Water Potability Prediction using Synthetic Minority Oversampling Technique with KNN and MICE Imputer

Ranjeetsingh Suryawanshi[1], *Aryan Karande*[2], *Harshita Bhagat*[2], *Parth More*[2], *Vaishali More*[2]

1. *Department of Multidisciplinary Engineering, Vishwakarma Institute of Technology, Pune*
2. *Department of Information Technology, Vishwakarma Institute of Technology, Pune*

***Abstract*—** Drinking safe and pure water is very essential to ensure public health and safety. Consumption of contaminated water can cause severe health problems and diseases like cholera, typhoid, dysentery, diarrhea, and polio etc. In this paper the implemented system is able to determine whether the water is potable or not to ensure pure and safe drinking water. Existing systems are implemented using algorithms like SVM, KNN, and Decision Tree etc. Machine learning models implemented in these systems face issues like overfitting of model with an imbalanced dataset and computational complexity issue with high dimensional dataset. To solve the above problems this paper gives an optimal solution by using Synthetic Minority Oversampling Technique for balancing imbalance dataset and proper feature engineering to reduce complexity of model. Furthermore, for handling NA values MICE and KNN Imputer have been used. This system is implemented using efficient machine learning algorithms like KNN, Random Forest, Logistic Regression, Decision Tree, SVM Linear, SVM Radial, and XGBoost. From the results, using Random Forest with KNN Imputer and SMOTE gives the highest accuracy of 91.9%.

Keywords- *Machine Learning, Features Engineering, SMOTE, MICE, KNN, Random Forest, Logistic Regression, Decision Tree, SVM, XGBoost.*

I. Introduction

Water quality has considerably declined over the past few decades due to pollution and numerous other problems. For the deterioration of water, the main reason is the release of pollutants into rivers.Polio, typhoid, hepatitis, dysentery, cholera, and diarrhea are just a few of the illnesses that have been linked to contaminated water and poor hygiene. Having access to clean, plentiful water sources is important since the quality of drinking water has a big impact on people's health. According to the United Nations Environment Program (2000), 50% of people worldwide lack access to good sanitation systems and 20% of people lack access to clean drinking water, which poses a severe danger to water shortages and diseases brought on by waterborne pathogens. The water reservoirs need to be refilled with 64 billion m3 of water a year to accommodate the annual rise of about 60 million people[3]. In hospitals around the world, 15% of patients contract a virus during their stay, though this percentage rises dramatically in lower regions. Drinkable water selection requires careful consideration.[4]

As a result, over the past ten years, forecasting water quality has become an extremely popular subject. Calcium, industrial waste and many other parameters affect the quality of water. Hardness of water also causes heart diseases. Many physical, chemical and radioactive sources cause water pollution[15]. People are exposed to avoidable health risks when water and sanitation services are unavailable, inadequate, or improperly handled. Many algorithms are not suited for an unbalanced dataset to solve this problem; the proposed system has implemented the SMOTE technique. Through Boxplot it was visualized that data containing outliers and outliers affect the performance of the model, so we removed the outliers using Interquartile range. The process of choosing, transforming, and extracting pertinent features from unprocessed data so that they can be used to teach machine learning models is known as feature engineering[9]. In order to increase the accuracy of the algorithms implemented in the proposed system, feature engineering is used.

Purpose of this project is to avoid health risks caused by drinking water using machine learning techniques. Following are the objectives of the proposed project:

- To understand the quality of water and distinguish between potable and non-potable water.
- To resolve overfitting issues by balancing the data using the Synthetic Minority Oversampling Technique (SMOTE).
- To Predict water potability by using KNN, Random Forest, XGBoost, Decision Tree, logistic Regression.
- To solve complexity problems using Feature Engineering.

