419	0.419	0.514	41.91
	SGD	0.340	0.340	0.340	0.501	34.08
	LR	0.422	0.422	0.422	0.512	42.21
	LDA	0.422	0.422	0.422	0.512	42.23
	QDA	0.314	0.314	0.314	0.519	31.40

Table 8. Performance comparison of Proposed AO Prediction model (Algorithm 2) with other machine learning models

Socio-economic Factor Selection Model	Tested Model	Performance Metrics				
		Precision	Recall	F1-Score	ROC-AUC	Accuracy
Logit Model[17,19]	RF	0.560	0.560	0.560	0.654	56.05
	KNN	0.566	0.566	0.566	0.659	56.69
	DT	0.564	0.564	0.564	0.660	56.44
	MLP	0.425	0.425	0.425	0.519	42.51
		0.416	0.416	0.416	0.508	41.63
	SGD	0.253	0.253	0.253	0.504	25.34
	LR	0.416	0.416	0.416	0.508	41.66
	LDA	0.416	0.416	0.416	0.508	41.65
	QDA	0.318	0.318	0.318	0.500	31.81
Probit Model[21]	RF	0.560	0.560	0.560	0.655	56.09
	KNN	0.566	0.566	0.566	0.659	56.68
	DT	0.564	0.564	0.564	0.660	56.43
	MLP	0.427	0.427	0.427	0.513	42.79
		0.416	0.416	0.416	0.508	41.63
	SGD	0.253	0.253	0.253	0.504	25.34
	LR	0.416	0.416	0.416	0.508	41.66
	LDA	0.416	0.416	0.416	0.508	41.65
	QDA	0.166	0.166	0.166	0.499	16.60
Proposed Model for Socio-economic Factors Selection (Algorithm 1)	Proposed AO Prediction Model (Algorithm 2)	0.878	0.878	0.878	0.919	87.86

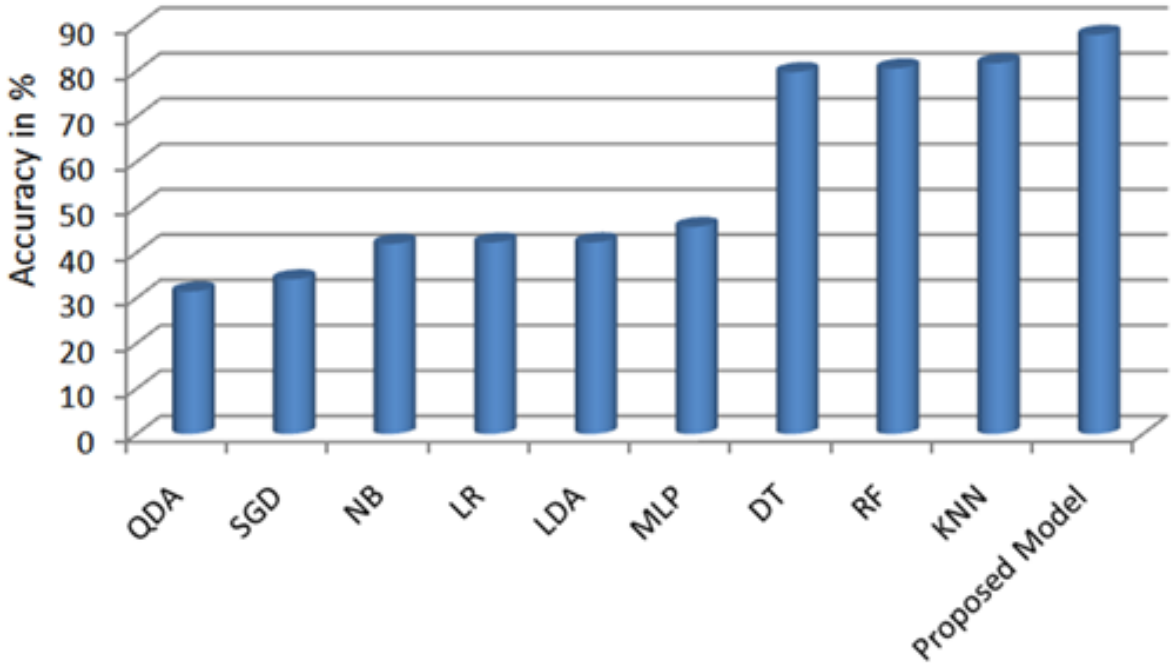

Fig 10. Overall performance comparison

The overall comparative analysis has been represented in Figure 10. Here the data has been split into 70% and 30 % using the stratified sampling method and the performance has been shown for the methods such as DT, QDA, MLP, SGD, NB, LR,

LDA, RF, KNN, and the proposed method. It is worthy to note that from all the result analysis, the performance of the proposed model is superior to all the other models.

