Elsevier required licence: © <2022>. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/
The definitive publisher version is available online at
[https://www.sciencedirect.com/science/article/pii/S0147651322001117?via%3Dihub]

Predicting sustainable arsenic mitigation using machine learning techniques

Sushant K. Singh[1, 2*], Taylor W. Robert[1], Biswajeet Pradhan[3,4], Ataollah Shirzadi[5], and Binh Thai Pham[6]

[1]Department of Earth and Environmental Studies, Montclair State University, New Jersey, USA

*E-mail: sushantorama@gmail.com

[2]Artficial Intelligence & Analytics | Health Care and Life Sciences, Virtusa Corporation, New York City, NY, USA (Current affiliation)

[3]Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), School of Information, Systems & Modelling Faculty of Engineering and IT, University of Technology Sydney, Australia

[4]Department of Energy and Mineral Resources Engineering, Sejong University, Choongmu-gwan, 209 Neungdong-ro Gwangjin-gu, Seoul 05006, Republic of Korea

[5]College of Natural Resources, Department of Rangeland and Watershed Management Sciences, University of Kurdistan, Sanandaj, Iran

[6]Department of Geotechnical Engineering, University of Transport Technology, 54 Trieu Khuc, Thanh Xuan, Ha Noi, VietNam

ABSTRACT: Several artificial intelligence techniques have been applied to developing an array of environmental prediction models for various environmental challenges. However, there is no such prediction models exist for sustainable arsenic mitigation technologies. This study evaluates the state-of-the-art artificial intelligence models such as linear, nonlinear, ensemble, tree-based, Naïve Bayes, and neural network machine learning classifiers in predicting the preference of the most sustainable arsenic mitigation technology and provides following insights: (a) which machine learning algorithm has the highest prediction accuracy and robustness, and (b) which machine learning models are best to fit socioeconomic-environmental data for developing prediction models of sustainable arsenic mitigation technology? We evaluated 19 machine learning models for their predictive accuracy and the robustness by comparing their overall prediction accuracy, precision, recall, and the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. A Gaussian distribution-based Naïve Bayes classifier outperformed the rest of the algorithms with the highest AUC of 0.825 on test data. The second two best models were Nu

Support Vector Classification (NuSVC) (AUC=0.800) (a radial basis function kernel-based support vector machine algorithm) and K-Neighbors (AUC=0.790). All the ensemble classifiers scored higher than 70% AUC, Random Forest being the top performer (AUC=0.769). We used only one tree-based classifier Decision Tree, and it produced promising results (AUC=0.769) after the three top classifiers. The neural network-based multilayer perceptron model, although ranked 9th position, also had a considerably good performance (AUC=0.748). Most linear classifiers did not perform well with the Ridge classifier at the top (AUC=0.727) and perceptron at the bottom (AUC=0.567). A Naïve Bayes-based classifier with Bernoulli distribution was the worst model (AUC=0.500). Socioeconomic, demographic, and psychological data may not be linearly associated with each other or with the outcomes. Therefore, nonlinear or ensemble classifiers could better understand these complex relationships and help develop the most accurate and robust prediction models. Gaussian NB is the best option for developing such prediction models on socioeconomic and psychological data with small sample size. The proposed methodological framework and the outcomes of the 19 machine learning models will help develop informed and intelligent research methods as well as in targeting the population who are ready to adopt sustainable arsenic mitigation technology.

Keywords: *Arsenic; Arsenic mitigation technologies; Machine learning; Linear classifier; Nonlinear classifier; Ensemble*

1. INTRODUCTION

Socioeconomic, demographic, psychological, and cultural aspects of the communities exposed to environmental contaminants, arsenic in this case, play a significant role in adopting and sustainably practicing mitigation technologies [1-4]. Groundwater arsenic contamination is a global environmental as well as a social threat to nearly 296 million individuals' lives in more than 100 countries, India and Bangladesh being the foremost victims [5, 6]. Arsenic is a human carcinogen, can adversely impact dermal, cardiovascular, respiratory, neurological, and genetic systems, and could lead to incurable varieties of cancers, if consumed for a prolonged period [5]. It is a well-known fact that arsenic concentration in those contaminated countries has increased multiple times of the standard norm of 10 µg/L stipulated by the World Health Organization (WHO) and the United Nations Food and Agricultural Organization's (FAO) standard of 100 µg/L for irrigation water [7]. Arsenic has also entered to the human, animal, and aquatic food chain evidenced from high concentrations of arsenic in rice, beans, pulses, vegetable, cereals, fruits, poultry, egg, fresh milk, milk powder, mother's milk, fish, shellfish, and algae [7-10]. There are studies that report a significant concentration of arsenic in the human urine, blood, hair, and nail samples, evidence of the arsenic exposure and accumulation in the human body [5, 11].

Low cost and simple arsenic mitigation techniques, such as arsenic treatment (filtration) units, deep tube wells, piped water supply system, and rainwater harvesting system have been the primary ways of providing arsenic-free water in the arsenic-contaminated areas globally [4, 12, 13]. However, because of the technical [4, 14-18], social, economic, and cultural challenges [1, 12-15, 19-22], these interventions could not be achieved sustainability. There are a fair number of studies that highlight the technical challenges of arsenic mitigation technologies, however, research on the socioeconomic, psychological, and cultural aspects of arsenic mitigation is still in the rudimentary stages [3, 19, 23-25]. Based on these handfuls of studies, the authors found that there is a lack of arsenic awareness and ownership of the implemented arsenic mitigation technologies; low willingness to pay for arsenic mitigation technologies; complicated operation and maintenance of arsenic mitigation technologies manuals; expensive technologies; long distance between the households and the arsenic-free water sources; and social resistance by a group of people to not let access the arsenic-free sources [1, 12-15, 19-22] have negatively impacted the sustainable adoption of arsenic mitigation technologies.

In some recent studies [19, 22], it was discovered that communities' trust in the local agencies and institutions as well as their social capital played a crucial role in their decision-making to adopt arsenic mitigation technologies. In other studies [19, 26, 27], the authors highlighted that people's perceived risk of health, income, and social discrimination due to arsenic contamination significantly impact their decision-making process to adopt arsenic mitigation technologies. The cost-effectiveness of a proposed arsenic mitigation technology also ensures their sustainable use by the beneficiaries [21].

Accurately capturing the socioeconomic, demographic, psychological, and cultural information of arsenic-affected communities is a challenging work for researchers, developing prediction models on these data is even more daunting [1]. The reasons are lack of empirical data, a complex relationship between the variables, socioeconomic, psychological, and cultural data are prone to multicollinearity, and lack of successful case studies [1]. These all may affect the selection of the most important predictors as in statistical analysis the model will only select the significant variables unless we enforce expert opinion and include the variables known to be important but not statistically significant [28]. In recent studies, the authors captured information on the socioeconomic, demographic, social trust and capital aspects from an arsenic-exposed community located in the middle-Ganga Plain of Bihar, India [19, 22]. While developing a logistic regression prediction model of the adoption of arsenic treatment units, the authors started with 19 statistically significant variables but end-up having eight variables in the final model with both significant and nonsignificant variables. The model accomplished an overall prediction accuracy of 80.2%, which looks promising [22]. However, since this model was not compared with other state-of-the art modeling techniques such as machine learning models, we cannot say this is the best socioeconomic model of predicting sustainable arsenic mitigation technologies. Also, considering the lack of such data, how a robust machine learning model can be developed that could help predict sustainable arsenic mitigation technologies in arsenic contaminated areas.

Several artificial intelligence techniques have been used in developing various prediction models on environmental data including landslide susceptibility [29-33], groundwater potential [34, 35], groundwater vulnerability [36], and groundwater contaminations [37, 38]. Pertaining to arsenic research, various machine learning algorithms have been used in predicting arsenic contamination in groundwater using the physical, chemical, hydrogeological, and topographical data [39-43].

However, to the best of our knowledge, these state-of-the-art technologies have never been applied to develop a socioeconomic model of arsenic mitigation. In a recent study by Singh et al. (2018), the authors have used various machine learning models including logistic regression (LR), support vector machine (SVM), decision trees (DT), k-nearest neighbor (k-NN), naïve Bayes (NB), and random forests (RF) to predict arsenic awareness as a function of various socioeconomic, sanitation, socio-behavioral, and social trust factors captured through an empirical study. In this study, the authors discovered that arsenic awareness is a nonlinear classification problem and the SVM and RF appeared to be the most appropriate machine learning algorithms in correctly classifying arsenic awareness [1]. The authors further suggested that survey-based complex environmental data may require advanced computational techniques opposed to traditional statistical approach for developing accurate and robust prediction models.

Therefore, this study is a founding step in filling the above research gaps through answering following questions: (a) which machine learning algorithm can achieve the highest prediction accuracy and robustness in predicting the preference of sustainable arsenic mitigation technology and (b) whether prediction of the preference of arsenic mitigation technology is a linear or a non-linear classification challenge?

2. METHODS

2.1. Study Area

The state Bihar, the study area, is the second worst arsenic affected states of India after West Bengal, which shares its geographical boundaries with other arsenic impacted regions including Uttar Pradesh state of India, Bangladesh, Nepal, and Tibet [7]. The groundwater used for drinking purposes is contaminated with elevated levels of arsenic in over 50% of the districts of Bihar. Groundwater used for irrigation is also found to be contaminated with arsenic in some areas along with a considerable amount of arsenic in agricultural soils and food materials [44-47]. Elevated levels of arsenic in urine, blood, hair, and nail samples are also detected and several arsenicosis victims are also diagnosed in the state [48-51]. Socioeconomic, health, and psychological aspects of arsenic in the study area are also investigated, but still confined in a few geographical regions of the state [19, 24, 25, 44, 48, 50, 51].

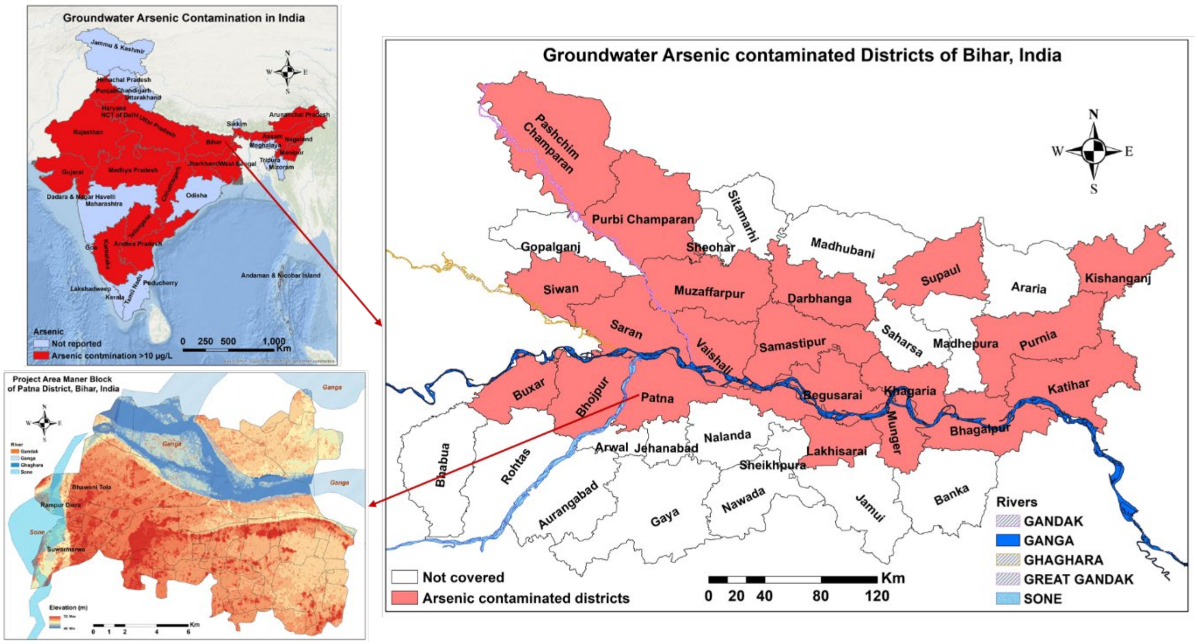


Figure 1. A map showing the arsenic affected districts of Bihar and the three villages selected for this study with their elevations.

In this study, we developed a five steps methodological framework to achieve our goals including: (1) data collection and storage, (2) data pre-processing, (3) data visualization, (4) model development and validation, and (5) prediction and recommendation.