II. Literature Review

The authors of this paper proposed a project to achieve better accuracy than the existing system as they fall short of the accuracy. To deal with the missing values this study makes use of KNN Imputer. This project is implemented using stacked ensemble H2O AutoML to handle the accuracy problem. They have used SHAP (SHapley Additive exPlanations) to understand the contribution of each feature.Maximum accuracy obtained in this project is 97%. Main problem faced during implementation is overfitting due to the imbalance dataset [1].This paper proposed an Artificial Neural Network (ANN) to predict water quality levels and water consumption. It has an accuracy of 96% and outperforms existing models in terms of prediction performance.It was also concluded that simple network architecture performed well in comparison to complex models (CNN,LSTM and GRU: Gated Recurrent Unit) [2].This project is designed to predict water quality accurately. To resolve the issue due to missing values a nine-layer MLP is implemented which deals with the same. Algorithms like RF, XGBoost, SGDC, LR are compared. Accuracy of the model using MLP is 99.9%. Overfitting is the main problem faced with the imbalanced dataset[3].Paper shows comparative analysis of different ML approaches like SVM, Decision Tree, Random Forest, Gradient Boost, and AdaBoost. SMOTE is used to balance the dataset. To determine important features Explainable AI (XAI) is used. RF and Gradient Boost have best accuracy i.e. 81% both. Lack of transparency is the major issue in this model[4].This paper studies the irrigation water quality indices in Nand Catchment area of Rajasthan where above 70% of agriculture and 94% of drinking water is dependant on ground water.Five ML models namely Linear Regression, Additive Regression, Random Subspace, REPTree and SVM were proposed by the authors.SVM showed highest efficiency in predicting the water quality in catchment[5].

Several ML models, such as QRF, Random Forest, radial SVM, Stochastic Gradient Boosting , and GBM H2O were investigated for predicting biochemical oxygen demand (BOD). The QRF model achieved the best performance, according to the results, among the developed models.[6].In this study, hybrid learning algorithms are used to simulate and analyze more than 30 years of observed data using two ML techniques, ANN and SVM. The findings show that the ANN is more effective for quick responses, whereas the SVM is more accurate but requires more training time[7]A hybrid model by combining ANN and process-based watershed model was proposed For the majority of watersheds, the hybrid model's prediction accuracy was greater.[8].In this research, a method was proposed to check Moroccan groundwater quality data for fourteen parameters for the development and evaluation of using Adaboost, RF, ANN, and SVR . During the validation performance, the ensemble models Adaboost and RF outperformed the conventional ANN and SVR models.[9].Urbanization has caused water quality to decline at an alarming rate, which has resulted in dreadful illness. To find a solution for this problem authors have implemented a system using supervised machine learning algorithms to estimate Water Quality Index. Algorithms are implemented with only four features; it decreases the complexity of the model.Accuracy obtained using MLP is 85.07%. This model can be implemented to handle large scale data obtained from IoT devices [10].

In this paper the water quality prediction system is designed using pattern extraction. Algorithms used are Naïve Bayes, Decision Tree, Random Forest, and Gradient Boosted Trees. DT has best accuracy 87.69%. Different data cleaning techniques can be used to handle issues due to missing values[11].This paper compared the performances of 10 learning models(7 traditional + 3 ensemble) on the water quality of major lakes in China.Conclusions derived from these study were that available big data(33,162 observations)

enhanced the effectiveness of traditional and ensemble models .Amongst all the models DCF(Deep Cascade Forest), DT, and RF showed significant prediction performance[12].According to the analysis in this paper's findings, the Bayesian network is the model that is best suited for predicting real-time water quality because it has a low cross validation error rates and can accurately predict the majority of mode red days. It is ideal for real-time prediction due to its capacity to handle missing values, outliers, and updates.[13]. In this paper they have used different ML techniques like RF, NN, MLR, SVM, and BTM. Random Forest is used for missing data management. MLR has a best accuracy of 99.83%. For further improvement to improve efficiency of the selection process different deep learning approaches can be used[14].In order to predict water quality, a novel approach based on long and short term memory neural networks (LSTM NN) is proposed in this research. Using Taihu Lake's monthly gathered data from 2000 to 2006, a prediction model was developed. Comparatively speaking, LSTM NN has a greater and more universal predictive accuracy. More needs to be done to shorten long training cycles[15].

