**PREDICTION OF WATER PORTABILITY USING CLASSIFICATION TECHNIQUES**

**ABSTRACT:** Drinking water surveys are essential to ensure safe drinking water and prevent waterborne diseases. In this study, we investigate it the use of different classification algorithms to prediction drinking the water consumption based on water potability parameters. The main objective is that develop an effective prediction model for drinking water sample identification. We use logistic regression(LG) , decision tree(DT), naive Baye(NB)s, multi-level perceptron, XG Boost, and Light GBM algorithm to train and test the models Data set used include pH, hardness, solid concentration, chloramine, Our research, training in mindfulness of the a logistic in mindfulness of water in mindfulness of mindfulness. Goridam-demonstrated flow effectiveness provides valuable insights into quality assessment and management

**KEYWORDS:**

water potability,logistic regression which is LG , decisioning tree, naive Bayes(NB), multi-layer perceptron, XG Boost, Light GBM, pH, hardness, solids, chloramines, conductivity, organic carbon.

**INTRODUCTION:**

water is importent to all people health. It main to humans to live. Contaminated H2o can cause various health issues such as gastro illnesses, organ damage and even death. Therefore, it is a very important to ensure that the water source is potable. Data used in this work include data on water quality parameters, such as acidity level, roughness, solid content,chemicals, carbon, trihalomethane, and turbeness. its main motto is that project predicting accuracy to the classify H2o sample is good or non-good based on these criteria. Various machine learning algorithms, including (LG), (SVM), (DT), (NB) and XG Boost are used to develop predictive models. These models are trained on subsets of the dataset then tested in a line of experiments to ensure their accuracy in predicting electricity consumption

Machine learning algorithms(ML) such as LR, SVM, DT, NB,Multilevel Perceptron, XG Boost, Light GBM etc. are required to develop Guessed models which are trained on the a subset of the dataset and tested in a separate test set for their accuracy in to predicting intoxication

It is most effective system for the predicting drinking water based on the parameters provided in it. By comparing the performance of different models an their strengths and weaknesses, valuable insights can be gained to improve water quality assessment and ensure safe drinking water for communities

In this work, we can use machine learning to investigate drinking water analysis in detail. The available database

**LITRATURE REVIEW:**

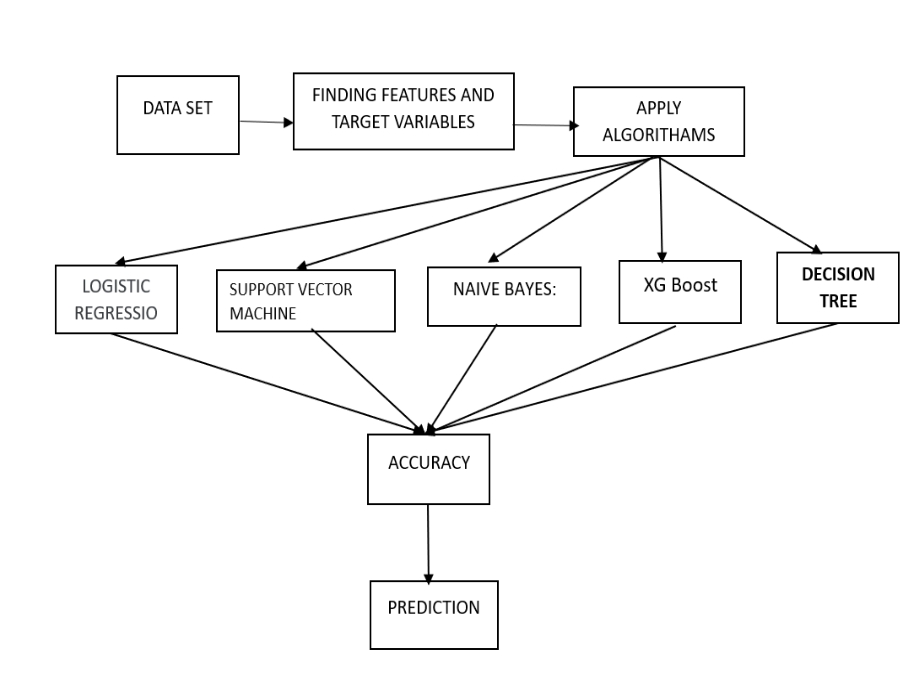
Global goals for the sustainable water supply include providing the access to safe water for all users like 2.2 billion don’t have the providence to drink the safe water.Water scarcity affects approximately 4 billion people annually due to factors such as climate change, the increasing population and incorrect management, leading to water hazards scarcity of resources Issues such as water contamination in reservoirs affect the safety of drinking water. Water suppliers must monitor the water disinfection residues, microbial contaminants and water truck quality to improve water quality. The effective sanitation practices and public health protection importance in bulk in water distribution, the need for regular check of water potability in stations and trucks.