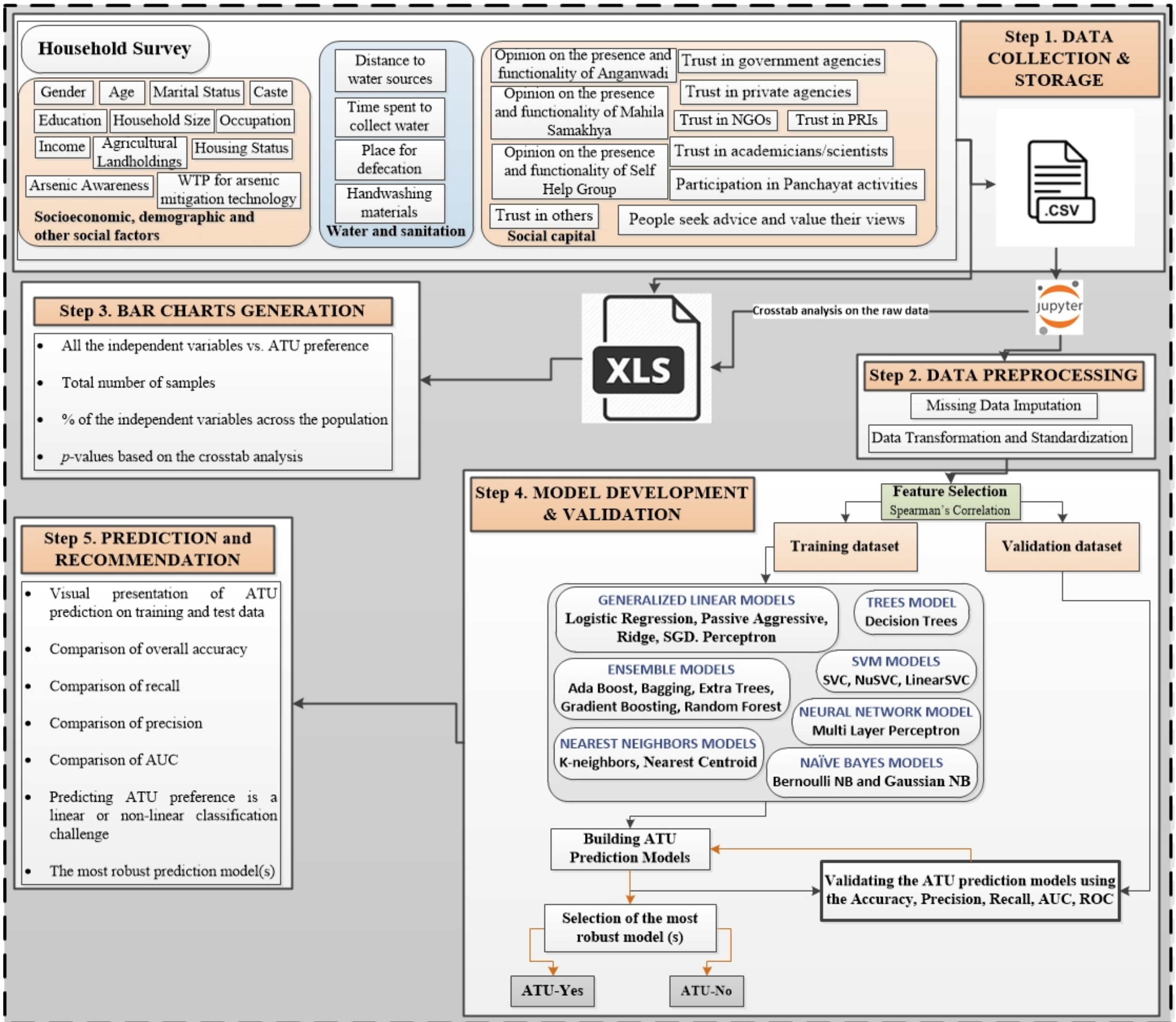


Figure 2. Methodological framework adopted for this study.

2.2. Data

The data was captured by interviewing 340 households, randomly selected and stratified by their caste, through a structured questionnaire in three villages Suarmarwa, Rampur Diara, and Bhawani Tola: all located in the severely arsenic contaminated block Maner of Patna district of Bihar, India. Survey methodological details are explained in Singh (2015). The socioeconomic and demographic information were captured through asking questions on gender, age, marital status, caste, education, household size, occupation, income, agricultural landholdings, and housing status. Water and sanitation behaviors were captured through asking questions on a number of households involved in water collection, distance travelled and time spent to collect water, place for defecation, and materials used for hand washing after defecation. Social capital and trust were captured through asking questions on communities' opinion on the presence and functionality of Anganwadi, Mahila Samakhya, Self-Help Group; trust in others, government agencies, NGOs, Panchayat Raj Institutions, private agencies, academicians and scientists; participation in PRI's activities, and whether people seek advice and value their views. Arsenic awareness was captured through 10 questions converted to arsenic awareness index (low awareness vs. high awareness). A detailed analysis on arsenic awareness is explained in [1]. Willingness to pay for arsenic mitigation technology was also captured through a structured question. Communities' preference for sustainable arsenic mitigation technologies was recorded through a structured questionnaire with options of arsenic treatment unit (ATU), piped water supply systems, deep tube wells, dug wells/open wells, and rainwater harvesting system. An in-depth analysis on these technological preferences is available in [22].

This study provides a comprehensive analysis of the most preferred sustainable arsenic mitigation technology (ATU) and investigates how the state-of-the-art machine learning technologies can efficiently predict communities' preference of sustainable mitigation technology. The survey data was transferred to a .excel file for frequency graphs generation and to a .csv file for data pre-processing, statistical analysis, and machine learning model development using Jupyter Notebook version 6.0.1 web application and Python 3 [52].

2.3. Data pre-processing

A majority of the variables were categorical and captured at the Likert-scale of five. Because of their imbalanced frequency distribution across the responses (strongly disagree to strongly agree), we reconstructed and recoded them for further analysis and model development. The transformation of the original categories of the variables to new categories is explained in SI-1. After screening the data, we found that one household did not answer the question on preference of sustainable arsenic mitigation technology; therefore, we had a total 339 samples. Using Pandas Python library, we imported the data to Jupyter Notebook and put it into a data frame for further analysis [53]. Scikit-learn, an open access machine learning library in Python programming language, was used for further analysis [54]. We also found 17 missing data that was imputed using the mode of each feature. A contingency analysis was performed between all the independent variables and ATU preference. Scipy library was used for all statistical analysis including Spearman's correlation [55].

2.4. Data visualization

The results of contingency analysis were graphically presented in Figure 2-4 in the result section using Excel Spreadsheet [56]. The graphs contain four important information including the number of data points, categories of each feature wherever applicable, percentage of different categories of features, and *p-value* to determine whether the responses across the categories were significantly different from each other.

2.5. Machine learning algorithm selection

Applying machine learning to efficiently predict communities' preference of sustainable arsenic mitigation technology is inspired by two recent researches where the first study [22] models the preferences using a traditional statistical technique logistic regression, but did not provide a comparison of how other statistical or machine learning techniques may fit the data to the various models. The second study [1] provides a great deal of insights that developing prediction models in the context of arsenic using complex socioeconomic, demographic, and other social factors may need very specific type of algorithms or a hybrid model. Applying all the algorithms together and comparing them in one study is not feasible therefore, we decided to select the state-of-the-art linear, nonlinear, ensemble, Naïve Bayes, and tree-based classifiers to develop the models and compare them for their prediction accuracy and robustness. A brief description of each algorithm applied in this study is described below.

2.5.1. Generalized Linear Models

Logistic Regression (LR)

The LR is a multivariate regression that provides the probability of the presence of an event at each response according to the predictors [57, 58]. It has some advantages that environmental researchers have encouraged to apply it, including; (1) the LR does not need to set normality for independent feature, (2) predictors can either be continuous or discrete or any combination of these types of data, (3) it is easy to implement in most statistical packages such as SPSS, SAS, STATA, R and so on [59-61]. The dependent variable in the LR should be binary (present/occurrence and absent/non-occurrence of an event) to achieve the probability values. In this study, we aim to predict the probability of adoption of arsenic treatment units (ATU) using LR model and some predictors. The LR can be formulated in its simplest form as follows:

$$P_{LR} = \frac{e^z}{1+e^z} \tag{1 }$$

where, P_{LR} is the probability of present/occurrence of an event that varies from 0 to 1as s-shaped curve, Z is a linear combination that varies from $-\infty$ to $+\infty$ and can be computed as bellow:

$$Z = c_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n \tag{2}$$

where, c_0 is the constant coefficient/intercept of the LR model, n is the number of predictors, a_i $(i = 1, 2, 3, \ldots, n)$ is the independent variables as input to the model, x_i $(i = 1, 2, 3, \ldots, n)$ is the coefficients of each predictors as input to the model [62].

Passive Aggressive (PA)

The passive-aggressive algorithm is similar to Perceptron as there is no learning rate required. However, it contain a regularization parameter "C" [54, 63]. Scikit learn machine learning library in python offers a passive aggressive classifier for binary classification such as in this case for classifying preference of ATU as a most sustainable arsenic mitigation technology. This technique is less explored though.

Ridge

The Ridge classifier is another linear classifier and is also known as least squares support vector machines where in Scikit learn this classifier first converts binary targets ATU-No and ATU-Yes to, respectively -1 and 1. The algorithm then regress the dependent variable against the independent variables, and the predicted class resembles to the sign of the regressor's prediction [54].

Stochastic Gradient Descent (SGD)

The SGD is another linear classifier and commonly used with large sample size. The SGD can fit both logistic regression model and support vector machine by selecting appropriate loss functions, respectively "log" and "hinge" [54]. It requires a learning rate, and the loss is estimated for each data point at a time. For a normally distributed data, the SGD provides a better result. Although SGD classifier is known for its efficiency on large datasets, it is known to be highly sensitive to feature scaling [54].

Perceptron

The Perceptron is another linear classifier that doesn't require learning rate, that's why it is a favorable classifier for large scale learning. Additionally, it doesn't penalize the learning and updates the model only on mistakes [54].

2.5.2. Trees Models

Decision Trees (DT)

The DT is a tree-base non-parametric classifier that learns decision rules from the data features. This is a very popular classifier for developing binary classification models as it is simple to understand and to interpret. Although it doesn't work well with missing data, it is a good classifier that efficiently handles both numerical and categorical data, and requires less assumptions [54].

2.5.3. Ensemble Models

Ada Boost

The AdaBoost is an ensemble classifier, first introduced in 1995 by Freund and Schapire. It fits a sequence of weak learners on modified versions of the data and produces a combined predicted class. In this entire process, the classifier makes sure that none of the data point is left in the training phase [54, 64].

Bagging

Bagging is another ensemble classifier that consolidates a final prediction based on the previous predictions on randomly selected subsets of the original training dataset. It works well with strong and complex models [54, 65].

Extra Trees (ET)

The ET is another ensemble model, well known to control over-fitting but less explored. In scikit-learn, this classifier fits several randomized decision trees i.e. "extra-trees" on several sub-samples of the dataset [54].

Gradient Boosting (GB)

The GB is another ensemble classifier offered by Scikit-learn library and is very popular among scientific computation community [54]. It builds an additive model in a forward stage-wise fashion and allows for the optimization of arbitrary differentiable loss functions.

Random Forest (RF)

The RF is another popular ensemble classifier available through Scikit-learn that fits several decision tree classifiers on many sub-samples of the dataset. The RF averages the probabilistic prediction value of each decision-tree and uses to improve the classification accuracy. It is also known to be prone to over-fitting [54, 66].

2.5.4. Support Vector Machines Models

Support vector machines (SVM) is known for efficiently classifying linearly separable data as well as non-linearly separable data by using a kernel function, such as sigmoid, radial, or polynomial. It is advantageous if the data is clearly separable and the ratio between the number of dimensions and the number of samples is greater. It is also memory efficient. However, the SVM takes more time in training the model therefore; it is not feasible for large data set. Likewise, with noisy data, the performance is poor.

SVC

The SVC is known as C-Support Vector Classification, a non-linear SVM classifier [54].

Nu-Support Vector Classification (Nu-SVC)

The Nu-SVC is another non-linear SVM classifier that uses a parameter to control the number of support vectors [54].

Linear Support Vector Classification (LinearSVC)

The LinearSVC uses ‘linear’ kernel and Scikit-learn library offers suppleness in choosing loss functions and penalties [54].

