International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 9 Issue 10 Series I || October 2020 || PP 01-08

Prediction of Arsenic (As) contamination in rice grain produced in Sindh using Machine Learning

Nazia Pathan[1], Mohsin Ali Memon[2] , Zulfiqar Ali Bhatti[3], Sania Bhatti[4]
Muhammad Owais Raza[5]

[1]*Department of Software Engineering Mehran university of Engineering and Technology Jamshoro, Pakistan.*
[2]*Department of Software Engineering Mehran university of Engineering and Technology Jamshoro, Pakistan.*
[3]*Department of Chemical Engineering Mehran university of Engineering and Technology Jamshoro, Pakistan.*
[4]*Department of Software Engineering Mehran university of Engineering and Technology Jamshoro, Pakistan.*
[5]*Department of Software Engineering Mehran university of Engineering and Technology Jamshoro, Pakistan.*

Abstract: *Arsenic (As) is a carcinogenic element and a potential threat to human life. In Sindh groundwater is contaminant with As at a larger scale, which essentially leaves its toxic effect on crops and vegetables. Rice is a staple food used all around Pakistan and an area where there is As in water and it produces rice crop. It is one of the major causes of the exposure of As to the human population, causing some serious illness, including cancer. As detection in the crop is conventionally an expensive and time-consuming process, so in this research we studied previous researches on As in water, rice in crop-producing districts in Sindh, collected the data and applied a machine learning approach to predict As in rice. We have used the azure platform to perform our machine learning tasks; using different algorithms for prediction. According to the results, linear regression outperformed among them. Finally, a web service is created so we can make a future prediction without going into details of the model.*

Keywords- Machine Learning, Arsenic, Supervised Learning, Regression, Microsoft Azure.

Date of Submission: 06-10-2020 Date of Acceptance: 20-10-2020

I. INTRODUCTION

Water, one of the most basic human needs used not only for the purpose of drinking but also for cultivating crops. Where without water life ceases to exist there, the need for clean water is also very important because we only exposed to water while drinking but our agriculture depends on water and any contamination of water cause contamination in crops we eat. In Sindh province of Pakistan, one of the major reasons for contamination of water is As[23], [19] a huge number of districts in Sindh are prone to As contamination [12], and when that water is used in cultivation that As is passed to crops [13], [20] and having an adverse effect on humans [7] consuming it [23]. In Sindh one of the most common staple foods is rice [24] and almost all major rice producing district such as Larkana, Shikarpur, Ghotki, Kashmore Badin and Thatta [26] have the problem of As in water causing As contamination in rice. For the large consumption of rice in the population of Sindh, it exposes many people to As from rice. This As contamination due to rice is problem all around the world [3]. To detect As in water or in rice [22], the process is very expensive and time-consuming and a very little region can be covered [16] at a time so we have used those studies to collect data from major rice-producing districts and we have used machine learning to predict the value of As in rice from As in water. We are using the Azure platform to create our training pipeline for this problem after that we will test different algorithms and see what works best for this purpose and use it to create web service. The web service that we are creating will provide an easy-to-use interface for performing future predictions. The web service we are working on will be helpful for the government and organizations who are working in the environment and health domain. This study will open gates for researchers and developers to perform an experiment on top of our web service and create solutions using our web service.

II. LITREATURE REVIEW

As, exceedingly harmful element, and its essence in food composites involves worry for the general wellbeing security, explicitly in Bangladesh which is viewed as the most As influenced nation all through the world[4].As may be a characteristic component of the earth's outside and is commonly dispersed all over the environment within the discus, water, and arrival. It is amazingly noxious in its inorganic shape [18].As presence in drinking water cause local health problems in different districts of Sindh.it causes more effect in the age group 11 to15[9].It is investigated that the usually nutritional ingestion of As and related health hazards materializing the population due to eating of regular foodstuffs in the Bangladesh[2]. The total safe limit for