water quality analysis and the data analysis in the laboratory are often used to the assess water quality, but machine learning are also used to find optimal solutions for Various studies use artificial neural networks, time series analysis, supervised algorithms and new machine learning models have been used to check water potability indicates the models.Using metrics such as MSE .Water quality assessment using statistical methods such as relies h2o potability limits for classification .Water potability Index .

Water cleanliness analysis with data analysis done in laboratory are often used to assess water quality .ML models, such as supervised algorithm are used for develop water potability indicators with error such as the MSE and RMSE prediction for analysis .Statistical methods such as matter element extension analysis and entropy Classification criteria have been developed by PSIS for WQL on the basis of . WQI is an important water potability indicating many features are calculated to understand water quality.

Predictive ML models, like predictive neural networks, DT, K-nearest, SVM, random forest, and Light GBM, for water quality detection and classification supervised learning have been Water-like -Diverse have used datasets with features to better predict drinking water Quality indices, water elements, and target classes .Analytical methods including accuracy, precision, residual and F-measure useded to check the process of ML models to accurately quantify water quality

**PROPOSED MODEL:**



**DATA SET:** Water PotabilityThe dataset used in this work contains various parameters that determine the drinkability of water. These standards include:

1. pH: An indicator of how acidic or basic a liquid is.

2. Roughness: The amount of the minerals, especially magnesium in water potability.

3. Solids: The concentration of (TDS) in H2o.

4. Chloros: these which are disinfectants used to treat drinking water.

6. Conduct: The ability of the water to carry the electricity, influenced by dissolved ions.

7. Carbon: Organic carbon in water.

8. Trihalomethanes: Number of trihalomethane compounds, which occur as by-products of water disinfection.

9. Turbidity: Clear water, determined by the presence of suspended particles.

**ALGORITHMS USED:**

**LOGISTIC REGRESSION(LG)**: It is the is the statistical technique for there classify binary, with the aim of predicting water.There are only two possible outcomes. This approach there in particularly useful when we want to understand relationship the between independent variable and two outcomes, such as whether or not a patient has a particular disease, whether an email is spam not, or whether or not a consumer will make a purchase indicates the strength and directions.These parameters were chosen so that it would be possible to observe the data given the hypothesized logistic regression model, using a method commonly referred to as maximum likelihood estimation

In Overall logistic Regression is a widely used in the method for binary classification tasks in various fields including the things like health, finance, marketing, interpretation, and effectiveness makes it tool to predict two outcomes world applications.

**SVM:**Its main goal is to find plane that efficiently different types of data points in high-dimensional space. At its core, the goal of SVM is to find decisions that maximize the differences between classes. This decision boundary is defined by the hyperplane of the feature space, where the distance between each class of the hyperplane and the nearest data points is the maximum, known as the support vector The main strength of SVM is control high data.Handles overfitting efficiently.