III. Proposed Methodology

Implementation methods for the proposed model are discussed in this section. Figure 1 represents the data flow of the model. The overall flow is as follows, First selection of the dataset, analysis of the data and data visualization. Feature selection was performed using Heatmap. Further data preprocessing contains imputation of missing data and data oversampling. Spilling of dataset in training data and testing data and modeling with training data. Model predicts the class whether water is potable or not. Further, testing the model using a test dataset. Then analysis of performance of the model using a confusion matrix.

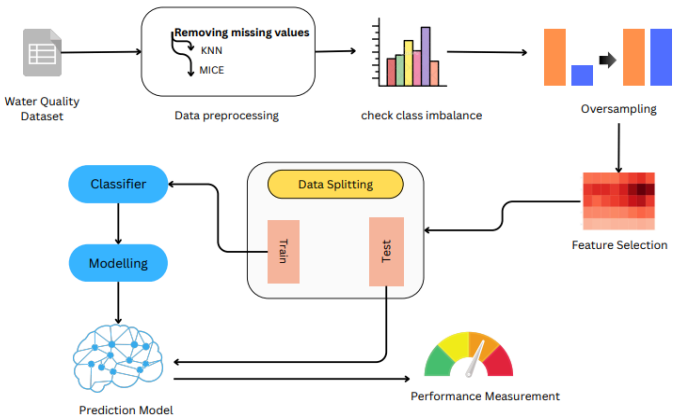

Figure 1: Architecture of the Implemented System

1. Dataset

The dataset used in this project is taken from Kaggle. This dataset contains records of nine features that help to decide potability of water like pH, Solids, Hardness, Turbidity, Conductivity, Chloramines, Sulfate, Organic_Canbon, Trihalomethanes,. Potability is a predicted class with factor data with two classes '1' or '0'. '1' means water is potable and '0' means not potable. 61% of the data are classified as non-potable, while 39% are classified as potable.

2. Data Visualization

The process of presenting data and information in a visual format to understand complex datasets easily is known as data visualization. Box plots are used to visualize dataset distribution and find potential outliers. Outliers are decided based on interquartile range. Data inside the box lies inside the quartile range and data outside the box are outliers. For all The data features most of the data points lie inside the box as shown in Figure 2 most of the pH values lie between 6 to 8.