contamination of Asi.e. target hazard quotient THQ in vegetables and cereals for both the rural and urban residents is defined as (>1) [5]. Rice contributes significantly to As exposure of humans through the diet, because As is mobilized from paddy fieldsoils and accumulated by the rice plant [1].The cancer risk in countries with large consumption of rice isincreasing. Cancer riskfrom As showed the comparative risk involvement from water to be 51%, from rice to be 44% and 5% from wheat intake as well [3]. Authors in study [8] collected720 Ground Water and Surface Water samples from 18 different sites of Sindh province, the estimates of As in groundwater and surface water was observed in the range of 0 to 125 and 0 to 35 µg/L with mean values of 46.8 and 15.43 µg/L respectively [8].Effects of As on human health become the global concentration because of increasing contamination in the water crop and soilin many areas [9].The results of investigation indicate that some rice varieties sold in the local markets of AlmadinahAlmunawarah,KSA contain hazardous levels of one or more of the toxic element (e.g.As, chromium,lead)[10]. The higher As rate in cereals was considered due to the higher growth of As from soil to cropin study [3].A survey is done in Khairpur andMatiariand increasing levels of As contamination wereobserved which is not good for human health. According to [13] 81% of water samples were not suitable for drinking in Qambar district and based on [14],[15] it can be said that quality of ground water has also worsen in Larkana due to As contamination. 24 fields of T.M khan were selected and each entrance of the water canals sample was collected.0.12-0.16 mg/kg of Aswas calculated by analytical methods [16]. The total mean of Ascontamination in irrigation water of Badin district was found 0.12-0.14 [17]. Researches discuss until now have detected the value of As in rice or water with the help of chemical process but researchers in [25] have used machine learning to create awareness regarding As.They have performed survey and questionnaire was filled based on which predictions were made regarding awareness of As contamination. A recent development in machine learning had provided great support in the earth and environmental science like in [27] researcher have collected various samples of groundwater and applied ANN and SVM for prediction where SVM performed well, apart from prediction forecasting is also one of the major techniques so in [18] authors forecasted the As value in water and cancer risk due to that for next years in Tando Mohammad khan district Pakistan using techniques such as ARIMA. Our work uses basics of the research presented in [18].We have created a machine learning model and use the As in water values and predict the values of As in rice.

III. DATASET

For the purpose of prediction datasets are collected from different research papers. As in rice for district Badin is collected from [17] and for district T.M khan is collected from [16] and water usage from study [5]. 200 samples of each feature are taken respectively.

IV. AZURE ML STUDIO

Azure Machine learning studio is GUI based IDE for machine learning workflow; creating and optimizing machine learning operation. It is made up of modulesfor performing variousmachine learning operations and these modules are drag and drop based with data flow from one module to another for example if you have dataset you want to clean it then first you will drag and drop dataset module cleaning module and then connect them accordingly. For our research we have used various algorithms which are also Azure ML studio modules [6].

1. Algorithms

From the dataset used in this research, the data is in continuous form, so we have used regression to predict the future values. We have used following regression algorithm modules from Azure Machine Learning Studio.

- Neural Network Regression
- Bayesian Linear Regression
- Boosted Decision Tree Regression
- Decision Forest Regression
- Linear Regression
- Poison Regression

1. 1. Neural Network Regression

It supervises neural Network Regression machine learning method, so it requires a tagged or labeled dataset, it creates regression model with customizable neural network algorithm. It is a suitable fit where the normal regression technique doesn't work so if regression is to be performed on continuous data and other techniques not seem to work then Neural Network Regressionis appropriate to use [28].

1.2. Bayesian Linear Regression

When we are creating a regression model on the continuous data with the help of Bayesian statistics, we are using Bayesian linear regression. In this module it uses linear regression with prior probability function [28].

1.3. Boosted Decision Tree Regression

We use this module when we want to create an ensemble of trees of regression with boosting that means that current tree depends on previous trees. This module improves accuracy with minimum risk [28].

1.4. Decision Forest Regression

We use this module when we want to create an ensemble of decision tree. It is better to use them because they provide efficient use of memory and computation [28].

1.5. Linear Regression

It is supervised machine learning technique; it works in such a way that a linear relationship is established between one or two independent variables to get predictions which will be numeric [28].

1.6. Poison Regression

It creates a regression model with the assumption that the data given has a poison distribution. The ideal case to use this is when predicting numeric values specially counts [28].

V. METHODOLOGY

The purpose of the research is to predict the in rice from the amount of As in water which is used in cultivation of rice so the dataset we have is continuous data so we are using regression for future predictions. For applying this whole workflow we are using Azure Machine Learning Studio. Following are the steps which we have followed in experiments.

