Water Quality Analysis

Sai Sailaja Policharla
Computer Science and Engineering
PES University
Bangalore, India
sailaja.policharla03@gmail.com

Nidhi S Sheth
Computer Science and Engineering
PES University
Bangalore, India
nidhisheth007@gmail.com

Baddela Divya Mallika
Computer Science and Engineering
PES University
Bangalore, India
divyamalika325@gmail.com

***Abstract*—Water is extremely important for humans to survive and function. The presence of certain contaminants in water can lead to health issues which include gastrointestinal illness, reproductive problems, and neurological disorders. Water is mostly contaminated by compounds like - cadmium, mercury, arsenic, lead, fluorine, zinc, barium, nitrates, chlorine etc. Certain concentrations of the minerals is tolerable, but when the concentration goes beyond a threshold value, it can be quite deadly that lead to disturbing the ecosystem that depends on these water bodies, cause various health problems and worse death. Learning underlying features which determine the safety of water and providing insights can help better understand the problem and solve it. Machine Learning can provide to be useful in learning these features and their weightage to be used for prediction in determining the water's safety.**

I. Introduction

The US Safe Drinking Water Act defines the term "contaminant" as meaning any physical, chemical, biological, or radiological substance or matter in water. Therefore, the "contaminant" definition very broadly applies as being anything other than water molecules. Drinking water may reasonably be expected to contain at least small amounts of some contaminants. Some drinking water contaminants may be harmful if consumed at certain levels in drinking water while others may be harmless. The presence of contaminants does not necessarily indicate that the water poses a health risk.

The CCL serves as the first level of evaluation for unregulated drinking water contaminants that may need further investigation of potential health effects and the levels at which they are found in drinking water:

Physical contaminants primarily impact the physical appearance or other physical properties of water. Examples of physical contaminants are sediment or organic material suspended in the water of lakes, rivers and streams from soil erosion.

Chemical contaminants are elements or compounds. These contaminants may be naturally occurring or man-made. Examples of chemical contaminants include nitrogen, bleach, salts, pesticides, metals, toxins produced by bacteria, and human or animal drugs.

Biological contaminants are organisms in water. They are also referred to as microbes or microbiological contaminants. Examples of biological or microbial contaminants include bacteria, viruses, protozoa, and parasites.

Radiological contaminants are chemical elements with an unbalanced number of protons and neutrons resulting in unstable atoms that can emit ionizing radiation. Examples of radiological contaminants include cesium, plutonium and uranium.

In this paper, we are looking at analyzing the chemical and some radiological contaminants and few biological components like viruses and bacteria which affect the quality of water and in turn it's safety. Our solution is to aid the automation of manual processes required to determine if water is safe or not.

The paper has been organised as per the following sections: II. Related Works; Literature survey corresponding to our problem statement, III Proposed Solution; Dataset description, preprocessing, descriptive analysis, training and testing, details on model building and methodologies, IV Results and Conclusion; Experimental results of models and inferences made along with concluding remarks have been detailed.

II. Related Work

Majority of the papers to date which correspond to water quality analysis and prediction deal mainly with features including the pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, biological oxygen demand (BOD), Turbidity to predict the potability of water - 1 means Potable and 0 means Not potable. (0) Water is not safe to drink and (1) Water is safe to drink.

The data pertaining to these tasks are most of the time collected over a particular time period extracted with the help of sensors by monitoring water levels. These sensors focus on features like electrochemical features and sometimes the chlorophyll level to predict the probability of algal bloom.

In [4]Levels of chlorophyll were also considered for algal bloom predictions which involves combination of automatic high-frequency monitoring (AFHM) systems with machine learning (ML) techniques to build a data-driven chlorophyll-a (Chl-a) soft-sensor. Massive data for water temperature, pH, electrical conductivity (EC) and system battery were taken for three years at intervals of 15 min.