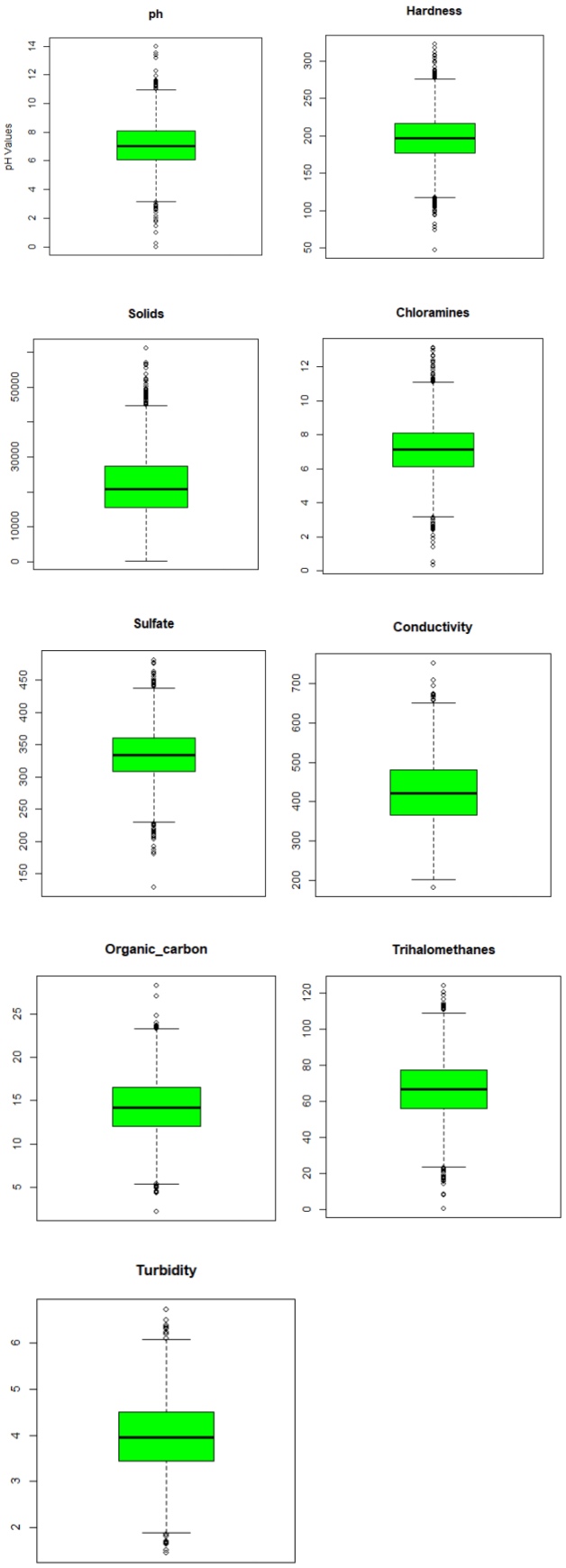

Figure 2: Boxplot for all Features of the Dataset

3. Correlation Analysis and Feature Selection

In this paper using correlation heatmap features are selected for prediction. Sulfate has a strong inverse relationship with potability. The Dataset contains 9 features so data becomes nine dimensional. It increases the computational complexity of the model and affects the accuracy, so in this system solids and chloramines features are removed so the dataset becomes seven dimensional.

4. Data Preprocessing

For data preprocessing following techniques are used:

a. Imputing missing values

There are various resources from where data is collected for analysis. Missing values can be generated during incomplete data entry, human errors and many more. Removing rows containing missing values affects the accuracy of the model as that record may have some important information. So it is important to handle missing values carefully. From the dataset it can be seen that data contains 491, 781, 162 NA values in pH , Sulfate and Trihalomethanes columns respectively. In this project two different imputation techniques are used:

- KNN Imputation

The k-nearest neighbor algorithm is used to impute the missing value.This project is implemented with k as 5.

- MICE Imputation

MICE (Multiple Imputation by Chained Equation) is used for quickly filling in missing values by examining the information from the other columns. This project is implemented using the PMM(Predictive Mean Matching) method.

b. Removing Outliers

Outliers are the points in the dataset which are considerably different from the other data points. These outliers significantly decrease the performance of the machine learning models. As shown in the Figure 2 boxplot shows the outliers in all feature points. So, in this project outliers based on interquartile range are removed from the dataset. It significantly increased the performance of the model.

c. Oversampling using SMOTE

Dataset in this study is imbalanced as shown in Figure 4 the count of class with 0 (not potable) is more than class 1 (potable). The problem with the majority class is that the model learns to predict the class more correctly for the majority class and it results in the poor performance of the model.So, to solve this problem this project is implemented with SMOTE.

5. Implementation of Machine Learning Models

Prediction is the most important part in classification of water potability. Accuray of the model completely depends on modeling. In R there are many libraries which include functions to train the model for prediction. This project is implemented with KNN, Random Forest, Logistic Regression, Decision Tree, XGBoost, Linear SVM and Radial SVM machine learning algorithms. Using k fold cross validation for prediction removes the randomness while splitting the data in training and testing sets. In this system we have used 10 fold cross validation.

IV. Results and Discussion

This project is implemented with various classifier algorithms like Random Forest, KNN, Decision Tree, SVM, XGBoost. Implementation of this project is done in R. In Figure 3 it can be observed that first, the dataset was imbalanced. Percentage of probability count of 1 and 0 class is 39% and 61% respectively. After SMOTE oversampling data barplot count of potable and non-potable class is shown in figure 4. Probability percentage of class 1 and 0 after oversampling is 56% and 43.9% respectively. Balancing the dataset significantly improved accuracy of the model.