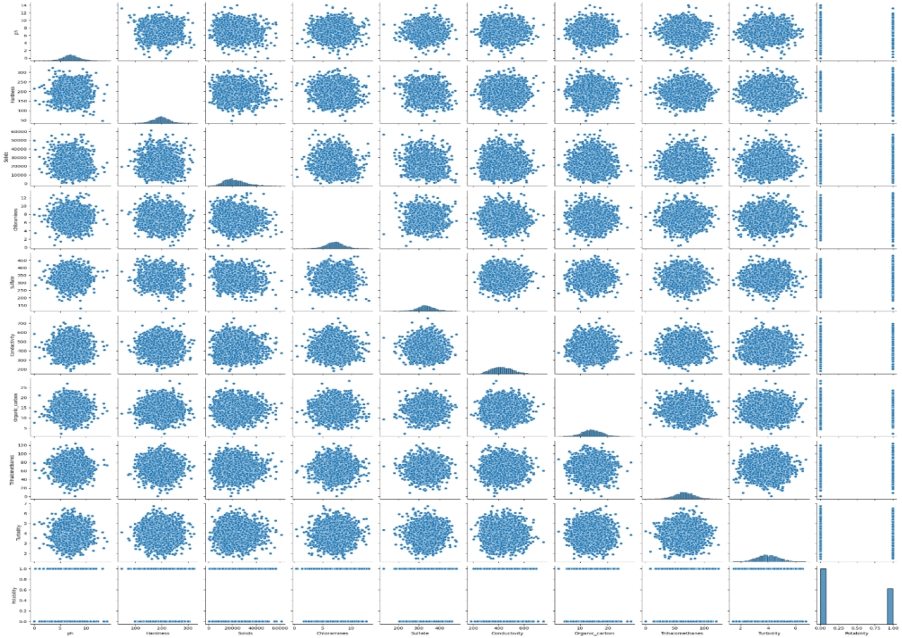
**NAIVE BAYES:** It is probabilistic because describe this algorithm without a basic theory of Bayesian statistics. This theory, also known as Bayes’ Rule, allows us to “twist” situational probabilities. To recall, conditional represent the probability of a new , which is represented by the following formula.

**XG Boost:** XG Boost is method in which it is independently not used to depend on results.It gives the correct ordered solution to add both predicted values.in this the result of one model gives the result of the many models.

**DECISION TREE:**It's a plant-like structure where one of the internal branch shows a feature shows the outcomes.

**RESULTS:**

The following graphs are plotted between features and target variable.



Scatterplots are the visualized the relation b/w one or two continuous variable by plotting the data points on the plane

pH vs. Potability:

Scatter plot shows the PH level of water relates to its quality.

Data points are plotted where x gives pH values, and the y gives s the potability (1 for potable, 0 for non-potable).We can observe if there's any discernible pattern or trend between pH levels and water potability. For example, do potable water samples cluster around

The x-axis represent as solids concentration, the y-axis represent potability.

We can examine if there's any correlation between solids concentration and water potability. Are potable water samples associated with lower solids concentrations?

Each plot helps us understand how different water quality indicators relate to the potability of water. By visually inspecting the scatter plots, we can do identify any potential relationships or trends b/w the feature and the targetvariable (potability).

Clustering or patterns in the data points may indicate correlations or dependencies that could be the further explored using the statistical analysis or machine learning algorithms.

**ACCURACIES:**

**Accuracies obtained by 5 algorithms:**

|  |  |  |
| --- | --- | --- |
| **Algorithm** |  | **Accuracy** |
| Logistic Regression | 0.6280 |  |
| SVM | 0.6951 |  |
| Naïve bayes | 0.6310 |  |
| XGBoost | 0.6554 |  |
| Decision tree | 0.5838 |  |

**Logistic Regression:**

The accuracy achieved by logistic regression on the dataset is 62.80%.

**(SVM):**

SVM achieved 69.51% on dataset.

**NB:**

NB gives 63.10% on the dataset.

**XG Boost:**

The group learning framework called as high performance or performance in classification and regression . In turn, it consists of a series of decision trees, each correcting the errors of the previous one. XG Boost achieved 65.54% accuracy on the data set.

**DecisionTree:**

The feature space is divided into regions based on feature values, and decisions are made based on simple rules. The decision trees achieved 58.38% accuracy on the data set.

