Applying Machine Learning to Predict Water Quality

Final Project Introduction Machine Learning

Muhammad Ramadhani - Batch 9

Ramshall13@gmail.com

1. Introduction

Water scarcity is a social and environmental problem because water is a basic human need that is very important in our life because we need it for food, sanitation, health and energy. The COVID19 pandemic has shown the essentiality of sanitation to prevent disease. It’s clear that the impact of COVID19 is more risky for the poor who live in slums and don't have access to clean water for consumption. About 3 billion people worldwide lack basic sanitation facilities at home. Two of three of our earth is covered by water, but the fact is there is only a few amount of water that is good for us to consume. From the total amount of water, about 97% is salt water that can’t be consumed and the rest about 3% is fresh water. Two of the 3% of fresh water is in the form of frozen water that exists in the north pole and south pole and only 0.62% of clean water is able for us to consume from the remaining 1% of fresh water. The supply of clean water will continue to shrink due to the interference of water circulation for city developments that are made of concrete and asphalt which prevent raindrop from being absorbed by the soil. Therefore, water scarcity is predicted to displace about 700 million people by 2030.

According to the United Nations (UN) in 2019, 2.2 billion people or a quarter of the world's population still lack safe drinking water. Meanwhile, 4.2 billion people didn’t have safe sanitation services and 3 billion didn’t have basic hand washing facilities. From the Bappenas report, the availability of water in most parts of the islands of Java and Bali is currently classified as scarce to critical. Meanwhile, the availability of water in South Sumatra, West Nusa Tenggara and South Sulawesi is predicted to become scarce or critical in 2045. The scarcity of clean water also applies to drinking water. According to the 2020-2024 RPJMN, only 6.87% of households have access to safe drinking water. The 2020 National Socio-Economic Survey (Susenas) from BPS also shows that 90.21% of households have access to proper drinking water, even though the distribution is uneven.

Starting on the problem of the clean water crisis that is still occurring, the objective problem of this analysis is to ensure that the water is suitable for drinking or not. Based on historical data, the purpose of this analysis is to predict whether the water contained is fit for drinking or not.