- Data Gathering
- Importing Data
- Feature and Label Selection
- Splitting Into Train and Test set
- Applying Machine Learning algorithm
- Evaluates results from all algorithms
- Chose best model
- Create Web Service

The first step is data collection; which is done from the previous researches where either the value of As is detected or it is checked that how much As is present in rice crop and how much water is used. Table 1 shows that As in rice and water for T.M khan and Badin district are taken form paper [16],[17] respectively. Second step is to import data where we want to use it here in this case we are using azure ml studio so data is imported there, now we have data next step is to recognize what are the features and what are the labels in dataset, after getting appropriate features and label we are splitting the data into train and test in next step we will take train data and train machine learning model on that now we will evaluate the models using the test data now based on our evaluation we will select the model and in last step we will create web service using the best model.

Table 1 Data Collection Papers

Paper	District	As in Rice *mg/kg*	As in Water $\mu g/l$
Chohan, Muhammad, et al [16]	T.M Khan	0.12-0.16	51.47
Chohan, M., et al.[17]	Badin	0.12-0.14	57.12

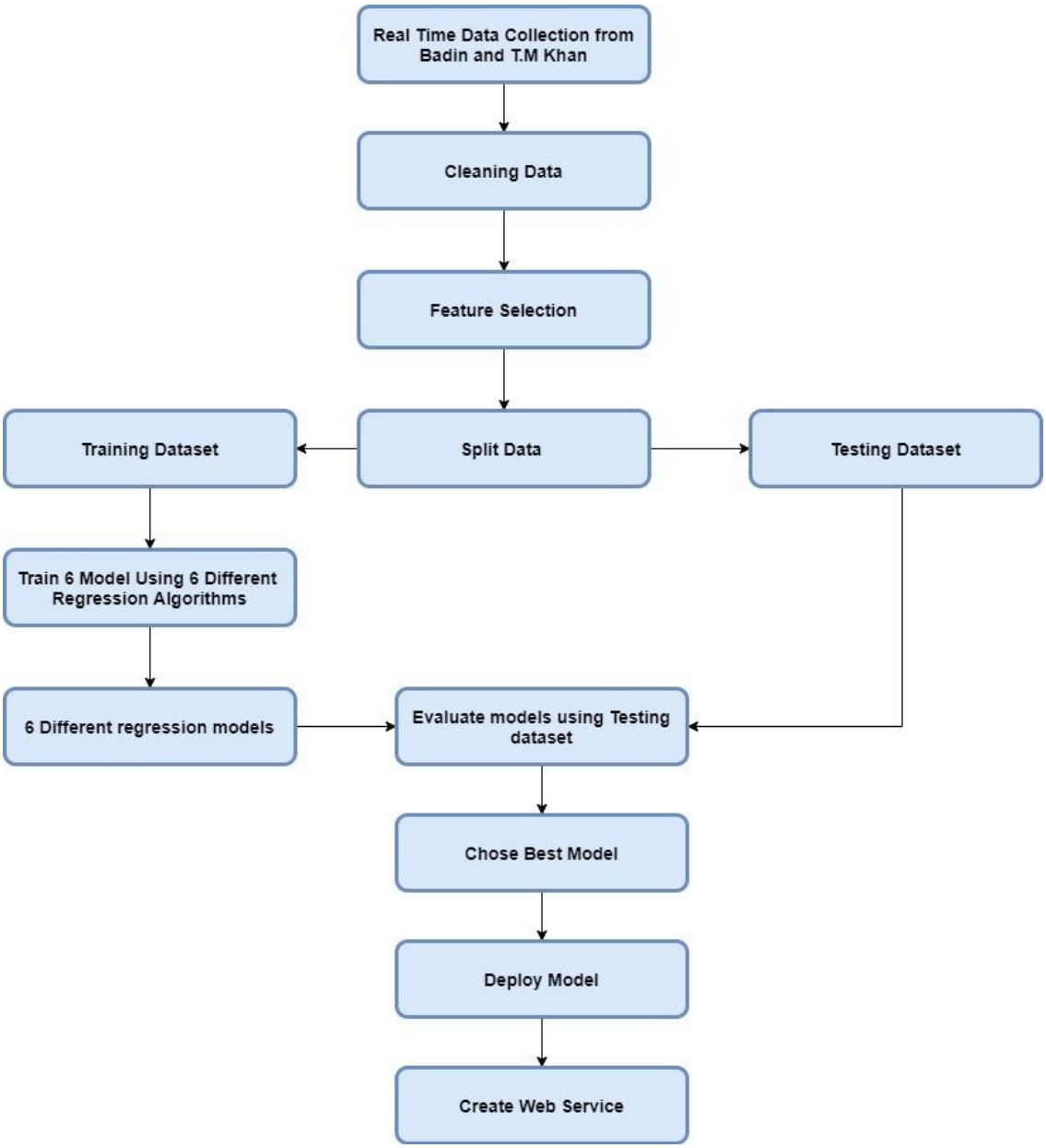

Fig.1. Methodology

VI. IMPLEMENTATION