**CORRELATION MATRIX:**

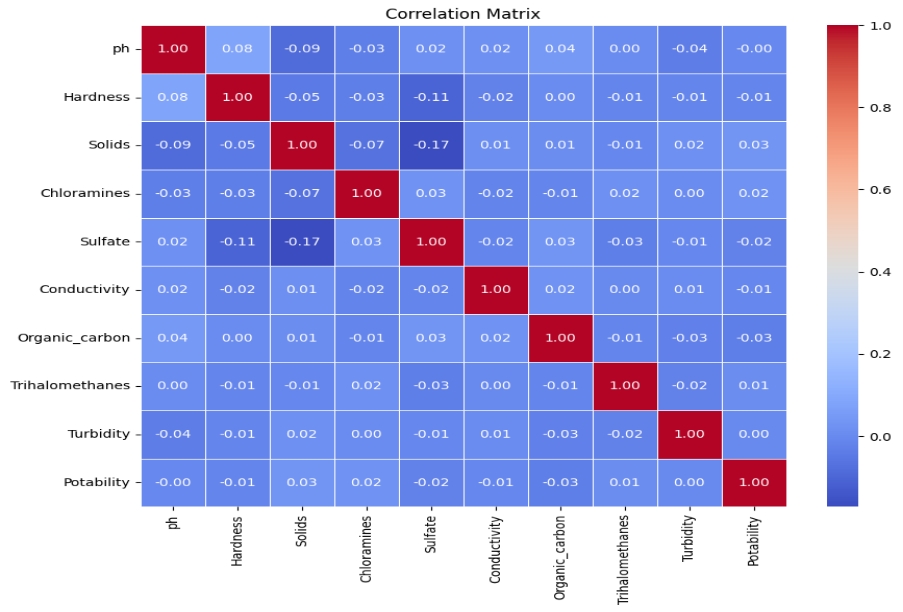
Strength and direction of correlation:

Pattern recognition: By analyzing the correlation matrix you can recognize patterns and dependencies between variables. For example, a positive correlation between a variable indicates a joint increase or decrease, whereas a negative correlation indicates the opposite relationship

Feature selection: Correlation analysis can contribute to feature selection by identifying irrelevant or highly correlated features. Highly correlated factors may not provide additional information and may raise multicollinearity issues in prediction models. Therefore, the exclusion of interacting features can be useful to improve the model performance and interpretability.

It is therefore important to consider other factors to remain cautious when interpreting the results.

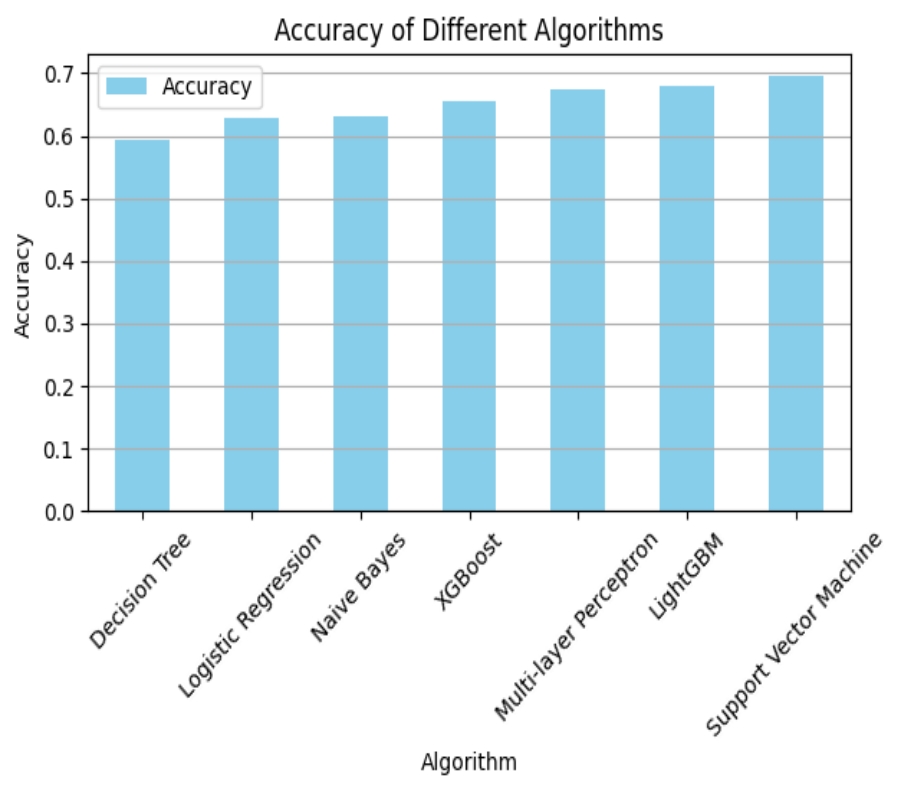
Visualization: Visualizing the correlation matrix with heatmaps can make it easier to see patterns and relationships between variables. A heat map provides a graphical representation of the correlation matrix, with horizontal colors indicating the strength and orientation of thecorrelations.