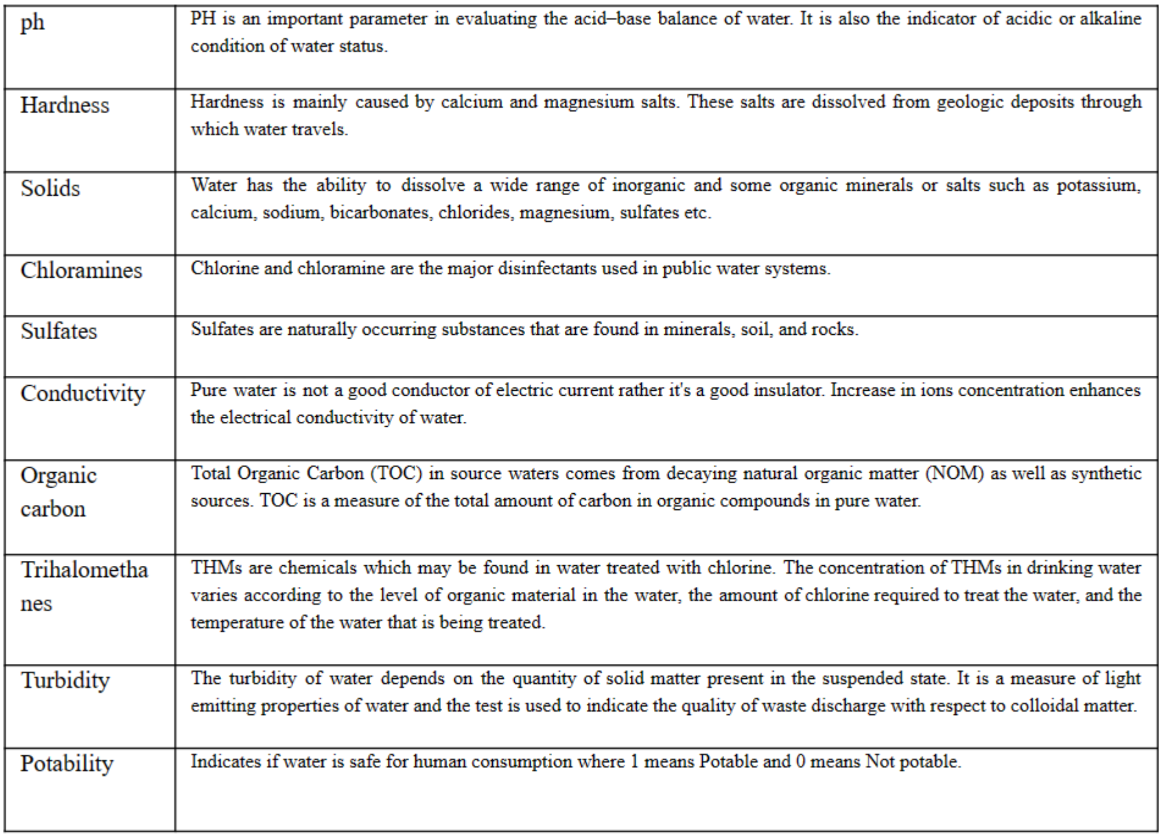
2. Dataset and Features

Context : Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions.

Features :

The water potability dataset contains water quality metrics for 3276 different water bodies, that also includes 10 variable below :

Table 1. Information of Features Dataset



ph	PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status.
Hardness	Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels.
Solids	Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc.
Chloramines	Chlorine and chloramine are the major disinfectants used in public water systems.
Sulfates	Sulfates are naturally occurring substances that are found in minerals, soil, and rocks.
Conductivity	Pure water is not a good conductor of electric current rather it's a good insulator. Increase in ions concentration enhances the electrical conductivity of water.
Organic carbon	Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water.
Trihalomethanes	THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated.
Turbidity	The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter.
Potability	Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

3. Methods

a. Clustering

In the Clustering part of the research, there are usually two methods we can use which are K-means and Hierarchical Clustering. k-means is method of cluster analysis using a pre-specified no. of clusters. It requires advance knowledge of ‘K’. Hierarchical clustering is also a method of cluster analysis which seeks to build a hierarchy of clusters without having a fixed number of clusters. but after finding the value of K and entering it into K-means Clustering, this clustering tends to be difficult to read and decipher because of the large amount of data we use in this study. Although hierarchical clustering is actually better used with less data, because it is easy to use and more applicable, this is the reason this research only uses hierarchical clustering.

All objects begin as singletons or individual clusters in hierarchical clustering. They are then consolidated utilizing one of the accompanying linkage techniques. Calculating the distances or similarities between all objects is how the linkage methods function. The remaining clusters are then reduced by combining the closest pair of clusters into a single cluster. There are couple of linkage method which are Average, Single, Complete, and Ward.

For the average linkage method, the average pair-wise proximity of all pairs of objects in various clusters. Based on their shortest average distances, clusters are combined. In the Single linkage two clusters with the shortest minimum distance are combined. This interaction rehashes until there is just a solitary group left. For the Complete linkage, two groups with the nearest greatest distance are blended. This procedure continues until there is only one cluster remaining. Furthermore, for Ward's linkage, the error sum of square (ESS) values of two clusters are used to merge them. The two groups with the most minimal ESS are consolidated. This procedure continues until there is only one cluster remaining.

b. Classification

Classification is a form of supervised learning in which the objectives are also served by the input data. Data that is either structured or unstructured can be classified. The primary objective of a classification problem is to determine the category or class that a new set of data will belong to. Credit approval, medical diagnosis, and targeted marketing are just a few of the many applications for classification. There are a couple of different classifiers which are Logistic regression, KNN, SVM, and Decision Tree and Random Forest.

Logistic Regression is the most widely used algorithm in Classification. It is a way to predict a categorical dependent variable from a set of independent factors. This objective—classification—inspired the development of logistic regression, which is particularly useful for determining how numerous independent factors influence a single outcome variable. The algorithm only works if the predicted variable is binary, all predictors must be independent of one another, and data must not contain missing values.