In [5] Another paper involves extracting time series data by monitoring water levels through sensors and analyzing data to identify changes or anomalies occurring on water quality time series data. This experiment mainly focuses on the electrochemical composition of water.

These papers mainly focus on training classical classification algorithms like Support Vector Machines, Random Forests, Decision Trees and Artificial Neural Networks for prediction and LSTMs or other time series forecasting models for time series data

In this study, we are focused towards how quantities of chemicals and minerals can affect the safety of water.

III. Solution Approach

The dataset for this study contains around 20 chemicals and minerals like aluminium, ammonia, arsenic, barium, cadmium, chloramine, chromium, copper, fluoride, bacteria, viruses, lead, nitrates, nitrites, mercury, perchlorate,radium,selenium,silver,uranium and a target class "is_safe" which contains binary values indicating if the water is safe for consumption(1 - safe, 0 - not safe). There are around 7999 rows and 21 columns as mentioned above.

A. Data Preprocessing

Dataset was checked for null values, duplicates and outliers. The null values were not present in a general format NA or a blank cell, instead were filled as #NUM!. There were only 3 null values in 2 columns out of which one column was the target. Hence,the corresponding were imputed with required values.

B. Exploratory Data Analysis

Before feeding the data into a model, the data was thoroughly analyzed to understand the significance of each feature, its distribution and resonability towards the target class.
From Fig.1 we observe that for some minerals like aluminium and chloramine majority of the values are below the threshold value which indicates if it is dangerous.

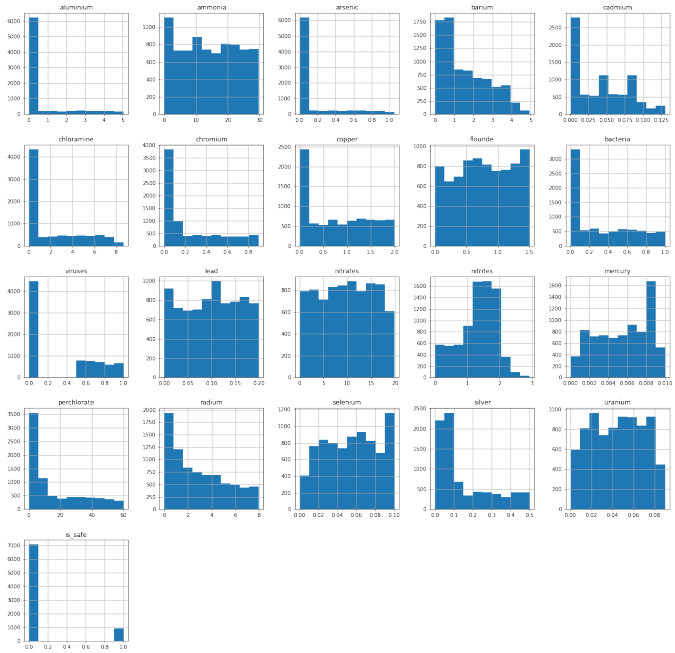

Fig. 1. Distribution of mineral values across the dataset

We also observe in Fig2 there is a class imbalance in the dataset with 88.59% corresponding to 0(not safe) and the rest 11.40% which belongs to 1(safe)

The dataset has around 20 features which are enormous for training and cause a penalty on the metrics as some of them do not contribute to the prediction significantly and there are possibilities of spurious correlation.