2.5.5. Nearest Neighbors Models

K-nearest neighbors

k-Nearest Neighbors (k-NN) is a nonparametric non-linear classifier that categorizes the intended event by a majority vote of its *k* nearest neighbors, which is a positive integer and can be derived through elbow-test [54, 67]. k-NN is advantageous in several ways, as it does not require assumptions, easily interpretable, good for nonlinear data, and can produce comparatively better accuracy [54, 67]. However, k-NN is very sensitive to irrelevant features, the scale of the variables, the dimensions of the dataset, and class imbalance. In addition, it can be highly computation intensive as it stores all the training data.

Nearest Centroid (NC)

In Scikit-learn, the NC classifier belongs to the nearest neighbor algorithms where each class is characterized by the centroid of its members [54].

2.5.6. Neural Network Model

Multi-Layer Perceptron

The ANN is organized and structured based on the skills of human brain cells to extract knowledge from the input dataset [32, 68]. It has some advantages that it has been a strong and promising technique to prediction environmental problems including, (i) it can efficiently detect a different subset of data within a whole dataset, (ii) it do not need to any experience and pre-knowledge process, and (iii) It do not need to a given statistical model in the training dataset [69]. The multilayer perceptron (MLP) and radial base function (RBF) are two popular and well-known as functions of ANN. Although the capability of these two function are different from one case study to another, in general the MLP is more popular and general than the RBF kernel function [70]. The MLP is more successfully and flexibility in modeling, especially on non-linear, imprecise and imperfect data so that it can extract the reliable results [71]. Therefore, in this study, we used of MLP function to construct a network for determining the relationship between the ATU and predictors. The MLP has a structure with three layers including a n input, an output and one or more hidden layers between them [72, 73]. In this study, ATU predictors are taking into consideration as inputs (neurons) and the weights for each predictor is output. In a simplest form, let x_i and w_i are input predictors and they obtained weights during the modeling process. In hidden layer, they are multiplied and then summed up to extract the final output or the final weights (y_i) by a non-linear activation function as follows:

$$\text{net} = \sum_{i=0}^{n} w_i x_i \tag{3}$$

$$y_i = f(\text{net}) \tag{4}$$

2.5.7. Naïve Bayes Models

Naïve Bayes (NB) is another state-of-the-art nonlinear classifier that works on the Bayes theorem, has a strong assumption that all the predictors are independent and not correlated to each other, and can be mathematically presented as below:

$$\mathit{Posterior\ Probability}[Y = P(c|x)] = \frac{[\mathit{Likelihood}: P(x_1|c) \times P(x_2|c) \times \ldots P(x_n|c)] \times [\mathit{Class\ Prior\ Probability}: P(c)]}{\mathit{Predictor\ Prior\ Probability}: P(x)} \tag{5}$$

NB is known to be outperforming other state-of-the-art classifiers, such as logistic regression [54]. It requires less training dataset and can quickly predict on test dataset. It is also known to be a good classifier for categorical variables [54]. However, the strong assumption of independence could be a challenge while applying NB on the dataset with multicollinearity.

Bernoulli Naïve Bayes

This is one of the NB classifiers that assumes the data has multivariate Bernoulli distributions. Therefore, this class requires samples to be represented as binary-valued feature vectors; if handed any other kind of data, a BernoulliNB instance may binarize its input (depending on the binarize parameter) [54].

The decision rule for Bernoulli naive Bayes is based on

$$P(x_i \mid y) = P(i \mid y) x_i + \big(1 - P(i \mid y)\big)(1 - x_i) \tag{6}$$

which differs from multinomial NB's rule in that it explicitly penalizes the non-occurrence of a feature i that is an indicator for class y, where the multinomial variant would simply ignore a non-occurring feature [54].

Gaussian Naïve Bayes

This is another NB classifier that assumes that the data has a Gaussian distribution, and the Gaussian distribution can be presented in Eq. 7.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \quad (7)$$

The parameters σ_y and μ_y are estimated using maximum likelihood [54].

2.5.8. Nearest Shrunken Centroids (NSC)

The NSC [74, 75] in classification issues calculates a standardized centroid for each class using the average value of each feature of each class divided by its class standard deviation of that feature. In the next step, the feature vector of a new conditioning factor as input is compared to the centroids of each of these classes. Consequently, the class with the closest centroid (in squared distance) is the predicted class for that new conditioning factor as input data [76]. In this model, using a threshold, each of the class centroids of the features shrinks toward the overall centroid. Mathematically, first a threshold value is assigned to the class centroids of the features, and if it is small for all classes, it is set to zero. Consequently, when shrinking the centroids for all classes is completed the new sample of the feature is classified by the usual nearest centroid rule [76].

2.6. Model Development and Validation

After imputing the missing values, except the response variable ATU, the data was scaled and centered, and 75% of the data was used for model development and 25% for testing. The data was split using model selection function of Scikit-learn [54].

2.6.1. Feature selection

Feature selection is an important step in developing any machine-learning model, and there are various ways of selecting the most appropriate predictors. Some machine learning models can handle feature selection, but not most of them. Therefore, we decided to apply Spearman's correlation to select all the predictors with a significant correlation with ATU [54].

2.6.2. Training machine learning models

Using train_test_split function of Scikit-learn the data was split into 75% for training the model and 25% of the data for model validation [54]. To train all 19 models together, we created a function by listing ensemble methods including ensemble. AdaBoostClassifier, ensemble.BaggingClassifier, ensemble.ExtraTreesClassifier, ensemble.GradientBoostingClassifier, ensemble.RandomForestClassifier; generalized linear models including linear_model.LogisticRegressionCV, linear_model.PassiveAggressiveClassifier, linear_model. RidgeClassifierCV, linear_model.SGDClassifier, linear_model.Perceptron; Navies Bayes methods including naive_bayes.BernoulliNB, naive_bayes.GaussianNB; Nearest Neighbor methods including neighbors.KNeighborsClassifier and neighbors.NearestCentroid; SVM techniques including svm.SVC, svm.NuSVC, svm.LinearSVC; Trees-based methods including tree.DecisionTreeClassifier, tree.ExtraTreeClassifier(), and Neural Network methods including neural_network.MLPClassifier [54].

2.6.3. Validation of machine learning models

All models were validated on 25% of the data using accuracy, precision, recall, and AUC score.

Predicted number of people preferred ATU (TP); Predicted number of people not preferred ATU (TN); incorrectly predicted number of people preferred ATU (FP); incorrectly predicted number of people not preferred ATU (FN)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (9)$$

$$Sensitivity = \frac{predicted\ number\ of\ people\ prefered\ ATU}{total\ number\ of\ people\ prefered\ ATU\ in\ the\ population} \quad (10)$$

Sensitivity is also known as recall, hit rate, and true positive rate.

$$Specificity = \frac{TN}{TN+FP} \quad (11)$$

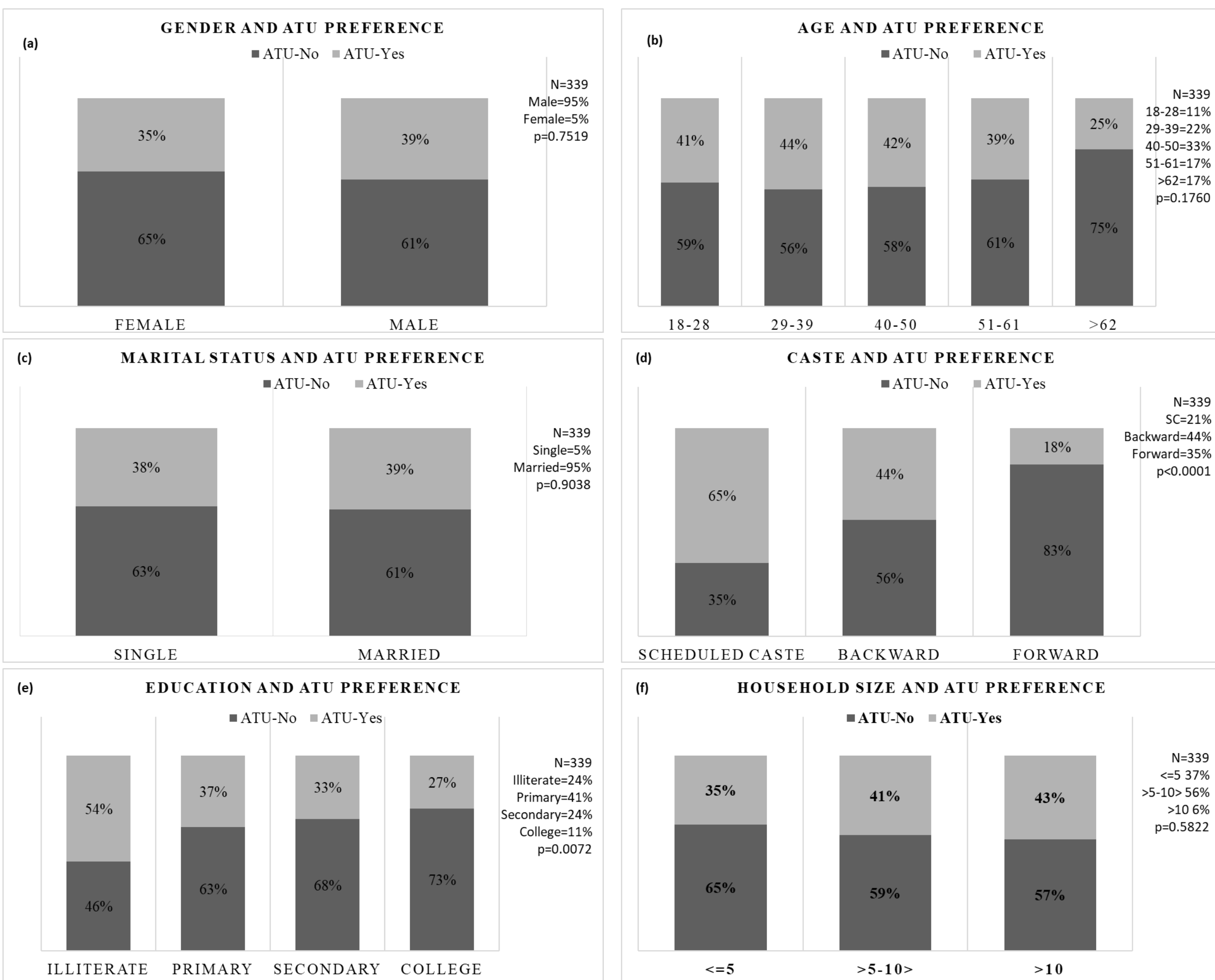
$$Specificity = \frac{predicted\ number\ of\ people\ not\ prefered\ ATU}{total\ number\ of\ people\ not\ prefered\ ATU\ in\ the\ population} \quad (12)$$

Specificity is also known as true negative rate and selectivity.

3. RESULTS

3.1. Socioeconomic, demographic and other social factors

With 39% of the population, ATU was the most preferred sustainable arsenic mitigation technology and significantly different from the other options including piped water supply system, deep tube wells, dug wells/open wells, and rainwater harvesting system [22]. Although there was a less participation of females in the survey than males, we did not find any significant difference (p=0.7519) among them preferring ATU as the most sustainable arsenic mitigation technology (Figure 3-a). People of age group of 40-50 (Figure 3-b) were more interested in adopting ATU as the sustainable arsenic mitigation technology than the other age groups, but not significantly different from each other (p=0.1760).