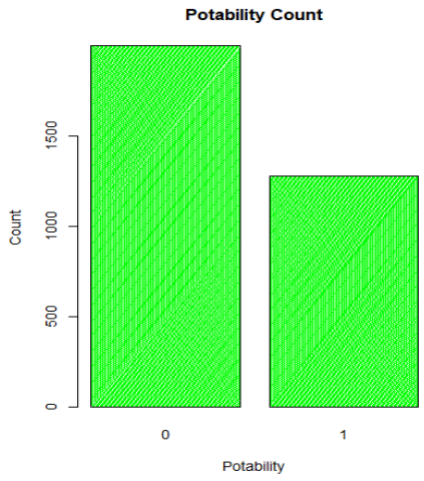

Figure 3: Bar Graph of Potability Count before oversampling

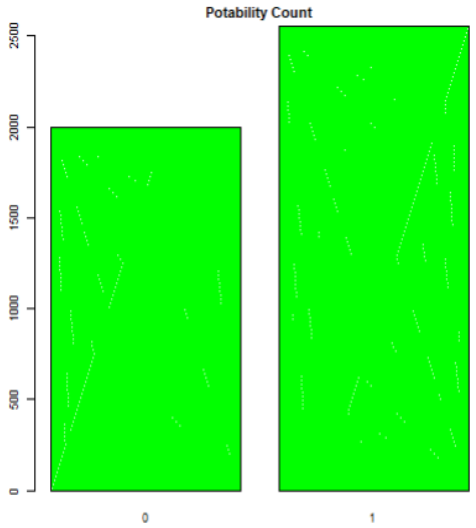

Figure 4: Bar Graph of Potability Count after SMOTE oversampling

Performance of the models are measured using different performance measures like accuracy, precision, recall and F1 score. In the initial series of experiments, two different imputation techniques were used.

Table 1 shows the performance of different models using MICE Imputation. From the results, accuracy values obtained using Random Forest and XGBoost are 91.3% and 82%. Logistic Regression gives the lowest accuracy of 50.3%.

Model	**Accuracy**	**Precision**	**Recall**	**F1**
Random Forest.	0.913	0.897	0.798	0.94
XGBoost	0.82	0.781	0.584	0.666
KNN	0.808	0.86	0.861	0.861
SVM Radial	0.76	0.762	0.946	0.844
Decision Tree	0.696	0.712	0.938	0.809
SVM Linear	0.684	0.709	0.912	0.798

Table 1

Results of the implemented algorithms using KNN imputer.