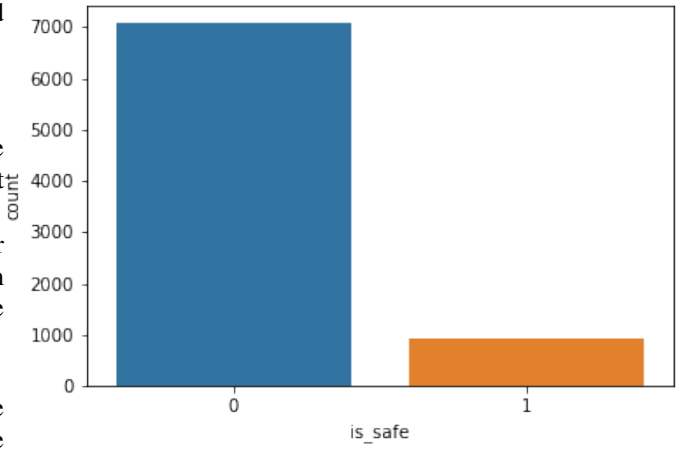

Fig. 2. Class Imbalance

For minerals like ammonia, selenium and uranium there are no values in the dataset that are greater than the threshold value for each of the element.

Due to the presence of a lot of features in the dataset, feature reduction techniques like PCA were applied on the dataset to derive the important.