Experimentis created in azure machine learning studio and collected dataset isimported using dataset module.After having data imported in our experiment; features and labels are selected from dataset using select column from dataset module of ML Studio.Two features are: Column 1 containing value of As in water and column 2 containing water consumption are chosenand one label is: column 3 containing values of As in rice is chosen.Then dataset is split into two parts:80%of data for training and 20% of data for testing using Split Data module.Next step is to apply machine learning algorithm to train our model.We have used Neural Network Regression, Bayesian Linear Regression, Boosted, Decision Tree Regression, Decision Forest Regression, Linear Regression, Poison Regression modules.Forevaluation ofour model we are using score and test model modules and got the following evaluation parameters

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Relative Absolute Error (RAE)
- Relative Squared Error (RSE)

We have selected (RMSE) as our primary metric and based on this we have selected the best model and after selecting the best model we have created web service from that model. Fig 2 shows the implementation of complete workflow and fig3 shows the complete web service machine learning workflow.

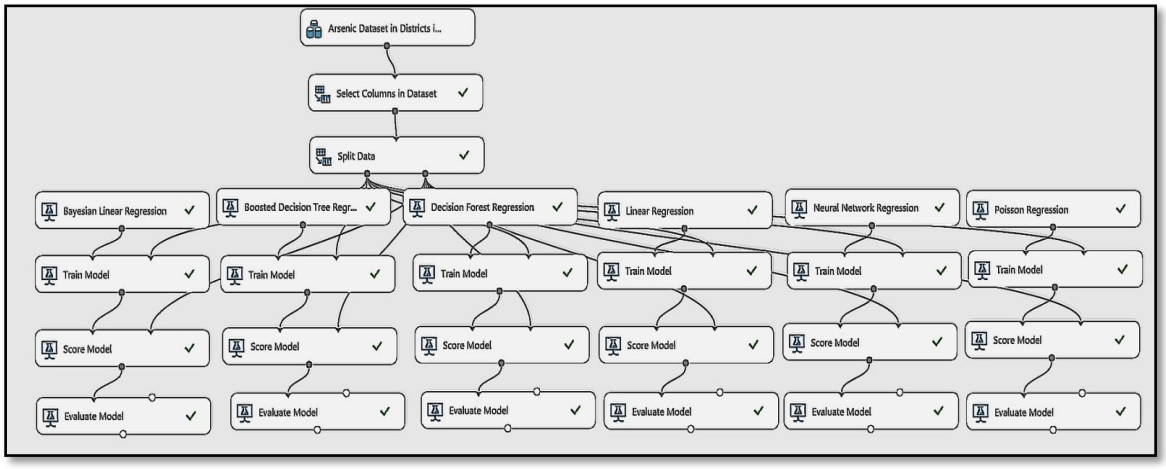

Fig.2. Implementation of machine learning workflow

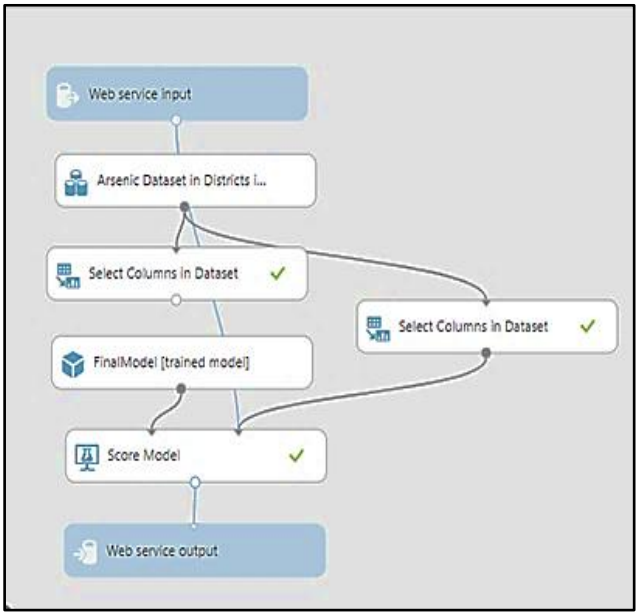

Fig.3. Implementation of Web service using Azure machine learning Studio

VII. RESULT

We have selected RMSE as a primary metric because it contains the normal distribution and does not use absolute values as these values are not convenient for the calculation.