It is observed that the proposed Extra-tree learning and multi-class adaptive boosting meta-estimator based socio-economic factor analysis model is found better for analyzing and predicting agricultural productivity. However, it requires large and complex computation as compared to Logit and Probit model. On the other hand, the Logit and Probit model is inherently better for the identification of the correlation between socio-economic factors. But, the Logit model has the limitation of representing random variation for the unobserved factors, and the Probit model has the issues of temporally correlated errors. The simulation results show that the proposed approach has better performance in the prediction of agricultural productivity.

5 Conclusion

Although Probit and Logit model for factor analysis and its application to socio-economic factor analysis has been found suitable to infer the relationships between socio-economic factors (independent variables) and agricultural productivity (dependent variable), it is found poor in terms of prediction of agricultural productivity (target variable). The machine learning model based on socio-economic factors selection by using Extra trees classifier is found capable to prune out the socio-economic factors that contain relevant information to the target variable agricultural productivity. However, this approach of socio-economic factor selection requires heavy and complex computation as compared to the Probit and Logit based model. In socio-economic study, usually, the data collected from respondents are highly unstructured and random. Hence, relying on a single model prediction is not sufficient to make a decision. Here in this study, an ensemble meta-learner has been used; which is a form of meta-learning that constructs a higher-level prediction model over the predictions of considered base classifiers. This ensemble learning-based approach is found better in terms of agricultural productivity prediction.

This work may be a framework for the further study of socio-economic factors and supplement to the existing knowledge base for agricultural research in India and abroad, particularly in the area of agricultural productivity analysis. Further, it can be used as a system identification model for the identification of various social-economic factors of respondents (farmers) and the evaluation of these factors towards sustainable agricultural productivity. The expected outcome of this project may be a data-driven operational system for evaluation of socio-economic factors that influence sustainable agricultural productivity in India and can be extended to further study in Abroad.

Acknowledgment: This research work is supported by Science and Engineering Research Board (SERB), Department of Science and Technology (DST), New Delhi, Govt. of India, under the research project Grant No. EEQ/2017/000355.

Conflict of interest: The authors declare that this manuscript has no conflict of interest with any other published source and has not been published previously (partly or in full). No data have been fabricated or manipulated to support our conclusions.