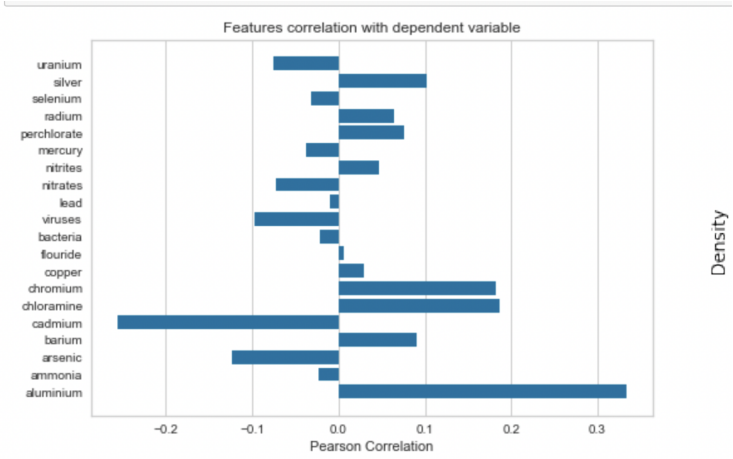

Fig. 3. Feature Correlation with dependent variable

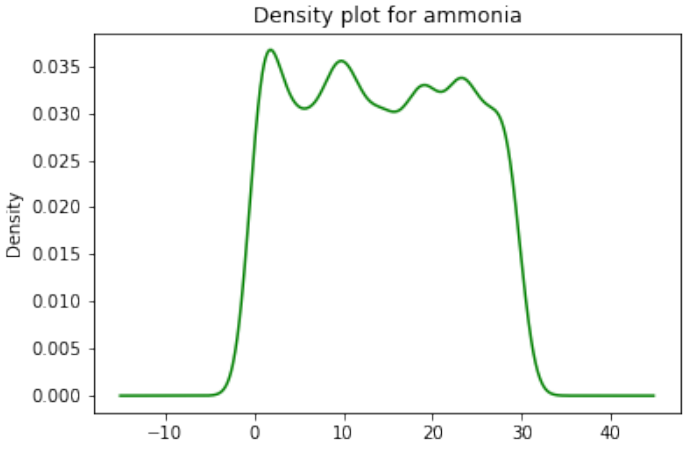

Fig. 4. Density plot for ammonia

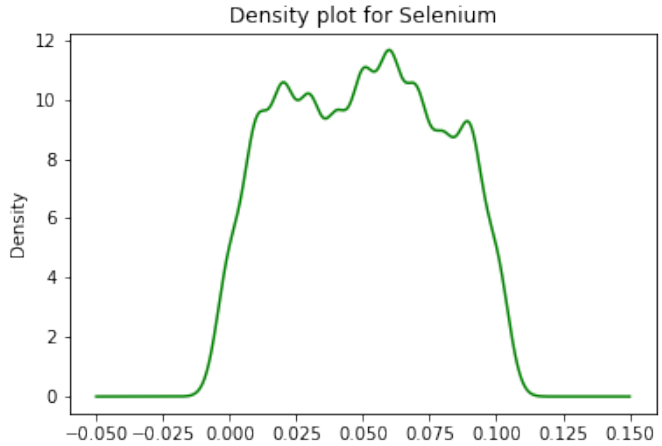

Fig. 5. Density plot for selemium

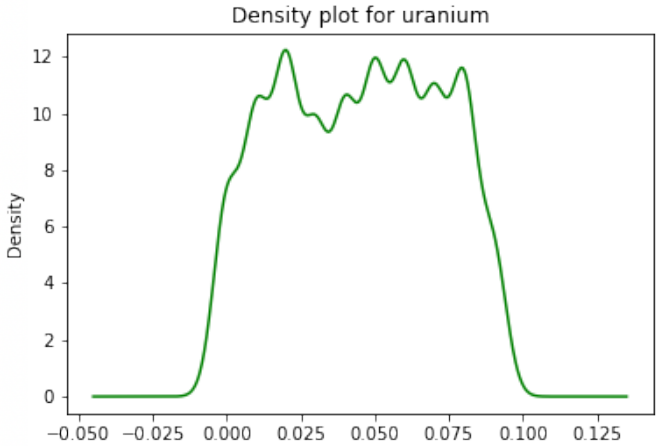

Fig. 6. Density plot for uranium

To deal with the imbalance in the dataset, Synthetic Minority Over-sampling Technique(SMOTE) was used to oversample the minority class i.e 1 and applied to models which were not robust or were sensitive to imbalance classes.

C. Training and Testing

The given dataset was trained and tested on 6 models and were compared with respect to accuracy, recall, precision and various other metrics local to that model and that were relevant to the given problem statement and most robust for water quality analysis.

The train and test dataset was split on a 80-20 ratio and scaling was performed using standard scaler.

Logistic Regression

A logistic regression model was first trained on the plain scaled dataset which gave an accuracy of 90.3%

To deal with high dimentionality of the dataset, PCA was performed on the scaled dataset to extract features that explain 90% of the dataset and around 5 features were obtained. The accuracy dropped after performing PCA, as applying PCA to imbalanced classes

Due to an imbalance in the class, the recall of the minority class is very less though precision is of acceptable number, this led to a lower F1 score. This imbalance not only affected the the metrics of the model but also the PCA done.

Synthetic Minority Over-sampling Technique(SMOTE) was applied after running the PCA which improved the accuracy and other metrics which. The reason for applying PCA before SMOTE is to leverage maximization of within data variability by projecting your data on the planes/axes where the variance of the data will be highest. Now once the variability within your data is really high, it is convenient for a nearest neighbors like algorithm, to over sample the data. This provides much

better quality of synthetic samples which would be more closer to your actual data.

We observe that the recall, specificity and F-1 score of Logistic Regression trained using PCA+SMOTE is higher as compared to the other variants.