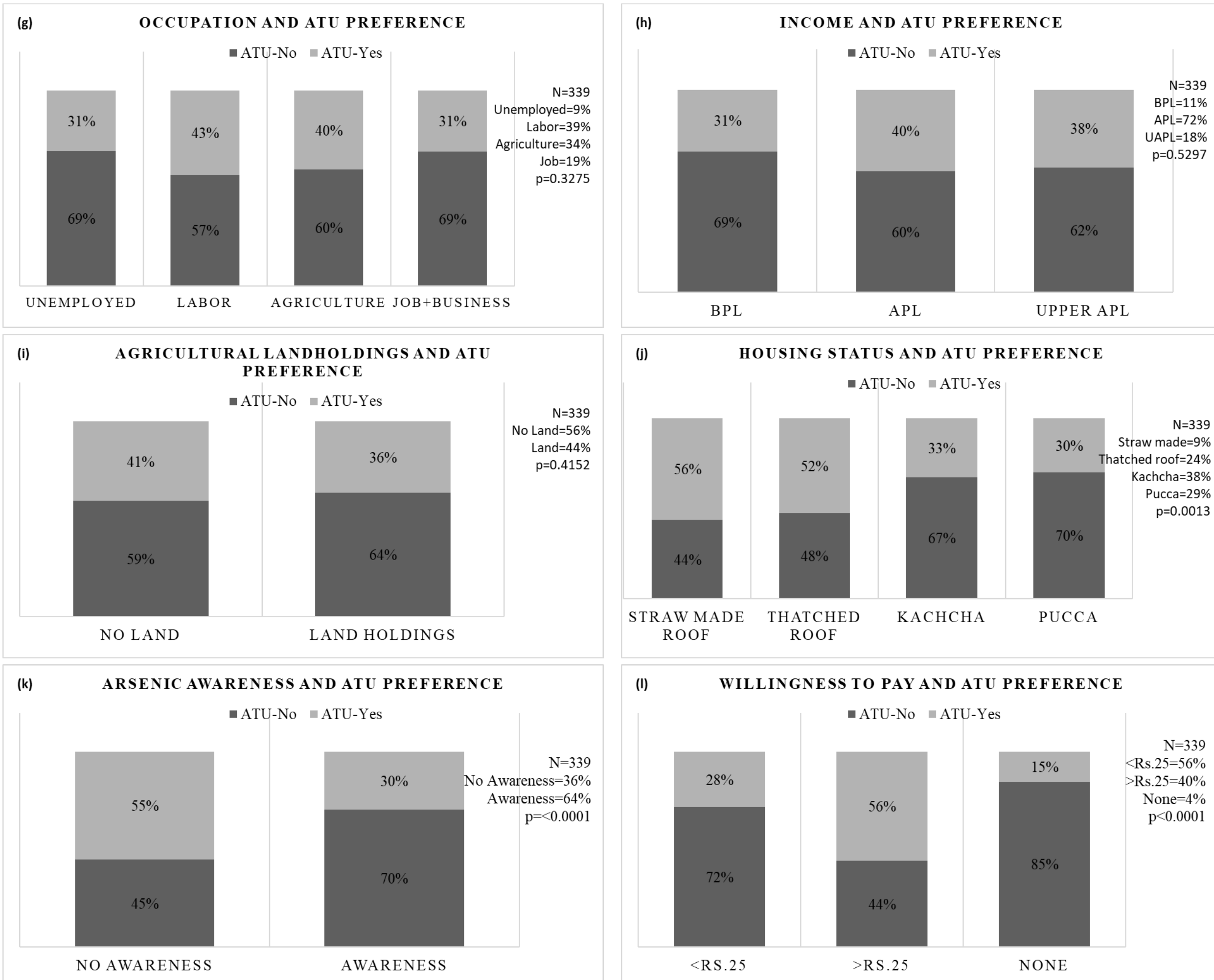
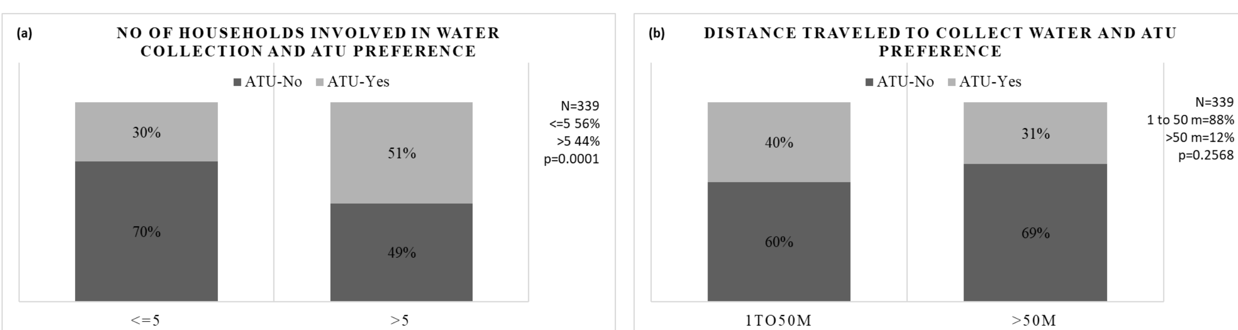


Figure 3. (a) gender and ATU preference; (b) age and ATU preference; (c) marital status and ATU preference; (d) caste and ATU preference; (e) education and ATU preference; (f) household size and ATU preference; (g) occupation and ATU preference; (h) income and ATU preference; (i) agricultural landholdings and ATU preference; (j) housing status and ATU preference; (k) arsenic awareness and ATU preference; (l) willingness to pay for arsenic mitigation and ATU preference.

A similar trend was observed with marital status (Figure 3-c) of the respondents, where there was no difference in the response of preferring ATU ($p=0.9038$) among single and married people. This could be also because of the low participation of unmarried people (5%) in the survey. Caste was appeared to be one of the most important features of the respondents that distinguish the preference of ATU among various castes. A majority of SC (65%) preferred ATU (Figure 3-d) followed by BC (44%) and FC (18%) and their preferences were significantly different ($p<0.0001$) from each other. A similar trend was observed across various education levels of the respondents (Figure 3-e) where people with no education found to be more likely (54%) to adopt ATU than people with various levels of education and was significantly different from each other (0.0072). There was no association between household size and preference of ATU (p=0.5822) (Figure 3-f). A similar trend was observed with the occupation (p=0.3275) (Figure 2-g), income (p=0.5297) (Figure 3-h), and agricultural landholdings (p=0.4152) (Figure 3-i) of the respondents where their preference of ATU as the sustainable arsenic mitigation technology was not different. People live in straw-made roofed houses were more likely (56%) (Figure 3-j) to prefer ATU than the people live in better housing structures and the responses were significantly different (p=0.0013) from each other. The respondents less aware of arsenic (55%) were more likely to prefer ATU than the respondents with arsenic awareness (Figure 3-k) and their responses were significantly different (p<0.001). This further interprets that the people perceive technology/filters as a better solution to purify any water contaminants. The respondents with a WTP >Rs. 25 were more likely (56%) to prefer ATU and their responses were different (p<0.001) across various WTP levels (Figure 3-l).

3.2. Water and sanitation factors

When it comes to respondents' water and sanitation behaviors, the number of households involved in water collection was significantly (p<0.0001) associated with ATU preference (Figure 4-a). The households with more than five members involved in water collection were more likely to prefer ATU, which may indicate their concern about dependency on many people for water collection and collective time loss. The distance travelled (Figure 4-b) and the time spent to collect water (Figure 4-c) were not significantly associated with ATU preference. However, people's sanitation habits were significantly associated with ATU preference. The respondents who defecate in the open field were more likely (43%) to prefer than who uses their own toilets (23%). The similar trend was observed for the materials used for hand washing after the defecation (Figure 4-c).



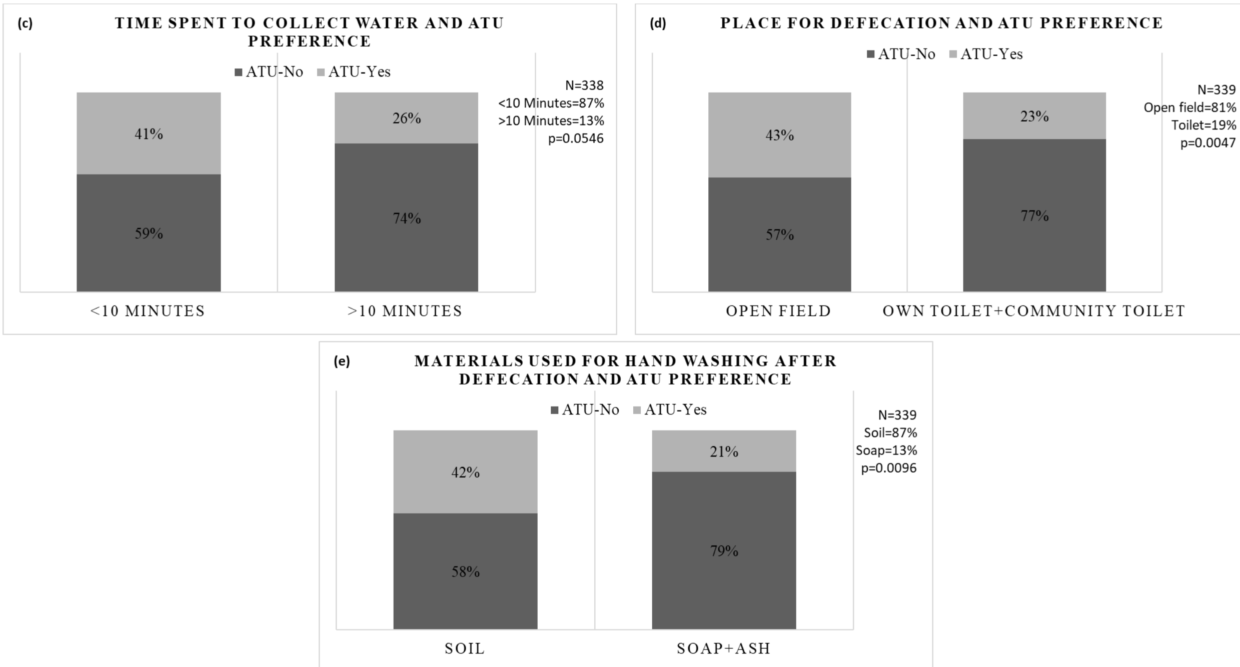
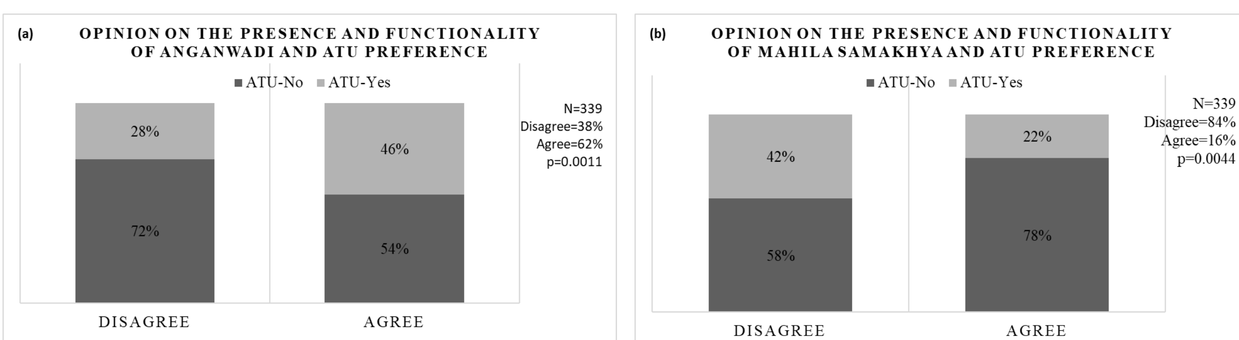


Figure 4. (a) number of households involved in water collection and ATU preference; (b) distance travelled to collect water and ATU preference; (c) time spent to collect water and ATU preference; (d) materials used for hand washing after defecation and ATU preference.

3.3. Social capital

Social capital found to be a major player in guiding the decision making to adopt arsenic mitigation technology. The respondents, who agreed on the presence and functionality of Anganwadi, were more likely to prefer ATU with a significant difference (p=0.0011) from who disagreed (Figure 5-a). A contrasting pattern was found with the respondents' response on the presence and functionality of Mahila Samakhya (Figure 5-b) and Self-Help Groups (Figure 5-c) with the respondents who disagree were more likely to prefer ATU than who agreed on this question.