References

1) Sengupta S. The food chain in fertile India, growth outstrips agriculture. 2008. Available from: http://www.nytimes.com/2008/06/22/business/22indiafood.html.
2) Fusi A, González-García S, Moreira MT, Fiala M, Bacenetti J. Rice fertilised with urban sewage sludge and possible mitigation strategies: an environmental assessment. *Journal of Cleaner Production*. 2017;140:914–923. Available from: https://dx.doi.org/10.1016/j.jclepro.2016.04.089.
3) Pingali PL, Roger PA. Impact of pesticides on farmer health and the rice environment. In: and others, editor. Springer Science & Business Media. Springer. 2012. Available from: https://doi.org/10.1007/978-94-011-0647-4_1.
4) Tuong TP, Bouman BA. Rice production in water-scarce environments. In: and others, editor. Water productivity in agriculture: limits and opportunities for improvement;vol. 1. 2003;p. 13–42. Available from: https://EconPapers.repec.org/RePEc:ags:iwmibo:138054.
5) Wassmann R, Jagadish SV, Sumfleth K, Pathak H, Howell G, Ismail A, et al. Regional vulnerability of climate change impacts on Asian rice production and scope for adaptation. *Advances in agronomy*. 2009;102:91–133. Available from: https://doi.org/10.1016/S0065-2113(09)01003-7.
6) Masutomi Y, Takahashi K, Harasawa H, Matsuoka Y. Impact assessment of climate change on rice production in Asia in comprehensive consideration of process/parameter uncertainty in general circulation models. *Agriculture, Ecosystems & Environment*. 2009;131:281–291. Available from: https://dx.doi.org/10.1016/j.agee.2009.02.004.
7) Greig L. An Analysis of the Key Factors Influencing Farmer's Choice of Crop, Kibamba Ward, Tanzania. *Journal of Agricultural Economics*. 2009;60(3):699–715. Available from: https://dx.doi.org/10.1111/j.1477-9552.2009.00215.x.
8) Nkonya E, Schroeder T, Norman D. Factors affecting adoption of improved maize seed and fertiliser in Northern Tanzania. *Journal of Agricultural Economics*. 1997;48(1-3):1–12. Available from: https://dx.doi.org/10.1111/j.1477-9552.1997.tb01126.x.
9) Feuerstein H. Factors affecting farmland prices in schleswig-holstein (Germany) - an econometric analysis from 1954 to 1968. *Journal of Agricultural Economics*. 1974;25(1):65–76. Available from: https://dx.doi.org/10.1111/j.1477-9552.1974.tb00527.x.
10) Osuntogun A. An econometric analysis of some factors influencing the loyalty of members in the Western Nigeria marketing co-operatives. *Journal of Agricultural Economics*. 1972;23(3):299–310. Available from: https://dx.doi.org/10.1111/j.1477-9552.1972.tb01453.x.
11) Gómez-Limón JA, Riesgo L, Arriaza M. Multi-Criteria Analysis of Input Use in Agriculture. *Journal of Agricultural Economics*. 2004;55(3):541–564. Available from: https://dx.doi.org/10.1111/j.1477-9552.2004.tb00114.x.
12) Hamade K, Malorgio G, Midmore P. Contrasting Quantitative and Qualitative Approaches to Rural Development Analysis: The Case of Agricultural Intensification in Lebanon. *Journal of Agricultural Economics*. 2015;66(2):492–518. Available from: https://dx.doi.org/10.1111/1477-9552.12095.