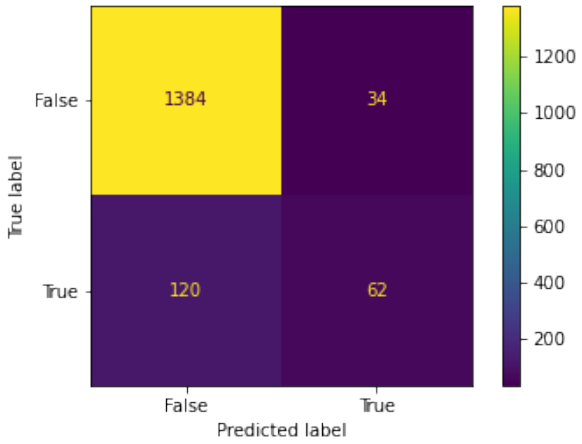

Fig. 7. Confusion Matrix without PCA

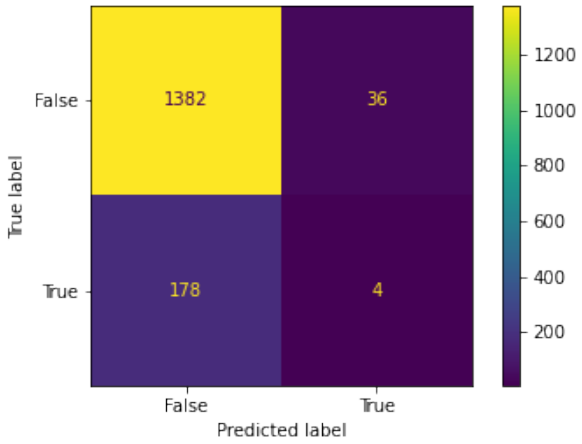

Fig. 8. Confusion Matrix with PCA

Logistic Regression Metrics			
Metrics	Vanilla Logistic regression	PCA	PCA+ SMOTE
Accuracy	0.90375	0.895625	0.895625
Precision	0.64583	0.58823	0.54601
Recall	0.34065	0.27472	0.48901
Specificity	0.97602	0.97531	0.947813
F-1 Score	0.44604	0.37453	0.51594

Support Vector Machine

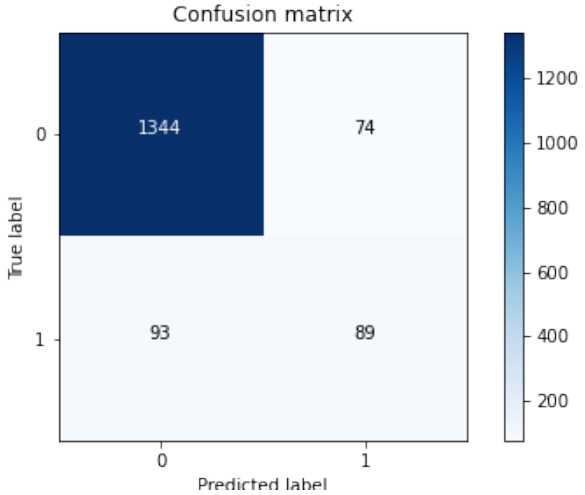

Fig. 9. Confusion Matrix with PCA+SMOTE

Support Vector machines are known to be very effetive in training high dimension data as it uses a hyperplane to separate the n-dimensional data points. It used the support vectors to construct the decision boundary for classification, hence it is both space and time efficient. Even though this model gives an accuracy of 89.5% the other metrics like precision, recall, f-1 and ROC-AUC score is a good improvement from the logisitic regression models.

SVM Metrics	
Metrics	SVM
Accuracy	0.895
Precision	0.525
Recall	0.8076
ROC-AUC score	0.8569
F-1 Score	0.6363

Naive Bayes Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. Naive Bayes usually performs well with high dimensional data due to the independence assumption of attributes, but in our model due to indirect correlation between the attributes and the fact that some attributes contributed more than the others (further proved using Random Forests) caused it to give us an accuracy of 85.4375%.

Decision Tree A decision tree is a flowchart-like tree structure where an internal node represents feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome.Decision trees are affected by outliers and high dimensional data. In order to avoid over-fitting of data, we perform PCA and extract meaningful features with the help of

dimensionality reduction. Decision tree performs better after PCA and gives an accuracy of 0.8 and F1-score of 0.88.

Ensemble Ensemble methods is a machine learning technique that combines several base models(weak learners) in order to produce one optimal predictive model. The EnsembleVoteClassifier is a meta-classifier for combining similar or conceptually different machine learning classifiers for classification via majority or plurality voting. This classifier implements "hard" and "soft" voting. In hard voting, we predict the final class label as the class label that has been predicted most frequently by the classification models. In soft voting, we predict the class labels by averaging the class-probabilities.

The ensemble model was trained on data that had feaure reduction and SMOTE performed.