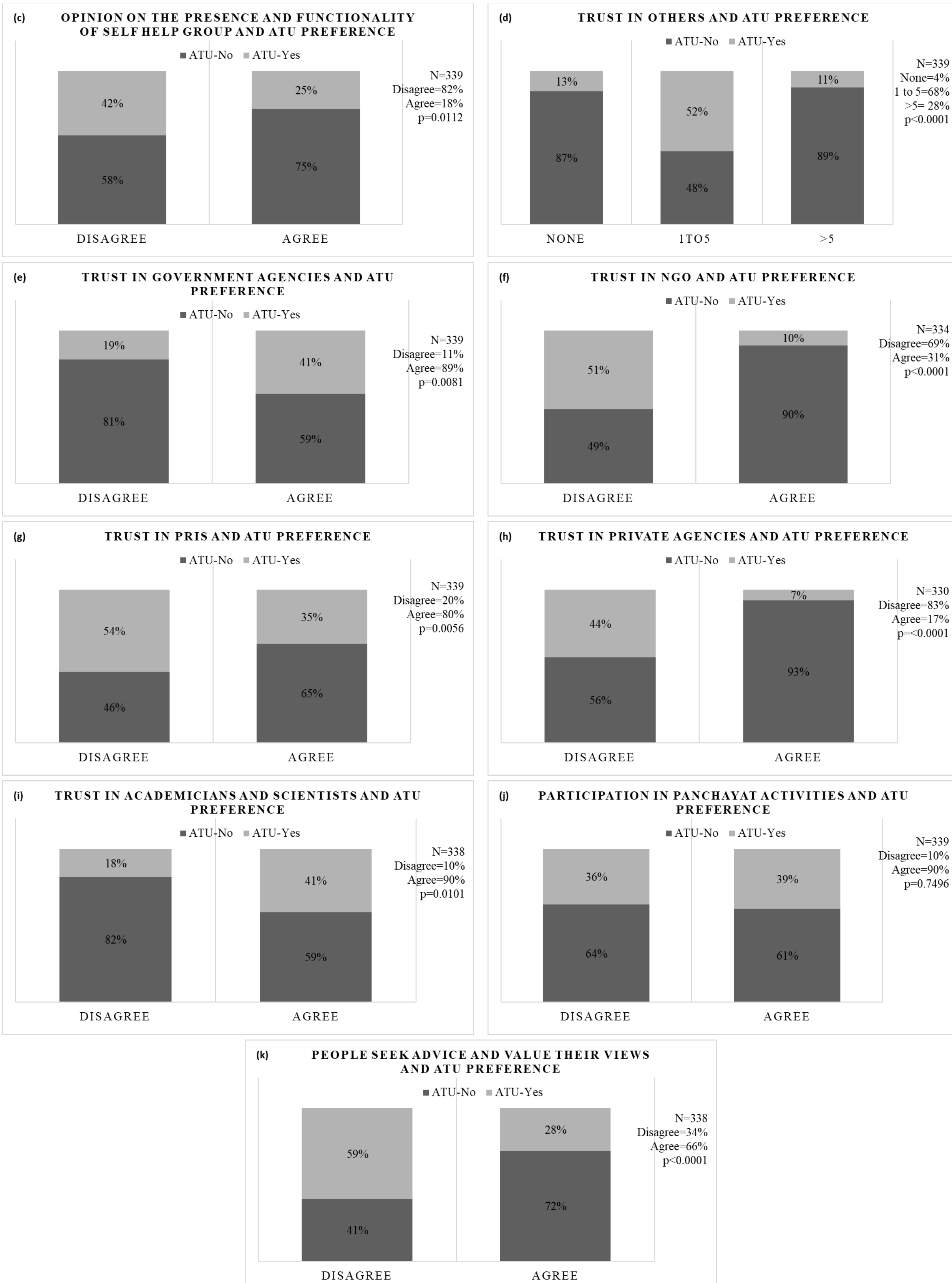

(c) OPINION ON THE PRESENCE AND FUNCTIONALITY OF SELF HELP GROUP AND ATU PREFERENCE
ATU-No ATU-Yes
42%
58%
25%
75%
DISAGREE
AGREE
N=339
Disagree=82%
Agree=18%
p=0.0112
(d) TRUST IN OTHERS AND ATU PREFERENCE
ATU-No ATU-Yes
13%
87%
52%
48%
11%
89%
NONE
1TO5
>5
N=339
None=4%
1 to 5=68%
>5= 28%
p<0.0001
(e) TRUST IN GOVERNMENT AGENCIES AND ATU PREFERENCE
ATU-No ATU-Yes
19%
81%
41%
59%
DISAGREE
AGREE
N=339
Disagree=11%
Agree=89%
p=0.0081
(f) TRUST IN NGO AND ATU PREFERENCE
ATU-No ATU-Yes
51%
49%
10%
90%
DISAGREE
AGREE
N=334
Disagree=69%
Agree=31%
p<0.0001
(g) TRUST IN PRIS AND ATU PREFERENCE
ATU-No ATU-Yes
54%
46%
35%
65%
DISAGREE
AGREE
N=339
Disagree=20%
Agree=80%
p=0.0056
(h) TRUST IN PRIVATE AGENCIES AND ATU PREFERENCE
ATU-No ATU-Yes
44%
56%
7%
93%
DISAGREE
AGREE
N=330
Disagree=83%
Agree=17%
p=<0.0001
(i) TRUST IN ACADEMICIANS AND SCIENTISTS AND ATU PREFERENCE
ATU-No ATU-Yes
18%
82%
41%
59%
DISAGREE
AGREE
N=338
Disagree=10%
Agree=90%
p=0.0101
(j) PARTICIPATION IN PANCHAYAT ACTIVITIES AND ATU PREFERENCE
ATU-No ATU-Yes
36%
64%
39%
61%
DISAGREE
AGREE
N=339
Disagree=10%
Agree=90%
p=0.7496
(k) PEOPLE SEEK ADVICE AND VALUE THEIR VIEWS AND ATU PREFERENCE
ATU-No ATU-Yes
59%
41%
28%
72%
DISAGREE
AGREE
N=338
Disagree=34%
Agree=66%
p<0.0001

Figure 5. (a) opinion of the presence and functionality of Anganwadi and ATU preference; (b) opinion of the presence and functionality of Mahila Samakhya and ATU preference; (c) opinion of the presence and functionality of Self Help Group and ATU preference; (d) trust in others and ATU preference; (e) trust in government agencies and ATU; (f) trust in NGOs and ATU; (g) trust in PRIs and ATU preference; (h) trust in private agencies and ATU preference; (i) trust in academicians and scientists and ATU preference; (j) participation in panchayat activities and ATU preference; (k) people seek advice and value their views and ATU preference.

The respondents who trust between 1 and 5 people were more likely (Figure 5-d) to prefer ATU and significantly different ($p<0.001$) than other trust categories. The people who trust in government agencies (Figure 5-e) and academicians (Figure 4-i) were more likely to prefer ATU and their responses were significantly different ($p=0.0081$ and $p=0.0101$ respectively). In contrast, the respondents who do not trust NGO (Figure 5-f, $p<0.0001$), PRIs (Figure 5-g, $p=0.0056$), and private agencies (Figure 5-h, $p<0.0001$) were more likely to prefer ATU. People's participation in panchayat activities had no significant association ($p=0.7496$) with ATU preference (Figure 5-j). On the other hand, people who reported that other people in the society do not seek advice and value their views were more likely (59%, $p<0.0001$) to prefer ATU than who agreed on this question.

3.4. Comparing machine learning models

We developed 19 different models and all of them were trained on the 75% of the data set (Figure 6). Among these 19 models, Decision Tree (accuracy=0.988), Extra Trees (accuracy=0.988), Bagging (accuracy=0.967), Random Forests (accuracy=0.972), and Gradient Boosting (accuracy=0.953) achieved the highest accuracy, all above 95%, Decision Tree being the top performer (Figure 6).

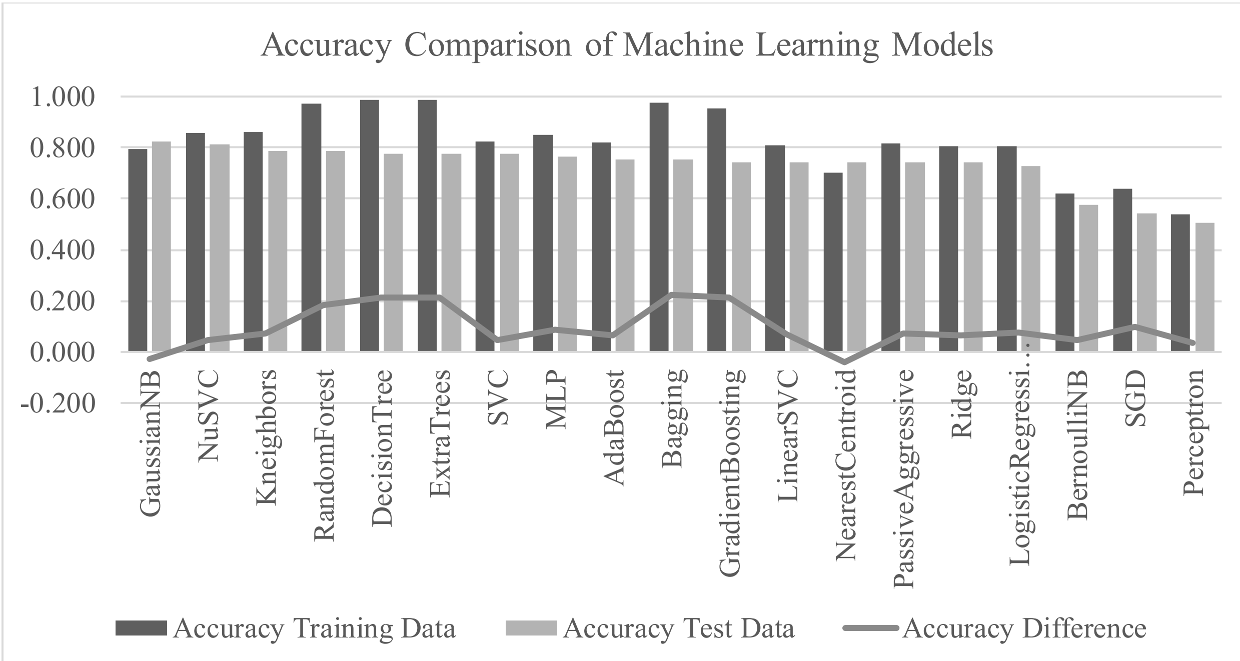


Figure 6. Model comparison for accuracy: the graph is in descending order based on the accuracy on test data.

Nine models, including K-neighbors (accuracy=0.862), NuSVC (accuracy=0.858), MLP (accuracy=0.850), SVC (accuracy=0.823), AdaBoost (accuracy=0.819), PassiveAggressive (accuracy=0.815), LinearSVC (accuracy=0.811), LogisticRegression (0.807), and RidgeClassifier (accuracy=0.807) had the accuracy above 80%. Gaussian NB and Nearest Centroid also had the satisfactory accuracy of 0.795 and 0.704, respectively (Figure 6). However, SGD (accuracy=0.638), Bernauli NB (accuracy=0.622), and Perceptron (accuracy=0.539) had the poorest performance (Figure 6).

When the accuracy of 19 models on test data was compared it was apparent that a majority of the models had high variance in the accuracy on training and test datasets (Figure 6). Considering the accuracy as a model performance criterion, Gaussian NB model was found to have less variance in the accuracy where the overall accuracy on test data was 0.824, 0.028 greater than the accuracy on the training data. The second most stable model was NuSVC (accuracy=0.812) with a difference between training and testing dataset of 0.047. Other models with good performance on test data after Gaussian NB and NuSVC were, respectively, K-neighbors (accuracy=0.788), Random Forests (accuracy=0.788), Decision Trees (accuracy=0.777), Extra Trees (accuracy=0.777), SVC (accuracy=0.777), MLP (accuracy=0.765), AdaBoost (accuracy=0.753), Bagging (accuracy=0.753), Gradient Boosting (accuracy=0.741), LinearSVC (accuracy=0.741), the Nearest Centroid (accuracy=0.741), Passive Aggressive (accuracy=0.741), Ridge (accuracy=0.741), and Logistic Regression (accuracy=0.729). Bernouli NB (accuracy=0.577), SGD (accuracy=0.541), and Perceptron (accuracy=0.506) models were the poorest performers. After comparing the accuracy of all the models of training and test data, Gaussian NB was found to be the clear winner and Perceptron was as good as an intuition, a random model.

As mentioned in the methodology section, Precision measures the type-I error i.e., false positive. It is also known as how sensitive the model is correctly predicting the positive events. While comparing the precision score on the training data, we found that Extra Trees, Random Forests, and Decision Trees have the perfect precision of "1." Bagging (precision=0.979) and Gradient Boosting (precision=0.957) almost equally precise after Extra Trees, Random Forests and Decision Trees. MLP (precision=0.854), NuSVC (precision=0.849), Passive Aggressive (precision=0.836), SVC (precision=0.831), and K-neighbors (precision=0.828) achieved good precision scores all above 80%. Ridge (precision=0.776), Logistic regression (precision=0.776), Ada Boost (precision=0.772), Linear SVC (precision=0.767) and Gaussian NB (precision=0.683) also had good precision score ranging between 70% and 80%. The Nearest Centroid (precision=0.581), SGD (precision=0.511), and Perceptron (precision=0.451) scored less precision. Bernoulli NB (precision=0) couldn't produce any precision score. It is clear that Bernoulli failed to precisely predict the positive outcome, i.e. the preference of ATU as a sustainable arsenic mitigation technology (Figure 7).