Mean absolute error (MAE)

Is a measure of errors between paired observations expressing the same phenomenon [29].

$$\text{MAE}=\frac{\sum_i^m \boldsymbol{ai}-\boldsymbol{bi}}{\boldsymbol{m}} \qquad (1)$$

Root mean squared error (RMSE)

Creates a single value that summarizes the error in the model. By squaring the difference, the metric disregards the difference between over-prediction and under-prediction [29].

$$\text{RMSE}=\sqrt{\frac{\sum_i^m(\widehat{\boldsymbol{ai}}-\boldsymbol{ai})^2}{\boldsymbol{m}}} \qquad (2)$$

Relative absolute error (RAE)

Is the relative absolute difference between expected and actual values; relative because the mean difference is divided by the arithmetic mean [29].

$$\delta = \left|\frac{\vartheta_A-\vartheta_E}{\vartheta_E}\right|.100\% \qquad (3)$$

Relative squared error (RSE)

Similarly normalizes the total squared error of the predicted values by dividing by the total squared error of the actual values[30].

$$\text{RSE} = \frac{\sum_{i=1}^{m}(Q_{im} - R_i)^2}{\sum_{i=1}^{m}(R_m - \bar{R})^2} \qquad (4)$$

$$\bar{R} = \frac{1}{m}\sum_{i=1}^{m} R_m \qquad (5)$$

Data of As in rice, As in water and water usage for rice crop per 1kg of two districts Badin and Muhammad Khan is used as a dataset. Both targeted districts produce an extensive amount of rice crop in Sindh. Six machine learning algorithms are used to predict the value of As in rice. The results of those algorithms are evaluated based on four parameters. But to choose the best model the primary metric is MSE which is one of the most common way to evaluate a regression problem.

Table.2 shows the value of RMSE for each algorithm, the minimum error value shows the best predictive algorithm and its clear from the table poison regression, Linear Regression and Neural Network Regression have outperformed the others.

In the figure 4, the graph shows the evaluation of parameter's value for each model. From these results the best model for our work is linear regression with MSE of 0.1144 although it can also be seen that poison regression is not far away. In addition the MAE, RAE and RSE values of both algorithms are equal. RMSE, RAE and RSE of Decision tree regression and decision forest regression are highest out of all algorithms. Figure 5 shows the interface on which we input As value in water and amount of water used to irrigate a kg of rice and it will give the predicted value of As in rice using logistic regression algorithm because it outperformed all algorithms, it is being used by the web service we have developed. The purpose of this interface is to provide an easy to use interface for peoplewithout letting them access the web service directly.

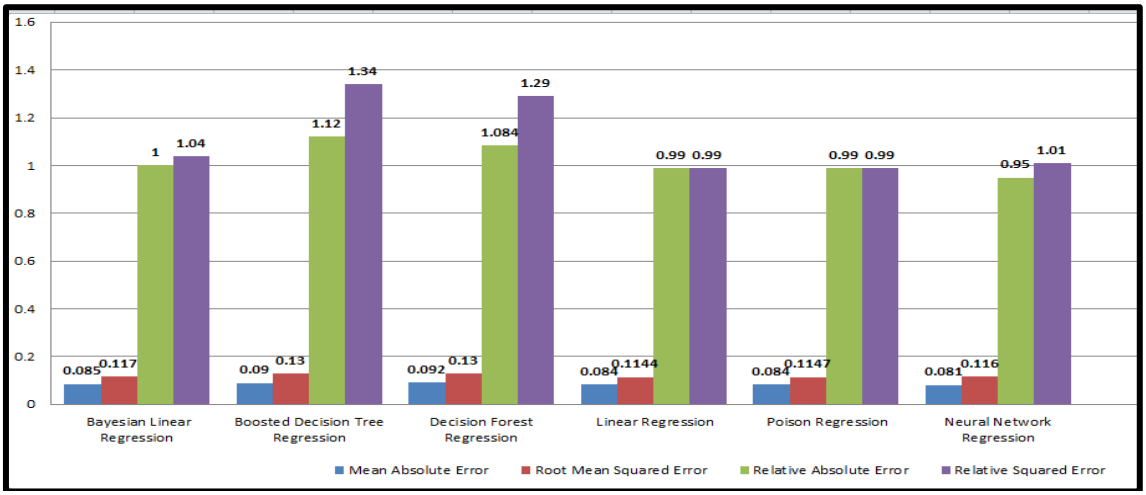

Fig.4. Results for various metrics and Algorithms

Table.2 Algorithms and Primary metric value

S.NO.	Algorithm	Root Mean Squared Error
1.	Bayesian Linear Regression	0.117
2.	Boosted Decision Tree Regression	0.13
3.	Decision Forest Regression	0.13
4.	Linear Regression	0.1144
5.	Poison Regression	0.1147
6.	Neural Network Regression	0.116