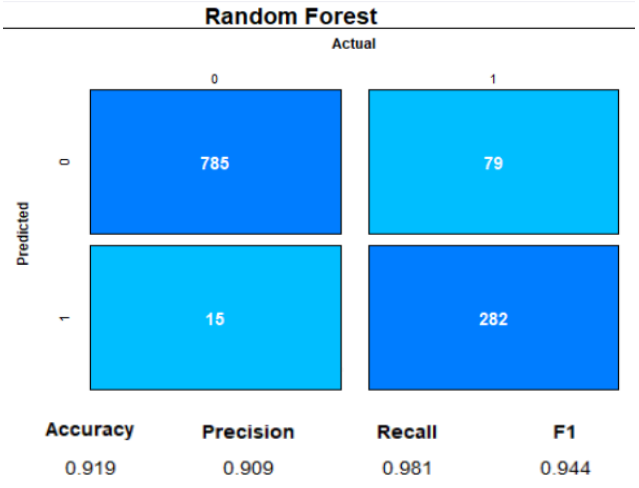

Figure 5: Confusion matrix for Random Forest

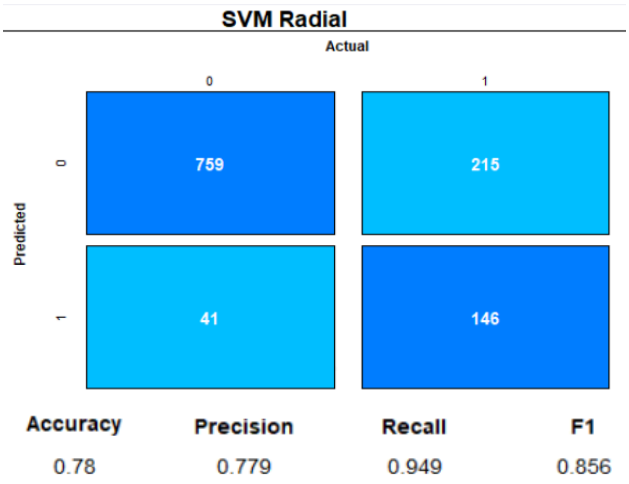

Figure 8: Confusion matrix for SVM Radial

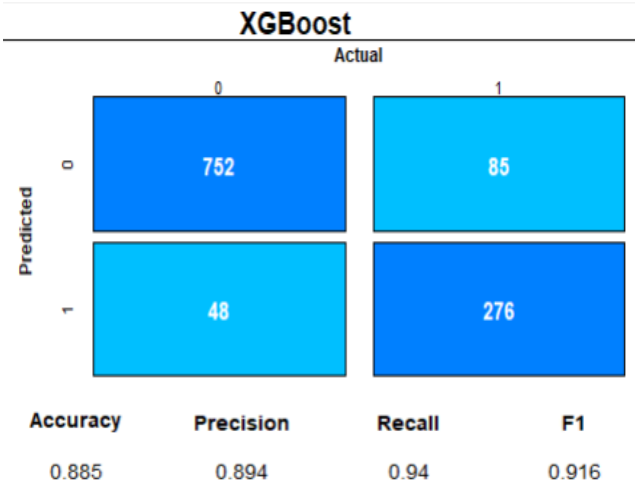

Figure 10: Confusion matrix for XGBoost

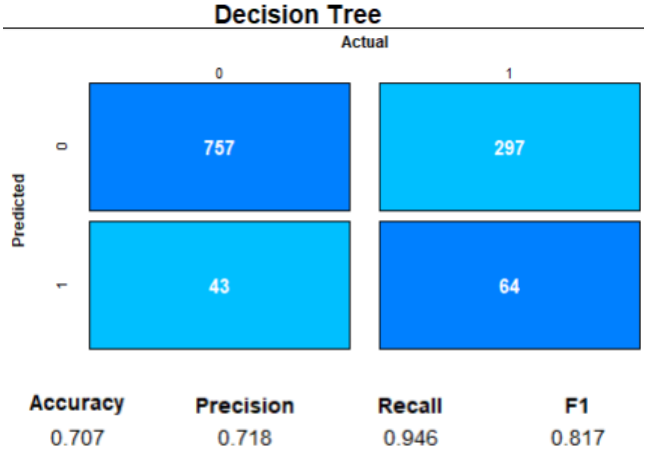

Figure 7: Confusion matrix for Decision Tree

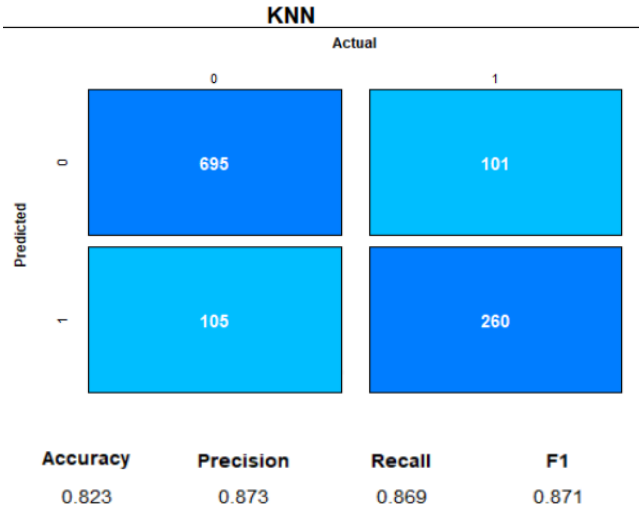

Figure 6: Confusion matrix for KNN

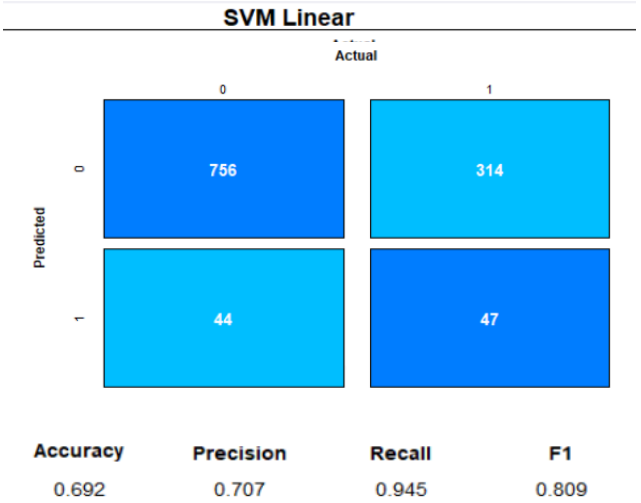

Figure 9: Confusion matrix for SVM Linear

Figure 5-10 shows confusion matrix for Random Forest, XGBoost, KNN, SVM Radial, SVM Linear, Decision Tree algorithms implemented with KNN imputation and SMOTE oversampling technique. Accuracy obtained using Random Forest is highest; it is about 91.9%. After the Random Forest, XGBoost classifier gives the better accuracy of 88.5%. Precision, Recall, F1 score values for random forest are 0.909, 0.981, 0.944 respectively. Linear regression gives the lowest accuracy of 54.2%.

Comparative analysis of the Performance of model using MICE and KNN imputation reveals that the KNN imputation gives better accuracy for all the algorithms compared with the MICE imputation. Table 2 shows comparison of results obtained using KNN and MICE Imputer.

Model	Accuracy	
	KNN imputer	MICE Imputer
Random Forest.	0.919	0.913
XGBoost	0.885	0.82
KNN	0.823	0.808
SVM Radial	0.78	0.76
Decision Tree	0.707	0.696
SVM Linear	0.692	0.684
Logistic Regression	0.542	0.503

Table 2

V. Conclusion and Future Scope