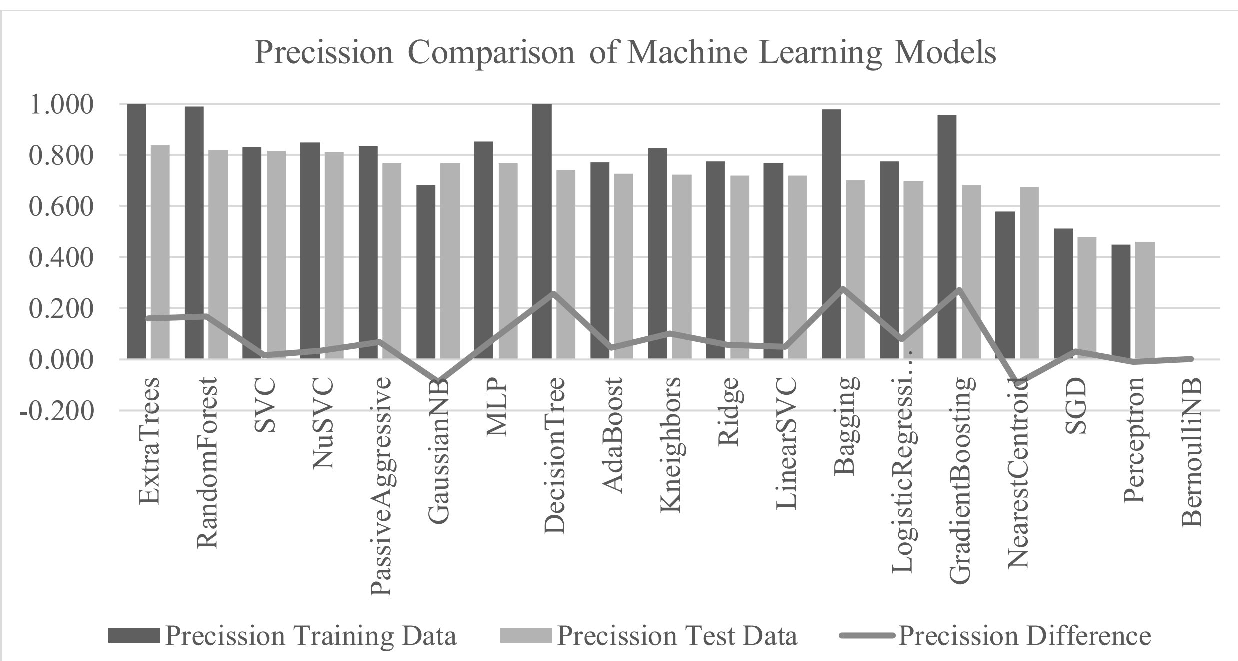


Figure 7. Model comparison for precision: the graph is in descending order based on the precision on test data.

The precision on test data for Extra Trees got reduced from 1 to 0.840, still the highest among rest of the models. Random Forest model had 0.821 precision value on test data, followed by SVC (precision=0.815), NuSVC (precision=0.813), Passive Aggressive (precision=0.769), GaussianNB (precision=0.769), MLP (precision=0.767), Decision Tree (precision=0.743), Ada Boost (precision=0.727), K-neighbors (precision=0.725), Ridge (precision=0.719), Linear SVC (precision=0.719), and Bagging (precision=0.703), all above 70%. Logistic regression (precision=0.697), Gradient Boosting (precision=0.684), and Nearest Centroid (precision=0.675) scored precision between 60% and 70%. SGD (precision=0.479), Perceptron (precision=0.461), and Bernoulli NB (precision=0.000) had the lower precision scores, Bernoulli NB being the lowest similar to its precision score on training data. That further states that when Bernoulli NB predicts ATU as a sustainable arsenic mitigation technology, it is correct 0% of the time (Figure 7).

As opposed to Precision, Recall measures the type-II error i.e., false negative. It is also known as how specific the model correctly predicts the negative events. While comparing the precision score on the test data, we found that Perceptron had the perfect Recall of 1, followed by SGD (recall=0.990), Decision Tree (recall=0.969), Extra Trees (recall=0.969), Bagging (recall=0.958), Random Forest (recall=0.938), and Gradient Boosting (recall=0.917), all above 90%. Gaussian NB (recall=0.854), K-neighbors (recall=0.802), NuSVC (recall=0.760), the Nearest Centroid (recall=0.750), Ada Boost (recall=0.740), MLP (recall=0.729), and Linear SVC (recall=0.719) had recall score above 70% but below 90%. Logistic Regression (recall=0.688), Ridge (recall=0.688), SVC (recall=0.667), and Passive Aggressive (recall=0.635) had comparatively lower recall value among all 19 models, and Bernoulli NB (recall=0.000) being at the bottom of this sequence. (Figure 8).

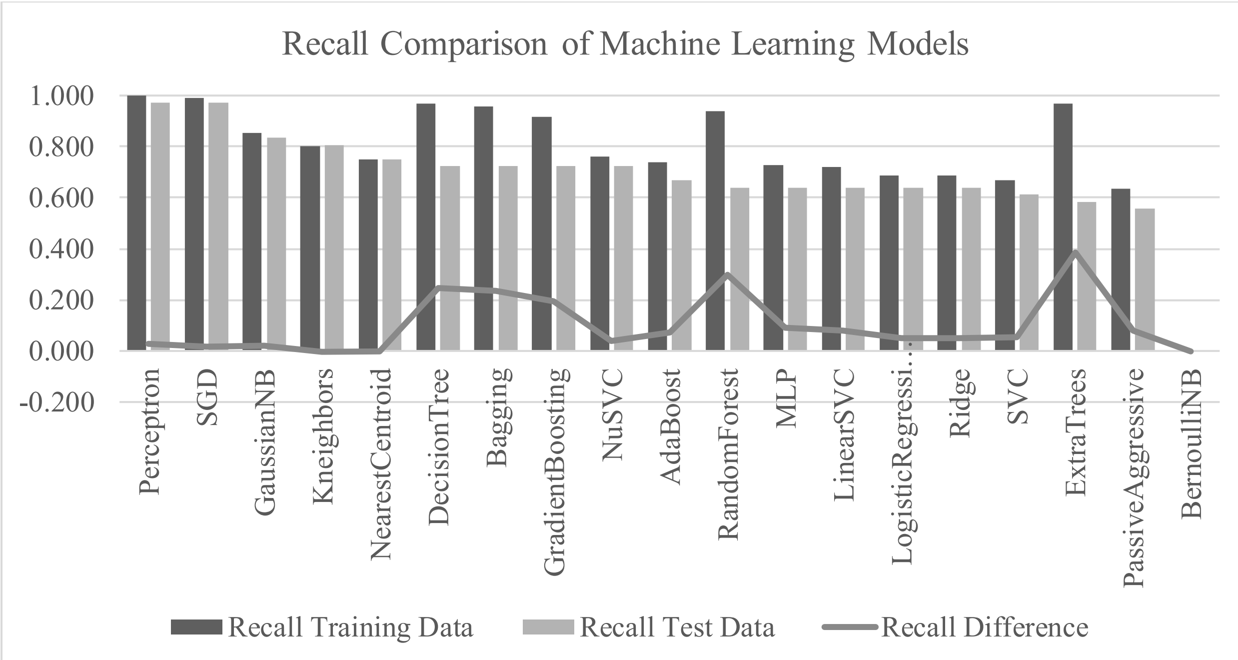


Figure 8. Model comparison for recall: the graph is in descending order based on the recall on test data.

The recall on test data for Perceptron got reduced from 1 to 0.972, still the highest among rest of the models. SGD achieved 0.972 recall score on the test data, followed by Gaussian NB (recall=0.833), K-neighbors (recall=0.806), the Nearest Centroid (recall=0.750), Decision Tree (recall=0.722), Bagging (recall=0.722), Gradient Boosting (0.722), Nu SVC (recall=0722), Ada Boost (recall=0.667), Random Forest (recall=0.639), MLP (recall=0.639), Linear SVC (recall=0.639), Logistic regression (recall=0.639), Ridge (recall=0.639), SVC (recall=0.611), Extra Trees (recall=0.583), Passive Aggressive (recall=0.556), and Bernoulli NB (recall=0.000) (Figure 8). Perceptron and SGD had the highest recall on both training and testing data. On the other hand, Bernoulli NB achieved "0" recall score (Figure 8).

AUC is considered as the most preferable machine learning model performance metrics to evaluate the accuracy and the robustness of any machine learning models. We found that Decision Tree (AUC=0.984) and Extra Trees (AUC=0.984) achieved the highest AUC score on training data. Bagging (AUC=0.973) secured the third place, followed by Random Forest (AUC=0.966), Gradient Boosting (AUC=0.946), K-neighbors (AUC=0.850), NuSVC (AUC=0.839), MLP (AUC=0.827), Gaussian NB (AUC=0.807), Ada Boost (AUC=0.803), Linear SVC (AUC=0.793), SVC (AUC=0.792), Ridge (AUC=0.784), Logistic Regression (AUC=0.784), Passive Aggressive

(AUC=0.780), Nearest Centroid (AUC=0.710), SGD (AUC=0.707), Perceptron (AUC=0.630), and Bernoulli NB (AUC=0.500) (Figure 9).

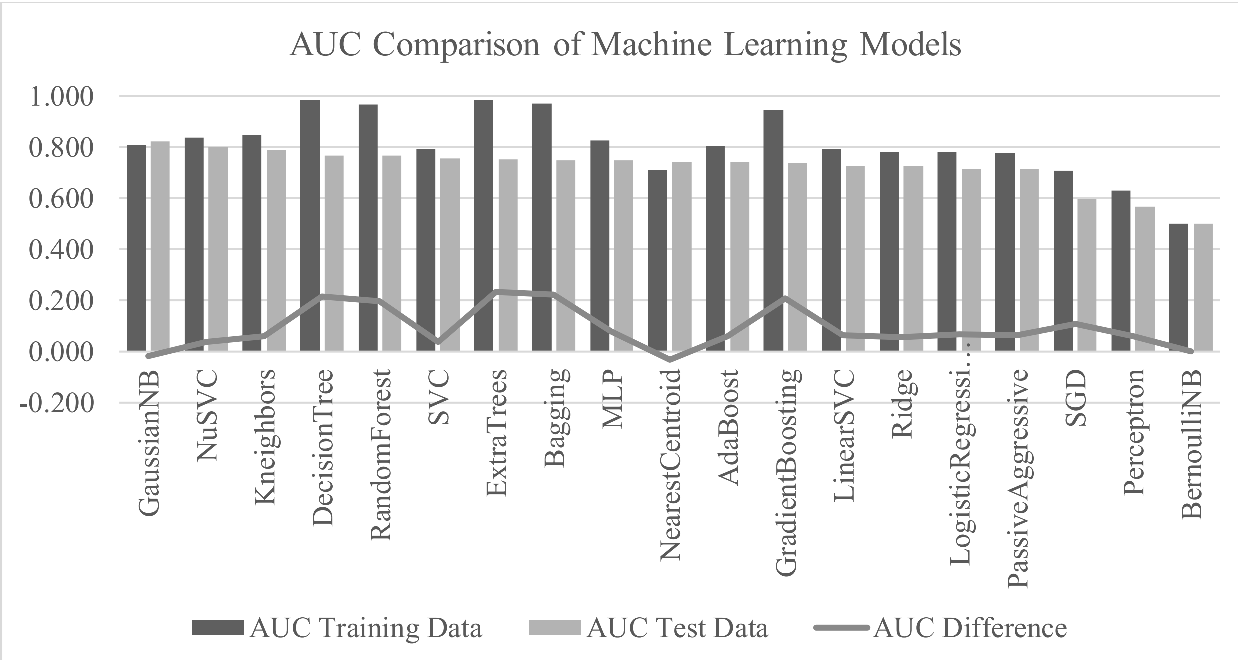


Figure 9. Model comparison for AUC: the graph is in descending order based on the AUC on test data.