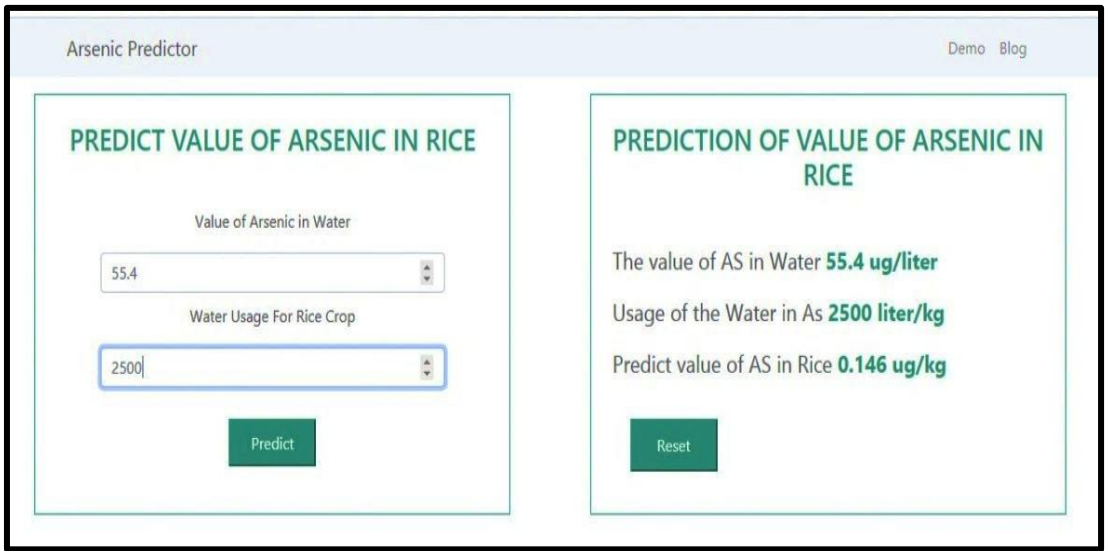

Fig.5. Web Service Prediction Interface

VIII. CONCLUSION

As is a heavy metal whose presence in water and crops causes cancer.Presence of As in water is common, which makes As detection and prediction an important research in this region. In related work we have seen various researches done in south Asia and Pakistan and in Sindh regarding As or detection or prediction of As which gives us strong base to present our research. Our research uses machine learning to predict the value of As in rice from water usage and value of As in water. We have chosen rice because it is extensively used in this region. We have used Azure Machine learning platform to perform to implement six different algorithms for prediction and then evaluated their performance. Linear Regression performed best based on the primary metric, which is RMSE with the value of 0.114. We have also created a web service so that researchers and developers can use our model. Government and other organization can also build their solutions on the top of our web service. Our research opens the gate for future research in prediction of As in food and vegetables.

IX. FUTURE WORK

In this research we have focused the As contamination in rice grain of two different districts as the water containing As[21] is used in rice crops. Our future study will focus on the As contamination in different grains and vegetables [11] produced in different districts of Sindh.

ACKNOWLEDGEMENT

Microsoft has approved this research proposal, for the Earth Grant. Microsoft Azure service area is used in this research to build a prediction model with machine learning techniques and methods for As contamination in rice in Pakistan

REFERENCES

[1]. Sultana, Nusrat. "Studies on Arsenic and Lead Contamination in Crops and Vegetables of Haris chandrapur Village of Jessore District in Bangladesh." Diss. Khulna University of Engineering & Technology (KUET), Khulna, Bangladesh, 2017.

[2]. Ahmed, Md Kawser, Nazma Shaheen, Md Saiful Islam, Md Habibullah-Al-Mamun, Saiful Islam, and Cadi Parvin Banu. "Trace elements in two staple cereals (rice and wheat) and associated health risk implications in Bangladesh." *Environmental monitoring and assessment* , Vol.187, No. 6 , pp.326, 2015 .