Access to clean and safe drinking water is crucial for the survival and well-being of human beings. Unfortunately, in many areas, the majority of water sources usually don't meet the requirements for water. Consuming impure water can cause many health problems. This project is implemented to solve this problem by predicting where water is potable or not. Experiment was performed for imputing missing values first MICE and then KNN imputation was performed. From the results it can be concluded that KNN imputation gives better performance compared to mice. By feature engineering dataset dimensions were reduced it helped to decrease the complexity of the dataset and Accuracy was significantly increased. To solve the overfitting problem with an imbalance dataset SMOTE technique is used and also 10 fold cross validation helped to solve the overfitting problem. Using SMOTE oversampling and KNN imputation Random Forest obtained best accuracy of 91.9%. In MICE and KNN we need to keep the whole training set on the server hence it becomes computationally complex with high dimensional dataset. In the future, this project can be implemented to increase the accuracy. Some algorithms which work by assigning weights based on importance of features can be implemented.

VI. References

[1] Madni HA, Umer M, Ishaq A, Abuzinadah N, Saidani O, Alsubai S, Hamdi M, Ashraf I., “Water-Quality Prediction Based on H2O AutoML and Explainable AI Techniques." Water. 2023; 15(3):475. https://doi.org/10.3390/w15030475

[2] Rustam F, Ishaq A, Kokab ST, de la Torre Diez I, Mazón JLV, Rodríguez CL, Ashraf I, “ An Artificial Neural Network Model for Water Quality and Water Consumption Prediction”, Water. 2022; 14(21):3359. https://doi.org/10.3390/w14213359

[3] Afaq Juna,Muhammad Umer ,Saima Sadiq ,Hanen Karamti,Ala’ Abdulmajid Eshmawi,Abdullah Mohamed and Imran Ashraf, “Water Quality Prediction Using KNN Imputer and Multilayer Perceptron”, Water, Volume 14, August 2022 https://doi.org/10.3390/w14172592