Gaussian NB had the best AUC of 0.825 on test data, followed by NuSVC (AUC=0.800), K-neighbors (AUC=0.791), Decision Trees (AUC=0.769), Random Forest (AUC=0.768), SVC (AUC=0.755), Extra Trees (AUC=0.751), Bagging (AUC=0.749), MLP (AUC=0.748), Nearest Centroid (AUC=0.742), Ada Boost (AUC=0.741), Gradient Boosting (AUC=0.739), Linear SVC (AUC=0.728), Ridge (AUC=0.728), Logistic Regression (AUC=0.717), Passive Aggressive (AUC=0.717), SGD (AUC=0.589), Perceptron (AUC=0.568), and Bernoulli NB (AUC=0.500) (Figure 10).

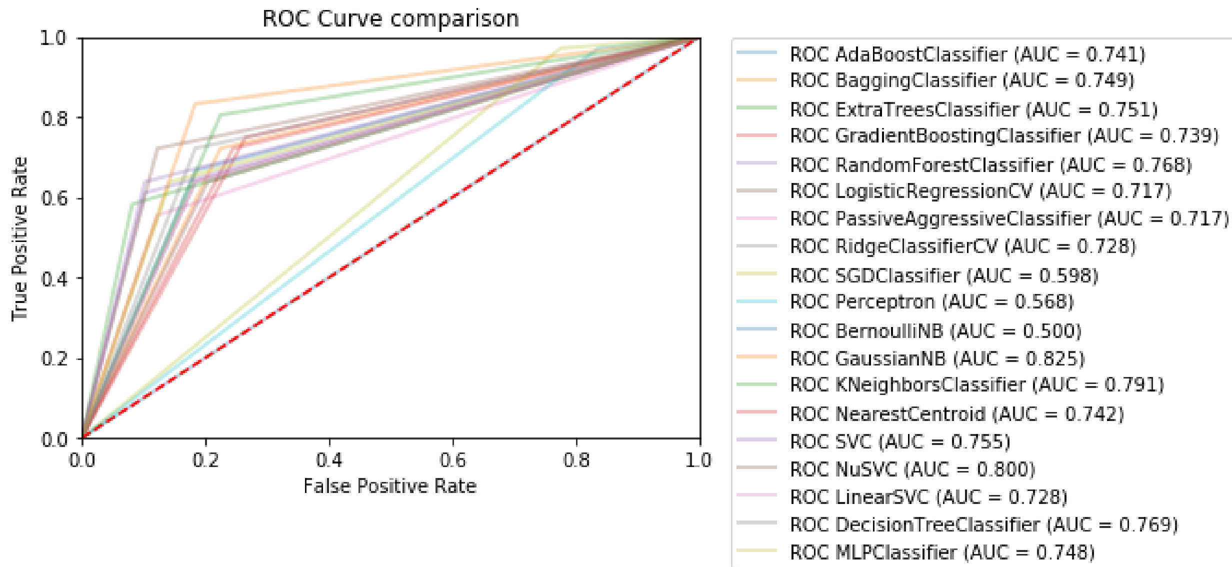


Figure 10. The AUC of all models on test data.

With the highest AUC of 0.825 on test data, GaussianNB can be declared as the best model in predicting ATU as a sustainable arsenic mitigation technology (Figure10).

4. DISCUSSION

The DT (21.4%), ExtraTrees (21.4%), Bagging (22.9%), Gradient Boosting (22.2%), and Random Forest (18.9%) models found to be over fitted on the training data while comparing on the test data (Figure 6). There were other ten models that overfitted above 5% on the training data. It appeared that these models learned the noise in the training data as a natural trend in the data that unfortunately could not be applied to test data. Therefore, the accuracy on test data went down by as high as 22.9%. That further indicates that these models cannot be generalized to unseen dataset. On the other hand, only two models including GaussianNB (-3.5%) and NearestCentroid (-5.8%) seem to be under-fitted on the training data, being very conservative in learning only the meaningful trend in the data and predicted with slightly less accuracy on test data. Each algorithm has its own advantages and disadvantages and have been developed to address specific scenario. Tree-based and ensemble algorithms are known to be prone to over-fitting if the sample size is not appropriate, assumptions are not met, or features are varied in nature [77, 78]. Among all, GaussianNB was found to be the most robust model. The reason could be that NB is a fast, highly scalable algorithm and is a good choice for binary classification problems [79, 80]. It can easily be updated on new data. In this study, we only had 339 samples, which is not a large sample size and NB is known to be easily get trained on a small dataset that reflected from its highest AUC score (Figure 10). Since it is highly scalable, this technique can easily be applied to other similar arsenic contaminated areas [79, 80]. In a recent study, the authors found that the GaussianNB outperformed the SVM as it is statistically robust, neutrally reasonable, and could reproduce across unseen datasets [79]. Socioeconomic and psychological data comprises both continuous data (age and income) and discrete data (gender, education level, marital status, etc.) and Gaussian NB works pretty well with multidimensional data. In reality, survey-based data also suffers from missingness, GaussianNB is also prone to missing data and over-fitting, and it also ignores irrelevant features in the model. A majority of machine learning outcomes are difficult to interpret. Gaussian NB provides predictive ability to users as it can make probabilistic predictions [80, 81].

5. CONCLUSION

Application of various techniques of artificial intelligence in environmental data modeling and prediction is a recent phenomenon. Its application on building prediction models on socioeconomic and psychological data collected from communities living in environmental contaminated regions has just began. This study is a founding step in providing insights on how various state-of-the-art artificial intelligence can be used for developing accurate prediction models of sustainable arsenic mitigation technologies. Selecting an appropriate method is key in developing meaningful prediction models as the machine (computational instruments) only understands the data as numbers. In this study, we have evaluated several cutting edges linear, nonlinear, ensemble, tree-based, and Naïve Bayes-based machine learning algorithms for predicting sustainable arsenic mitigation technology. Gaussian NB found to be the best model to fit to such multidimensional data. Achieving greater than 70% of AUC on test data by other 15 models is also promising. The top three models Gaussian NB, NuSVC, and K-neighbors are nonlinear classifiers that considers a nonlinear association between independent and dependent variables. The bottom performers including logistic regression, passive aggressive, SGD, and perceptron are all linear classifiers and could not do a better justice with this data. Bernoulli NB was clearly a bad choice of model as it assumes the features should be binary and that's the reason it failed to develop a meaning prediction model. Nonlinear and ensemble models are the better choice of models for multidimensional data where the association between the features are not linear, but complex. We also understand that if a few top linear and ensemble models can be further explored and finetuned, we could probably enhance the model performance. Where funding to generate such data in a large number is a challenge, Gaussian NB model is like a life-saving method that requires less data to learn, can handle missing data, prone to over-fitting, and easy to interpret. We evaluated a few less common algorithms that provided hope for exploring more for developing prediction models on socioeconomic-environmental data including NuSVC, Extra Tree, and Nearest Centroid. A larger sample size, careful feature selection, feature engineering may also help improve the performance of these models. Some models do require hyper parameter tuning, thus selecting the optimum hyper parameters for such models will also help improve the model performance.

References

[1] S.K. Singh, R.W. Taylor, M.M. Rahman, B. Pradhan, Developing robust arsenic awareness prediction models using machine learning algorithms, Journal of environmental management, 211 (2018) 125-137.

[2] S.K. Singh, E.A. Stern, Global Arsenic Contamination: Living with the Poison Nectar, Environment: Science and Policy for Sustainable Development, 59 (2017) 24-28.

[3] S.K. Singh, R.W. Taylor, Assessing the role of risk perception in ensuring sustainable arsenic mitigation, Groundwater for Sustainable Development, (2019) 100241.

[4] J. Bundschuh, M. Litter, V.S. Ciminelli, M.E. Morgada, L. Cornejo, S.G. Hoyos, J. Hoinkis, M.T. Alarcon-Herrera, M.A. Armienta, P. Bhattacharya, Emerging mitigation needs and sustainable options for solving the arsenic problems of rural and isolated urban areas in Latin America–A critical analysis, water research, 44 (2010) 5828-5845.

[5] D. Chakraborti, S.K. Singh, H.M. Rashid, M.M. Rahman, Arsenic: Occurrence in Groundwater, Encyclopedia of Environmental Health. Burlington: Elsevier, 2 (2017) 1-17.

[6] S. Murcott, Arsenic contamination in the world, IWA publishing, 2012.

[7] D. Chakraborti, S.K. Singh, M.M. Rahman, R.N. Dutta, S.C. Mukherjee, S. Pati, P.B. Kar, Groundwater Arsenic Contamination in the Ganga River Basin: A Future Health Danger, International Journal of Environmental Research and Public Health, 15 (2018) 180.

[8] A. Heikens, Arsenic contamination of irrigation water, soil and crops in Bangladesh: Risk implications for sustainable agriculture and food safety in Asia, RAP Publication (FAO), (2006).

[9] M.M. Moriarty, I. Koch, R.A. Gordon, K.J. Reimer, Arsenic speciation of terrestrial invertebrates, Environmental Science & Technology, 43 (2009) 4818-4823.

[10] M. Bassil, F. Daou, H. Hassan, O. Yamani, J.A. Kharma, Z. Attieh, J. Elaridi, Lead, cadmium and arsenic in human milk and their socio-demographic and lifestyle determinants in Lebanon, Chemosphere, 191 (2018) 911-921.

[11] M. Molin, S.M. Ulven, H.M. Meltzer, J. Alexander, Arsenic in the human food chain, biotransformation and toxicology–Review focusing on seafood arsenic, Journal of Trace Elements in Medicine and Biology, 31 (2015) 249-259.

[12] M. Hossain, S.N. Rahman, P. Bhattacharya, G. Jacks, R. Saha, M. Rahman, Sustainability of arsenic mitigation interventions—an evaluation of different alternative safe drinking water options provided in MATLAB, an arsenic hot spot in Bangladesh, Frontiers in Environmental Science, 3 (2015) 30.

[13] M. Hossain, P. Bhattacharya, G. Jacks, M. von Brömssen, K.M. Ahmed, M.A. Hasan, S.K. Frape, Sustainable arsenic mitigation–from field trials to implementation for control of arsenic in drinking water supplies in Bangladesh, in: Best Practice Guide on the Control of Arsenic in Drinking Water, IWA Publishing UK, 2017, pp. 99-116.

[14] A. Kabir, G. Howard, Sustainability of arsenic mitigation in Bangladesh: Results of a functionality survey, International Journal of Environmental Health Research, 17 (2007) 207-218.

[15] M. Shafiquzzaman, M.S. Azam, I. Mishima, J. Nakajima, Technical and social evaluation of arsenic mitigation in rural Bangladesh, Journal of health, population, and nutrition, 27 (2009) 674.

[16] N. Shibasaki, P. Lei, A. Kamata, Evaluation of deep groundwater development for arsenic mitigation in western Bangladesh, Journal of Environmental Science and Health Part A, 42 (2007) 1919-1932.

[17] L.H. Winkel, P.T.K. Trang, V.M. Lan, C. Stengel, M. Amini, N.T. Ha, P.H. Viet, M. Berg, Arsenic pollution of groundwater in Vietnam exacerbated by deep aquifer exploitation for more than a century, Proceedings of the National Academy of Sciences, 108 (2011) 1246-1251.

[18] M.M. Hira-Smith, Y. Yuan, X. Savarimuthu, J. Liaw, A. Hira, C. Green, T. Hore, P. Chakraborty, O.S. Von Ehrenstein, A.H. Smith, Arsenic concentrations and bacterial contamination in a pilot shallow dugwell program in West Bengal, India, Journal of Environmental Science and Health Part A, 42 (2007) 89-95.

[19] S.K. Singh, Assessing and mapping vulnerability and risk perceptions to groundwater arsenic contamination: Towards developing sustainable arsenic mitigation models (Order No. 3701365). Available from ProQuest Dissertations & Theses Full Text. (1681668682). in: Earth and Environmental Studies, Montclair State University, USA, 2015, pp. 392.

[20] S.K. Singh, N. Vedwan, Mapping Composite Vulnerability to Groundwater Arsenic Contamination: An Analytical Framework and a Case Study in India, Natural Hazards, 75 (2015) 1883-1908.

[21] S.K. Singh, An Analysis of the Cost-Effectiveness of Arsenic Mitigation Technologies: Implications for Public Policy, International Journal of Sustainable Built Environment, 6 (2017) 522-535.

[22] S.K. Singh, R.W. Taylor, H. Su, Developing Sustainable Models of Arsenic-Mitigation Technologies in the Middle-Ganga Plain in India, Current Science, 113 (2017) 80-93.