[3]. Rasheed, Hifza, Rebecca Slack, Paul Kay, and Yun Yun Gong. "Refinement of arsenic attributable health risks in rural Pakistan using population specific dietary intake values." *Environment international*,Vol. 99 pp. 331-342, 2017.

[4]. Ahmed, Md Kawser, Nazma Shaheen, Md Saiful Islam, Md Habibullah-Al-Mamun, Saiful Islam, Md Monirul Islam, Goutam Kumar Kundu, and Lalita Bhattacharjee. "A comprehensive assessment of arsenic in commonly consumed foodstuffs to evaluate the potential health risk in Bangladesh." *Science of the Total Environment* ,Vol.544, pp. 125-133, 2016.

[5]. Islam, Md Saiful, Md Kawser Ahmed, Md Habibullah-Al-Mamun, and Dennis Wayne Eaton. "Arsenic in the food chain and assessment of population health risks in Bangladesh." *Environment Systems and Decisions* , Vol. 37, No. 3, pp. 344-352, 2017.

[6]. Etaati, Leila. "Azure Machine Learning Studio." In Machine Learning with Microsoft Technologies, pp. 201-223. Apress, Berkeley, CA, 2019.

[7]. Bhatti, Zulfiqar, Khadija Qureshi, Inamullah Bhatti, Imran Nazir Unar, and Mohammad Yar Khuhawar. "Determination of Arsenic and Health Risk Assessment in the Ground Water of Sindh, Pakistan." 2017

[8]. Shahab, Asfandyar, Shihua Qi, and Muhammad Zaheer. "Arsenic contamination, subsequent water toxicity, and associated public health risks in the lower Indus plain, Sindh province, Pakistan." *Environmental Science and Pollution Research*, Vol, 26, No. 30 pp. 30642-30662, 2019.

[9]. Baig, Jameel Ahmed, Tasneem Gul Kazi, Muhammad Ayaz Mustafa, Imam Bakhsh Solangi, Mirza Junaid Mughal, and Hassan Imran Afridi. "Arsenic exposure in children through drinking water in different districts of Sindh, Pakistan." *Biological trace element research*, Vol.173, no. 1 pp.35-46, 2016

[10]. Shraim, Amjad M. "Rice is a potential dietary source of not only arsenic but also other toxic elements like lead and chromium." *Arabian Journal of Chemistry*,Vol 10 pp. S3434-S3443, 2017.

[11]. Arain, M. B., T. G. Kazi, J. A. Baig, M. K. Jamali, H. I. Afridi, A. Q. Shah, N. Jalbani, and R. A. Sarfraz. "Determination of arsenic levels in lake water, sediment, and foodstuff from selected area of Sindh, Pakistan: estimation of daily dietary intake." *Food and Chemical Toxicology*, Vol.47, no. 1,pp. 242-248, 2009.

[12]. Alamgir, Aamir, Moazzam Ali Khan, Janpeter Schilling, S. Shahid Shaukat, and Shoaib Shahab. "Assessment of groundwater quality in the coastal area of Sindh province, Pakistan." *Environmental Monitoring and Assessment*, Vol.188, No. 2 p.78.

[13]. Lanjwani, Muhammad Farooque, Muhammad Yar Khuhawar, Taj Muhammad Jahangir Khuhawar, Abdul Hameed Lanjwani, Muhammad Saqaf Jagirani, Abdul Hameed Kori, Imran Khan Rind, Aftab Hussain Khuhawar, and Jagirani Muhammad Dodo. "Risk assessment of heavy metals and salts for human and irrigation consumption of groundwater in Qambar city: a case study." *Geology,Ecology,and andscapes*, Vol.4, No. 1 pp. 23-39, 2020.

[14]. Kori, Abdul Hameed, Mushtaque Ali Jakhrani, Sarfaraz Ahmed Mahesar, Ghulam Qadir Shar, Muhammad Saqaf Jagirani, Abdul Raheem Shar, and Oan Muhammad Sahito. "Risk assessment of arsenic in ground water of Larkana city." *Geology, Ecology, and Landscapes*, Vol.2, no. 1, pp.8-14, 2018.

[15]. Ali, Waqar, Nisbah Mushtaq, Tariq Javed, Hua Zhang, Kamran Ali, Atta Rasool, and Abida Farooqi. "Vertical mixing with return irrigation water the cause of arsenic enrichment in groundwater of district Larkana Sindh, Pakistan." *Environmental Pollution*, Vol.245, pp.77-88 2019.