We combine two weak learners - svm and naive bayes and we use a strong voting classifier and observe that the metrics are better than the individual logistic regresion,svm and naive bayes used for classification

Ensemble(SVM+GNB+DecisionTree) Metrics	
Metrics	SVM+GNB+DecisionTree
Accuracy	0.9245
Precision	0.8058
Recall	0.8740
ROC-AUCscore	0.9066
F-1 Score	0.8385

Random Forest Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned.Random decision forests correct for decision trees' habit of overfitting to their training set.Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted tree. Random forest gave an accuracy of 0.9359%.

KNN K-nearest neighbours (KNN) is one of the simplest Machine learning algorithms based on Supervised Learning Algorithm.It is a lazy learner because it stores the result and performs action on it when requires. KNN is generally used for classification problems.In unbalanced data,KNN shows biased results towards the majority class. Unbalanced class gives an score of 0.94 for class 0 and 0.18 for class 1.After balancing the data, the score for minority class significantly improved to 0.57.

Recall and F-1 score for the minority class increases for PCA+SMOTE as compared to PCA alone. The F-1 score for minority class increased from 0.46 with PCA to 0.76 with PCA+SMOTE. The F-1 score for minority class increased from 0.32 with PCA to 0.88 with PCA+SMOTE.

KNN Metrics		
Metrics	PCA	PCA+ SMOTE
Accuracy	0.91	0.88

XGBoost XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. The XGBoost algorithm performs well in machine learning because of its robust handling of a variety of data types, relationships and the variety of hyperparameters that we can fine-tune. XGBoost can be used for regression, classification, and ranking problems. XGBoost is not affected by the presence of unbalanced data. But its always better to balance data so the predictions are not biased to the minority class. For the water sample dataset, we get an accuracy of 94.14

Results and Conclusion

After comparing different models with respect to the accuracies and metrics, it can be observed that the metrics for models which were trained on PCA+SMOTE have an overall higher metric value.

Accuracies	
Model	Accuracies
Logistic Regression	0.895625
SVM	0.895
Naive Bayes	85.4375
Decision Tree	0.9525
Ensemble	0.9245
RandomForest	0.9359
KNN	0.88
XGBoost	94

Future Work

We would like to extend the problem as to find better methods to deal with the data imbalance as SMOTE still synthetically oversampling which will add to the variance. Exploring or training models that perform better in presence of class imbalance.

Acknowledgment

We would like to express out profound gratitude to Prof KS Srinivas for always encouraging us to learn more and make classes very interesting everyday. We would also like to thank the Teaching Assistants for helping us learn and apply the theoretical concepts taught in class practically and make the course even more interesting.

References

[1] Naseriparsa, Mehdi and Kashani, Mohammad Mansour Riahi "Combination of PCA with SMOTE resampling to boost the prediction rate in lung cancer dataset" , 2014.

[2] https://www.epa.gov/ccl/types-drinking-water-contaminants

[3] Y. Khan and C. S. See, "Predicting and analyzing water quality using Machine Learning: A comprehensive model," 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT), 2016, pp. 1-6, doi: 10.1109/LISAT.2016.7494106.

[4] Mozo, A., Morón-López, J., Vakaruk, S. et al. Chlorophyll soft-sensor based on machine learning models for algal bloom predictions. Sci Rep 12, 13529 (2022). https://doi.org/10.1038/s41598-022-17299-5

[5] Fitore Muharemi, Doina Logofătu, Florin Leon (2019) Machine learning approaches for anomaly detection of water quality on a real-world data set, Journal of Information and Telecommunication, 3:3, 294-307, DOI: 10.1080/24751839.2019.1565653

[6] T. Maruthi Padmaja, Bapi S. Raju, Rudra N. Hota, and P. Radha Krishna. 2014. Class imbalance and its effect on PCA preprocessing. Int. J. Knowl. Eng. Soft Data Paradigm. 4, 3 (August 2014), 272–294. https://doi.org/10.1504/IJKESDP.2014.064265