[23] S.K. Singh, R.W. Taylor, Assessing and Mapping Human Health Risks Due to Arsenic and Socioeconomic Correlates for Proactive Arsenic Mitigation, in: Arsenic Water Resources Contamination, Springer, 2020, pp. 231-256.

[24] B.K. Thakur, V. Gupta, Arsenic concentration in drinking water of Bihar: health issues and socio-economic problems, Journal of Water, Sanitation and Hygiene for Development, 6 (2016) 331-341.

[25] B.K. Thakur, V. Gupta, Arsenic-Contaminated Drinking Water and the Associated Health Effects in the Shahpur Block of Bihar: A Case Study From Five Villages, in: Arsenic Water Resources Contamination, Springer, 2020, pp. 257-271.

[26] S. Priyadarshini, How the arsenic-affected perceive risk, in, 2014.

[27] S.K. Singh, R.W. Taylor, Likelihood of adoption of arsenic-mitigation technologies under perceived risks to health, income, and social discrimination to arsenic contamination, in: Y. Zhu, H. Guo, P. Bhattacharya, A. Ahmad, J. Bundschuh, R. Naidu (Eds.) Environmental Arsenic in a Changing World: Proceedings of the 7th International Congress and Exhibition on Arsenic in the Environment (AS 2018), CRC Press, Beijing, P.R. China, 2018, pp. 700.

[28] R.M. Warner, Applied statistics: From bivariate through multivariate techniques, Sage Publications, 2012.

[29] B.T. Pham, B. Pradhan, D.T. Bui, I. Prakash, M. Dholakia, A comparative study of different machine learning methods for landslide susceptibility assessment: a case study of Uttarakhand area (India), Environmental Modelling & Software, 84 (2016) 240-250.

[30] B.T. Pham, I. Prakash, J. Dou, S.K. Singh, P.T. Trinh, H. Trung Tran, T. Minh Le, V.P. Tran, D. Kim Khoi, A. Shirzadi, A Novel Hybrid Approach of Landslide Susceptibility Modeling

Using Rotation Forest Ensemble and Different Base Classifiers, Geocarto International, (2018) 1-38.

[31] B.T. Pham, I. Prakash, S.K. Singh, A. Shirzadi, H. Shahabi, D.T. Bui, Landslide susceptibility modeling using Reduced Error Pruning Trees and different ensemble techniques: Hybrid machine learning approaches, CATENA, 175 (2019) 203-218.

[32] T.V. Phong, T.T. Phan, I. Prakash, S.K. Singh, A. Shirzadi, K. Chapi, H.-B. Ly, L.S. Ho, N.K. Quoc, B.T. Pham, Landslide susceptibility modeling using different artificial intelligence methods: a case study at Muong Lay district, Vietnam, Geocarto International, (2019) 1-24.

[33] B. Pradhan, A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS, Computers & Geosciences, 51 (2013) 350-365.

[34] B.T. Pham, A. Jaafari, I. Prakash, S.K. Singh, N.K. Quoc, D.T. Bui, Hybrid computational intelligence models for groundwater potential mapping, Catena, 182 (2019) 104101.

[35] W. Chen, B. Pradhan, S. Li, H. Shahabi, H.M. Rizeei, E. Hou, S. Wang, Novel hybrid integration approach of bagging-based fisher's linear discriminant function for groundwater potential analysis, Natural Resources Research, 28 (2019) 1239-1258.

[36] W.S. Jang, B. Engel, C.M. Yeum, Integrated environmental modeling for efficient aquifer vulnerability assessment using machine learning, Environmental Modelling & Software, 124 (2020) 104602.

[37] L. Knoll, L. Breuer, M. Bach, Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning, Science of the total environment, 668 (2019) 1317-1327.

[38] H.M. Rizeei, O.S. Azeez, B. Pradhan, H.H. Khamees, Assessment of groundwater nitrate contamination hazard in a semi-arid region by using integrated parametric IPNOA and data-driven logistic regression models, Environmental monitoring and assessment, 190 (2018) 633.

[39] F.-J. Chang, L.-s. Kao, Y.-M. Kuo, C.-W. Liu, Artificial neural networks for estimating regional arsenic concentrations in a blackfoot disease area in Taiwan, Journal of hydrology, 388 (2010) 65-76.

[40] B. Purkait, S. Kadam, S. Das, Application of Artificial Neural Network Model to Study Arsenic Contamination in Groundwater of Malda District, Eastern India, Journal of Environmental Informatics, 12 (2008).

[41] K.H. Cho, S. Sthiannopkao, Y.A. Pachepsky, K.-W. Kim, J.H. Kim, Prediction of contamination potential of groundwater arsenic in Cambodia, Laos, and Thailand using artificial neural network, Water research, 45 (2011) 5535-5544.

[42] J.D. Ayotte, B.T. Nolan, J.A. Gronberg, Predicting arsenic in drinking water wells of the Central Valley, California, Environmental science & technology, 50 (2016) 7555-7563.

[43] Y. Park, M. Ligaray, Y.M. Kim, J.H. Kim, K.H. Cho, S. Sthiannopkao, Development of enhanced groundwater arsenic prediction model using machine learning approaches in Southeast Asian countries, Desalination and Water Treatment, 57 (2016) 12227-12236.

[44] S. Singh, Arsenic contamination in water, soil, and food materials in Bihar, Lambert Academic Publishing, Germany, 2011.

[45] S.K. Singh, A.K. Ghosh, Entry of arsenic into food material–a case study, World Appl Sci J, 13 (2011) 385-390.

[46] S.K. Singh, A.K. Ghosh, Health Risk Assessment due to Groundwater Arsenic Contamination: Children are at High Risk, Human and Ecological Risk Assessment: An International Journal, 18 (2012) 751-766.

[47] S.K. Singh, A. Ghosh, A. Kumar, K. Kislay, C. Kumar, R. Tiwari, R. Parwez, N. Kumar, M. Imam, Groundwater Arsenic Contamination and Associated Health Risks in Bihar, India, International Journal of Environmental Research, 8 (2014) 49-60.

[48] D. Chakraborti, S.C. Mukherjee, S. Pati, M.K. Sengupta, M.M. Rahman, U.K. Chowdhury, D. Lodh, C.R. Chanda, A.K. Chakraborti, G.K. Basu, Arsenic groundwater contamination in Middle Ganga Plain, Bihar, India: a future danger?, Environmental Health Perspectives, 111 (2003) 1194-1198.

[49] S. Ahamed, D. Chakraborti, Groundwater Arsenic contamination and Health Effects in Bihar and UP, LAP Lambert Academic Publishing, Germany, 2012.

[50] D. Chakraborti, M.M. Rahman, S. Ahamed, R.N. Dutta, S. Pati, S.C. Mukherjee, Arsenic contamination of groundwater and its induced health effects in Shahpur block, Bhojpur district, Bihar state, India: risk evaluation, Environmental Science and Pollution Research, 23 (2016) 9492-9504.

[51] D. Chakraborti, M.M. Rahman, S. Ahamed, R.N. Dutta, S. Pati, S.C. Mukherjee, Arsenic groundwater contamination and its health effects in Patna district (capital of Bihar) in the middle Ganga plain, India, Chemosphere, 152 (2016) 520-529.

[52] T. Kluyver, B. Ragan-Kelley, F. Pérez, B.E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J.B. Hamrick, J. Grout, S. Corlay, Jupyter Notebooks-a publishing format for reproducible computational workflows, in: ELPUB, 2016, pp. 87-90.

[53] W. McKinney, Python for data analysis: Data wrangling with Pandas, NumPy, and IPython, " O'Reilly Media, Inc.", 2012.

[54] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: Machine learning in Python, Journal of machine learning research, 12 (2011) 2825-2830.

[55] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, SciPy 1.0: fundamental algorithms for scientific computing in Python, Nature methods, (2020) 1-12.

[56] W. Winston, Microsoft Excel 2010 Data Analysis and Business Modeling: Data Analysis and Business Modeling, Pearson Education, 2011.

[57] P.M. Atkinson, R. Massari, Generalised linear modelling of susceptibility to landsliding in the central Apennines, Italy, Computers & Geosciences, 24 (1998) 373-385.

[58] W. Chen, X. Zhao, H. Shahabi, A. Shirzadi, K. Khosravi, H. Chai, S. Zhang, L. Zhang, J. Ma, Y. Chen, Spatial prediction of landslide susceptibility by combining evidential belief function, logistic regression and logistic model tree, Geocarto International, (2019) 1-25.

[59] D. Westreich, J. Lessler, M.J. Funk, Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression, Journal of clinical epidemiology, 63 (2010) 826-833.

[60] S.-B. Bai, J. Wang, G.-N. Lü, P.-G. Zhou, S.-S. Hou, S.-N. Xu, GIS-based logistic regression for landslide susceptibility mapping of the Zhongxian segment in the Three Gorges area, China, Geomorphology, 115 (2010) 23-31.

[61] A. Shirzadi, L. Saro, O.H. Joo, K. Chapi, A GIS-based logistic regression model in rock-fall susceptibility mapping along a mountainous road: Salavat Abad case study, Kurdistan, Iran, Natural hazards, 64 (2012) 1639-1656.

[62] S.Z. Mousavi, A. Kavian, K. Soleimani, S.R. Mousavi, A. Shirzadi, GIS-based spatial prediction of landslide susceptibility using logistic regression model, Geomatics, Natural Hazards and Risk, 2 (2011) 33-50.

[63] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, Y. Singer, Online passive-aggressive algorithms, Journal of Machine Learning Research, 7 (2006) 551-585.
[64] Y. Freund, R.E. Schapire, A desicion-theoretic generalization of on-line learning and an application to boosting, in: European conference on computational learning theory, Springer, 1995, pp. 23-37.
[65] L. Breiman, Bagging predictors, Machine learning, 24 (1996) 123-140.
[66] L. Breiman, Random forests, Machine learning, 45 (2001) 5-32.
[67] N.S. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, The American Statistician, 46 (1992) 175-185.
[68] S. Shanmuganathan, Artificial neural network modelling: An introduction, in: Artificial Neural Network Modelling, Springer, 2016, pp. 1-14.
[69] L. Jing, J. Hudson, Numerical method in rock engineering, International Journal of Rock Mechanics and Mining Sciences, 39 (2002) 409-427.
[70] I. Yilmaz, O. Kaynar, Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils, Expert systems with applications, 38 (2011) 5958-5966.
[71] D. Kanungo, M. Arora, S. Sarkar, R. Gupta, A comparative study of conventional, ANN black box, fuzzy and combined neural and fuzzy weighting procedures for landslide susceptibility zonation in Darjeeling Himalayas, Engineering Geology, 85 (2006) 347-366.
[72] C. Polykretis, C. Chalkias, Comparison and evaluation of landslide susceptibility maps obtained from weight of evidence, logistic regression, and artificial neural network models, Natural hazards, 93 (2018) 249-274.
[73] A. Shirzadi, H. Shahabi, K. Chapi, D.T. Bui, B.T. Pham, K. Shahedi, B.B. Ahmad, A comparative study between popular statistical and machine learning methods for simulating volume of landslides, Catena, 157 (2017) 213-226.
[74] R. Tibshirani, T. Hastie, B. Narasimhan, G. Chu, Diagnosis of multiple cancer types by shrunken centroids of gene expression, Proceedings of the National Academy of Sciences, 99 (2002) 6567-6572.
[75] R. Tibshirani, T. Hastie, B. Narasimhan, G. Chu, Class prediction by nearest shrunken centroids, with applications to DNA microarrays, Statistical Science, 18 (2003) 104-117.
[76] M. Pardo, G. Sberveglieri, Random forests and nearest shrunken centroids for the classification of sensor array data, Sensors and Actuators B: Chemical, 131 (2008) 93-99.
[77] M. LeBlanc, J. Crowley, A review of tree-based prognostic models, in: Recent advances in clinical trial design and analysis, Springer, 1995, pp. 113-124.
[78] M. Re, G. Valentini, 1 Ensemble methods: a review 3, (2012).
[79] R.D. Raizada, Y.-S. Lee, Smoothness without smoothing: why Gaussian naive Bayes is not naive for multi-subject searchlight studies, PloS one, 8 (2013).
[80] F. Pereira, M. Botvinick, Information mapping with pattern classifiers: a comparative study, Neuroimage, 56 (2011) 476-496.
[81] K.P. Murphy, Naive bayes classifiers, University of British Columbia, 18 (2006) 60.