[16]. Chohan,Muhammad, Mehrunisa Memon, Inayatullah Rajpar, and Muhammad Saleem Akhtar. "48. Arsenic III, V and total in canal irrigation water and transported load to rice fields of district Tando Muhammad Khan, Sindh-Pakistan." *Pure and Applied Biology (PAB)*, Vol.9, No. 1, pp.491-500, 2020.

[17]. Chohan, M., M. Memon, I. Rajpar, A. A. Khooharo, M. I. Kumbhar, and H. Kakar. "Arsenic transport in canal water and across rice fields in district Badin." *Indian Journal of Science and Technology*, Vol. 13, No. 14 ,pp.1505-1511, 2020.

[18]. Bhatti, Sania, Mohsin A. Memon, and Zulfiqar A. Bhatti. "Groundwater Arsenic and Health Risk Prediction Model using Machine Learning for TM Khan Sindh, Pakistan." 2020.

[19]. Naseem, Sadaf, and John M. McArthur. "Arsenic and other water-quality issues affecting groundwater, I ndus alluvial plain, P akistan." *Hydrological Processes*, Vol 32, No. 9 , pp.1235-1253, 2018.

[20]. Baig, Jameel A., Tasneem G. Kazi, Abdul Q. Shah, Hassan I. Afridi, Ghulam A. Kandhro, Sumaira Khan, Nida F. Kolachi et al. "Evaluation of arsenic levels in grain crops samples, irrigated by tube well and canal water." Food and chemical Toxicology, Vol.49, No.1,pp. 265-270, 2011.

[21]. Shahab, Asfandyar, Qi Shihua, Audil Rashid, Faizan Ul Hasan, and Muhammad Tayyab Sohail. "Evaluation of Water Quality for Drinking and Agricultural Suitability in the Lower Indus Plain in Sindh Province, Pakistan." *Polish Journal of Environmental Studies*, Vol.25, no. 6, pp. 2563-2574, 2016.

[22]. Zavala, Yamily J., and John M. Duxbury. "Arsenic in rice: I. Estimating normal levels of total arsenic in rice grain." *Environmental Science & Technology*, Vol 42, no. 10 pp. 3856-3860, 2008.

[23]. Uqaili, A. A., A. H. Mughal, and B. K. Maheshwari. "Arsenic contamination in ground water sources of district Matiari,Sindh." *International Journal of Chemical and Environmental Engineering*, Vol 3, no. 4, 2012.

[24]. Chandio, Abbas Ali, Yuansheng Jiang, Abrham Tezera Gessesse, and Rahman Dunya. "The nexus of agricultural credit, farm size and technical efficiency in Sindh, Pakistan: A stochastic production frontier approach." *Journal of the Saudi Society of Agricultural Sciences*, Vol.18,No.3, pp.348-354.

[25]. Singh, Sushant K., Robert W. Taylor, Mohammad Mahmudur Rahman, and Biswajeet Pradhan. "Developing robust arsenic awareness prediction models using machine learning algorithms." *Journal of environmental management*, Vol. 211 ,pp. 125-137, 2018.

[26]. Memon, Qurat Ul Ain, Shoaib Ahmed Wagan, Tufail Ahmed Wagan, Irfan Hussain Memon, Zohaib Ahmed Wagan, Hina Memon, Zahoor Ahmed Wagan et al. "Economic Analysis of Hybrid Rice in Taluka Golarchi District Baddin Sindh, Pakistan." International Journal of Business and Economics Research, Vol.4, no. 6, pp.250-255, 2015.

[27]. Park, Yongeun, Mayzonee Ligaray, Young Mo Kim, Joon Ha Kim, Kyung Hwa Cho, and Suthipong Sthiannopkao. "Development of enhanced groundwater arsenic prediction model using machine learning approaches in Southeast Asian countries." *Desalination and Water Treatment*, Vol 57, no. 26, pp.12227-12236.

[28]. ML Azure "Machine Learning Model Regression", Available At:docs.microsoft.com. Accessed on 7 July 2020.

[29]. ML Azure "Evaluate Model", Available At:docs.microsoft.com. Accessed on 7 July 2020

Nazia Pathan, et. al. "Prediction of Arsenic (As) contamination in rice grain produced in Sindh using Machine Learning." *International Journal of Engineering Science Invention (IJESI)*, Vol. 09(10), 2020, PP 01-08. Journal DOI- 10.35629/6734