[4] Jinal Patel, Charmi Amipara, Tariq Ahamed Ahanger, Komal Ladhva, Rajeev Kumar Gupta, Hashem O. Alsaab, Yusuf S. Althobaiti, Rajnish Ratna, "A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI", Computational Intelligence and Neuroscience, vol. 2022, Article ID 9283293, 15 pages, 2022.https://doi.org/10.1155/2022/9283293

[5] Dimple Dimple, Jitendra Rajput, Nadhir Al-Ansari, Ahmed Elbeltagi, "Predicting Irrigation Water Quality Indices Based on Data-Driven Algorithms: Case Study in Semiarid

Environment", Journal of Chemistry, vol. 2022, Article ID 4488446, 17 pages, 2022. https://doi.org/10.1155/2022/4488446

[6] Al-Sulttani, A. O., Al-Mukhtar, M., Roomi, A. B., Farooque, A. A., Khedher, K. M., & Yaseen, Z. M. (2021), "Proposition of New Ensemble Data-Intelligence Models for Surface Water Quality Prediction", IEEE Access, 9, 108527–108541. https://doi.org/10.1109/access

[7] Deng, T., Chau, K.-W., & Duan, H.-F. (2021), "Machine learning based marine water quality prediction for coastal hydro-environment management", Journal of Environmental Management, 284, 112051.doi:10.1016/j.jenvman.2021. 112051

[8] Noori, N., Kalin, L., & Isik, S. (2020) "Water Quality Prediction Using SWAT-ANN Coupled Approach", Journal of Hydrology, 125220.doi:10.1016/j.jhydrol.2020.125220

[9] Theyazn H. H Aldhyani, Mohammed Al-Yaari, Hasan Alkahtani,and Mashael Maashi, "Water Quality Prediction Using Artificial Intelligence Algorithms", Hindawi, Volume 2020, Article ID 6659314 https://doi.org/10.1155/2020/6659314, Dec 2020

[10] El Bilali, A., Taleb, A., & Brouziyne, Y. (2020), "Groundwater quality forecasting using machine learning algorithms for irrigation purposes", Agricultural Water Management, 106625. doi:10.1016/j.agwat.2020.106625

[11] Umair Ahmed ,Rafia Mumtaz ,Hirra Anwar,Asad A. Shah,Rabia Irfan and José García-Nieto, "Efficient Water Quality Prediction Using Supervised Machine Learning", October 2019, Water, Volume 11, https://doi.org/10.3390/w11112210 .

[12] Jefferson L. Lerios, Mia V. Villarica, "Pattern Extraction of Water Quality Prediction Using Machine Learning Algorithms of Water Reservoir", International Journal of Mechanical Engineering and Robotics Research Vol. 8, No. 6, November 2019.

[13] Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R.,Ren, H. (2019), "Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data" Water Research, 115454. doi:10.1016/j.watres.2019.115454

[14] Avila, R., Horn, B., Moriarty, E., Hodson, R., & Moltchanova, E. (2018) "Evaluating statistical model performance in water quality prediction", Journal of Environmental Management, 206, 910–919.doi:10.1016/j.jenvman.2017.11.049

[15] Md. Mehedi Hassan, Md. Mahedi Hassan, Laboni Akter, Md. Mushfiqur Rahman, Sadika Zaman, Khan Md. Hasib, Nusrat Jahan, Raisun Nasa Smrity, Jerin Farhana, M. Raihan, Swarnali Mollick, "Efficient Prediction of Water Quality Index (WQI) Using Machine Learning Algorithms", Human-Centric Intelligent Systems,vol. 1, Issue 3-4, pages 86-97, issn 2667-1336,doi 10.2991/hcis.k.211203.001

[16] Wang, Y., Zhou, J., Chen, K., Wang, Y., & Liu, L. (2017) "Water quality prediction method based on LSTM neural network" 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE). doi:10.1109/iske.2017.8258814