13) Yuguda RM, Girie AA, Dire B, Salihu M. Socio-economic factors and constraints influencing productivity among Cassava farmers in Taraba state Nigeria. *International Journal of Advances in Agricultural Science and Technology*. 2013;1(1):1–15.
14) Ajah J, Nmadu JN. Socio-economic factors influencing the output of small-scale maize farmers in. *Kasetsart Journal of Social Sciences*. 2012;33(2):333–341.
15) Masuku MB, Xaba B. Factors Affecting the Productivity and Profitability of Vegetables Production in Swaziland. *Journal of Agricultural Studies*. 2013;1(2):37–37. Available from: https://dx.doi.org/10.5296/jas.v1i2.3748.
16) Zalkuwi J. Socio-economic factors that affect Sorghum production in Adamawa State. *Nigeria International Journal of Science and Research (IJSR)*. 2015;4(2):1610–1614. Available from: https://www.ijsr.net/search_index_results_paperid.php?id=SUB151542.
17) Usman MA, Dodo H. Socio-Economic Factors Influencing Agricultural Insurance in Rice Production in Kano State, Nigeria. In: International Conference on Advances in Agricultural, Biological & Environmental Sciences (AABES-2014). Nigeria. 2014. Available from: http://dx.doi.org/10.15242/IICBE.C1014147.
18) U EGW, Williams E. Factors affecting sustainable agricultural productivity in Ebonyi State, Nigeria. *Global Journal of Agricultural Economics, Extension and Rural Development*. 2015;21(30):183–187.
19) Anigbogu TU, Agbasi OE, Okoli IM. Socioeconomic Factors Influencing Agricultural Production among Cooperative Farmers in Anambra State, Nigeria. *International Journal of Academic Research in Economics and Management Sciences*. 2015;4(3):43–58. Available from: https://dx.doi.org/10.6007/ijarems/v4-i3/1876.
20) Jiriko RK. Socio-economic factors affecting the performance of women in food production. *Global J Agr Res*. 2015;3(2):37–45.
21) Mittal S, Mehar M. Socio-economic Factors Affecting Adoption of Modern Information and Communication Technology by Farmers in India: Analysis Using Multivariate Probit Model. *The Journal of Agricultural Education and Extension*. 2016;22:199–212. Available from: https://dx.doi.org/10.1080/1389224x.2014.997255.
22) Office of the Registrar General & Census Commissioner I. 2011. Available from: http://agriodisha.nic.in/http_public/status%20of%20agriculture%20in%20orissa.aspx.
23) Office of the Registrar General & Census Commissioner I. Executive summary of Census in Odisha. 2011. Available from: http://www.censusindia.gov.in/2011census/PCA/PCA_Highlights/pca_highlights_file/Odisha/Executive_Summary.pdf.
24) Chambers R. The origins and practice of participatory rural appraisal. *World Development*. 1994;22(7):953–969. Available from: https://dx.doi.org/10.1016/0305-750x(94)90141-4.
25) Chambers R. Participatory rural appraisal (PRA): Analysis of experience. *World development*. 1994;22(9):1253–1268. Available from: https://doi.org/10.1016/0305-750X(94)90003-5. doi:10.1093/nar/22.2.124.
26) Chambers R. Participatory rural appraisal (PRA): Challenges, potentials and paradigm. *World Development*. 1994;22(10):1437–1454. Available from: https://doi.org/10.1016/0305-750X(94)90030-2.
27) Dougill AJ, Fraser EDG, Holden J, Hubacek K, Prell C, Reed MS, et al. Learning from Doing Participatory Rural Research: Lessons from the Peak District National Park. *Journal of Agricultural Economics*. 2006;57(2):259–275. Available from: https://dx.doi.org/10.1111/j.1477-9552.2006.00051.x.
28) Abrha BK. Factors affecting agricultural production in Tigray region, northern Ethiopia. . Available from: https://core.ac.uk/download/pdf/43177295.pdf.
29) Urgessa T. The determinants of agricultural productivity and rural household income in Ethiopia. *Ethiopian Journal of Economics*. 2015;24(2):63–91. Available from: https://www.ajol.info/index.php/eje/article/view/146625.
30) Cavestro L. PRA-participatory rural appraisal concepts methodologies and techniques. Padova University. Padova PD. Italia. 2003.
31) Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *The Journal of machine Learning research*. 2011;12:2825–2855. Available from: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.
32) Kent TJ. Information gain and a general measure of correlation. *Biometrika*. 1983;70(1):163–173. Available from: https://dx.doi.org/10.1093/biomet/70.1.163.
33) Lerman IR, Yitzhaki S. A note on the calculation and interpretation of the Gini index. *Economics Letters*. 1984;15(3-4):363–368. Available from: https://dx.doi.org/10.1016/0165-1765(84)90126-5.
34) Zhu J, Rosset S, Zou H, Hastie T. Multi-class AdaBoost. *Statistics and Its Interface*. 2009;2(3):349–360. Available from: https://dx.doi.org/10.4310/sii.2009.v2.n3.a8.
35) Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*. 1991;21(3):660–674. Available from: https://dx.doi.org/10.1109/21.97458.
36) Bhatia N. Survey of Nearest Neighbor Techniques. *Computer Vision and Pattern Recognition*. 2010;8(2):302–305. Available from: https://arxiv.org/abs/1007.0085.
37) Murphy KP. Naive bayes classifiers. University of British Columbia. 2006. Available from: https://www.ic.unicamp.br/~rocha/teaching/2011s1/mc906/aulas/naive-bayes.pdf.
38) Verikas A, Gelzinis A, Bacauskiene M. Mining data with random forests: A survey and results of new tests. *Pattern Recognition*. 2011;44:330–349. Available from: https://dx.doi.org/10.1016/j.patcog.2010.08.011.
39) Zhang GP. Neural networks for classification: a survey. In: and others, editor. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews);vol. 30 of 4. 2000;p. 451–462. Available from: https://doi.org/10.1109/5326.897072.
40) Erenguc SS, Koehler GJ. Survey of mathematical programming models and experimental results for linear discriminant analysis. *Managerial and Decision Economics*. 1990;11(4):215–225. Available from: https://dx.doi.org/10.1002/mde.4090110403.
41) Zanutto EL. A comparison of propensity score and linear regression analysis of complex survey data. *Journal of data Science*. 2006;4(1):67–91.
42) Tharwat A. Linear vs. quadratic discriminant analysis classifier: a tutorial. *International Journal of Applied Pattern Recognition*. 2016;3(2):145–145. Available from: https://dx.doi.org/10.1504/ijapr.2016.079050.
43) Clémençon S, Bertail P, Chautru E, Papa G. Survey schemes for stochastic gradient descent with applications to m-estimation. *Machine Learning*. 2015. Available from: https://arxiv.org/abs/1501.02218.