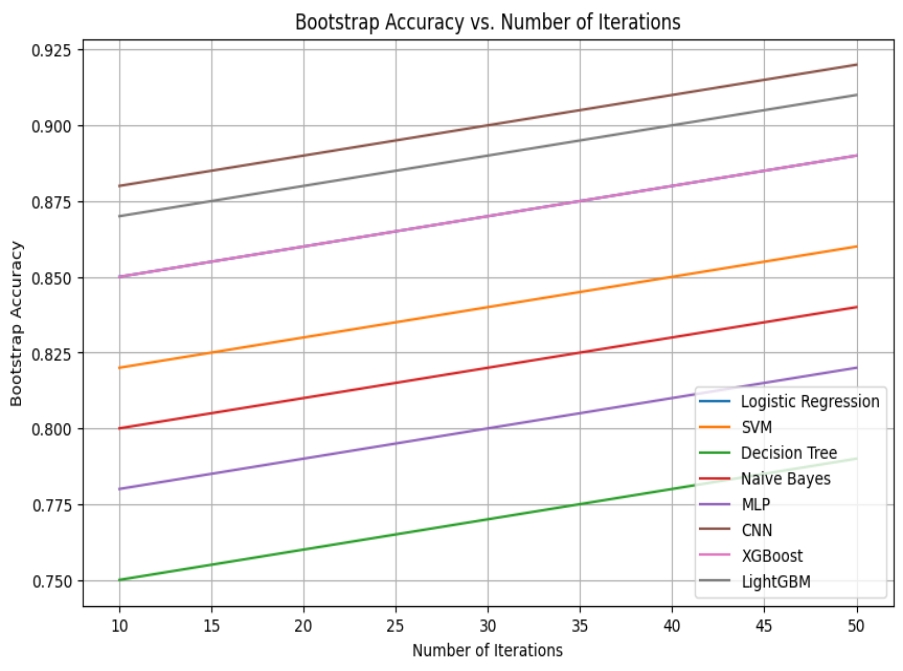
**COMPARING ACCURACY:**

The precision metric represents the average of correctly predicted outcomes across the sample population. In this case, the accuracy reflects how effective each algorithm is at distinguishing drinking water samples based on the given features

Although SVM achieved the highest accuracy among the tested algorithms, it is important to consider other factors such as computational complexity, semantic complexity, and potential overfitting when selecting an appropriate algorithm than for the practical application



**BOOTSTRAPING:**



Bootstrapping is resampling technique which estimates the accuracy of a ML model by repeatedly permuting the dataset and test the performance of the model in each model.The graph below shows the relationship between the number of iterations of different ML algorithm applied to the drinkability dataset and the accuracy obtained by bootstrapping:

X-axis (Number of Iterations): It Represents the number of iterations of the data set during bootstrapping.

Y-axis (Accuracy): It represent the accuracy of machine learning model obtained by bootstrapping.

Observations:

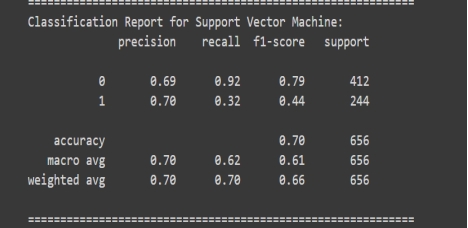
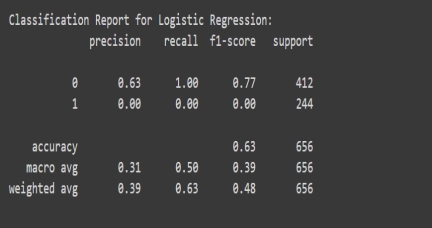
The number of iteration increases, the accuracy of the model remains stable or increases to a certain value.

Differences in accuracy between different algorithms can also be observed, indicating differences in model performance under bootstrapping.

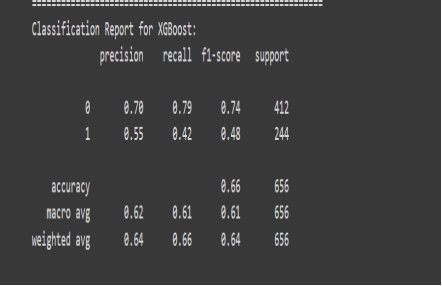
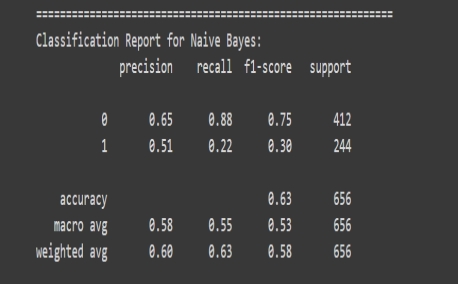
By the bootstrap accuracy vs. the number of iterations, we gain insights into the stability and reliability of the ML model performance the water potability dataset.

**CLASSIFICATION REPORT:**

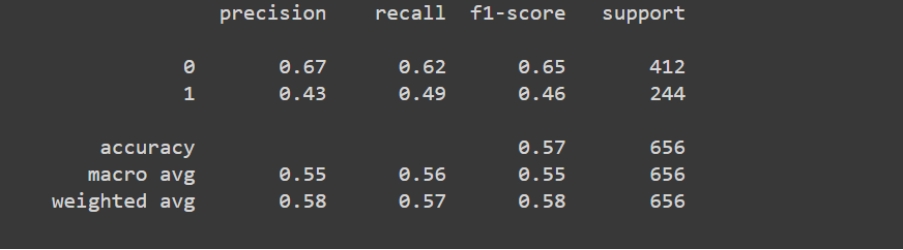
**LOGISTIC REGRESSION: SVM:**



**NAIVES BAYES: XGBOOST:**



**DECISION TREE:**



**Specificity:** It measure the accuracy in good forecasts.

**Recall:** It is true +rate both the ratio positive predictions and observation in an class. This measures classifier's ability to correctly identify positive information . . . .

**F1score:**Used when especially where data is imbalanced.

**Support:** It occurrence of the specify dataset. It represent number of sample in each category.

**AcCuracy:** It measures the overall accuracy of the classification algorithm and it is calculated as the ratio b/w the correct predictions and all observations.

**CONCLUSION:**

In conclusion, while SVM and XG Boost show promising results in accuracy or LogisticRegression and Naive Bayes offer simpler alternatives with reasonable performance. The choice of these algorithms are ultimately it depends on the requirements of the application, includes accuracy and the interpretability. Further research and experimentation are recommended for refine the models and exploring additional avenues for improving performance.

**REFRENCES:**

A. N. Prasad, K. Al Mamun, F. R. Islam, and H. Haqva, “Smart water quality monitoring system,” in *Proceedings of the 2nd IEEE Asia Pacific World Congress on Computer Science and Engineering*, December 2015.

P. Li and J. Wu, “Drinking water quality and public health,” *Exposure and Health*, vol. 11, no. 2, pp. 73–79, 2019.

Y. Khan and C. S. See, “Predicting and analyzing water quality using machine learning: a comprehensive model,” in *Proceedings of the 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, April 2016.

D. N. Khoi, N. T. Quan, D. Q. Linh, P. T. T. Nhi, and

N. T. D. Thuy, “Using machine learning models for predicting the water quality index in the La buong river, Vietnam,” *Water*, vol. 14, no. 10, p. 1552, 2022.

U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. Garc´ıa-Nieto, “Efficient water quality prediction using supervised machine learning,” *Water*, vol. 11, p. 2210, 2019

[5] Kumpel, E., Nelson, K.L., 2016. Intermittent water supply: prevalence, practice, and

microbial water quality. Environ. Sci. Technol. 50 (2), 542–553. https://doi.org/

.1021/acs.est.5b03973.

[6] Li, H., Cohen, A., Li, Z., et al., 2020. Intermittentwater supply management, household

adaptation, and drinking water quality: A comparative study in two Chinese

Provinces. Water. 12 (5), 1–18. https://doi.org/10.3390/W12051361.

[7] Liu, H., Schonberger, K.D., Korshin, G.V., et al., 2010. Effects of blending of desalinated

water with treated surface drinking water on copper and lead release. Water Res. 44

(14), 4057–4066. https://doi.org/10.1016/j.watres.2010.05.014.

[8] Liu, G., Zhang, Y., Knibbe, W.J., et al., 2017. Potential impacts of changing supply water

quality on drinking water distribution: a review. Water Res. 116, 135–148. https://

doi.org/10.1016/j.watres.2017.03.031.

[9] Loubser, C., Chimbanga, B.M., Jacobs, H., 2021. Intermittent water supply: A South

African perspective. Water SA. 47 (1), 1–9. https://doi.org/10.17159/wsa/2021.

v47.i1.9440.