K nearest neighbors (KNN) is a straightforward method that uses a similarity metric to classify new examples while keeping all existing examples. It's a type of languid learning since it doesn't attempt to construct a nonexclusive inner model; instead, it only stores training data instances. Each point's k closest neighbors vote by simple majority to determine the classification. A case is assigned to the class with the most members among its K closest neighbors, as determined by a distance function, by a majority vote of its neighbors. The case is simply assigned to the class of the nearest neighbor if K = 1.

A common Supervised Learning technique, the Support Vector Machine, or SVM, can be used to solve problems with classification and regression. However, the majority of its use is in Machine Learning for problems with classification. The goal of the SVM algorithm is to find the best line or decision boundary for classifying n-dimensional space into classes so that in the future, additional data points can be easily assigned to the right category. A three-dimensional classification model that goes beyond the X/Y predictive axes is created when SVM techniques classify data and train models within extremely limited degrees of polarity. SVM is used to select the extreme points and vectors that help create the hyperplane.

In machine learning, classification and regression are carried out with the help of random forest, a supervised learning technique. It is a classifier that, in order to improve the projected accuracy of a dataset, averages the outcomes of numerous decision trees applied to distinct subsets of the dataset. Because it uses the average to improve the forecast accuracy of the model and prevent over-fitting, it is also referred to as a meta-estimator. It fits a number of decision trees to various sub-samples of the dataset. From a collection of decision trees that are frequently trained by "bagging," it creates a "forest." The bagging method is based on the idea that combining multiple learning models improves the end result. The random forest uses forecasts from each tree rather than a single decision tree to predict the final result based on the majority of votes.

4. Data Exploration and Processing

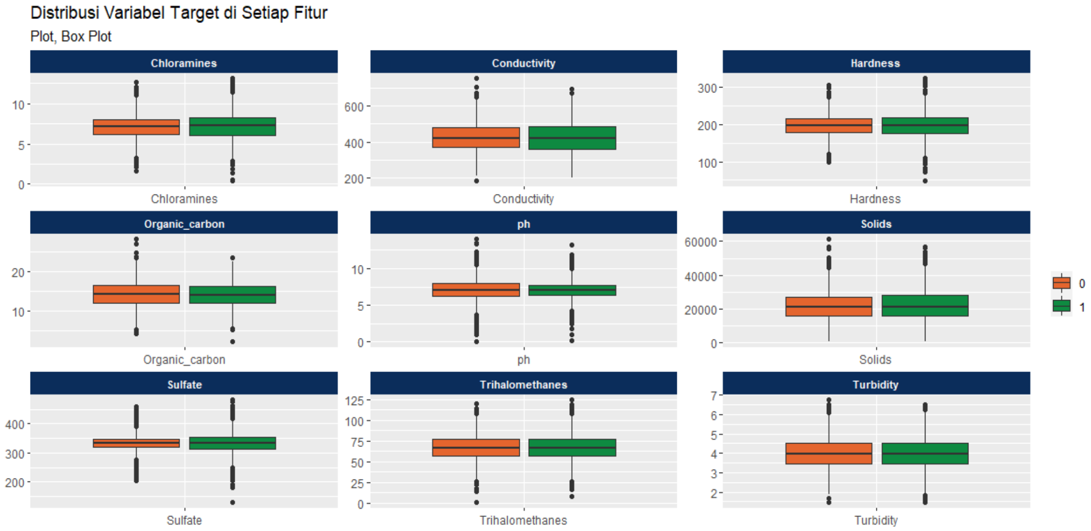
a. Cleansing & EDA

With so many features and so much data available, it is important to make sure that there are no missing values, mismatched data types, and that it helps to see patterns that facilitate analysis or early conclusions. As explained in the dataset and features section, the Potability column is factorized data. With the help of bar chart visualization, here is the factor composition of the Potability column.

Water with undrinkable quality (Potability=0) has more data than water that is suitable for drinking (Potability=1), where 61% of the data contains non-potable water data (see Appendix 1). There are three columns/features that contain missing values, namely Sulfate, ph, and Trihalomethanes. Three features (columns in the dataset) out of all those in the dataset have a total missing value percentage that is below 50% (see appendix 2). Therefore, we decided not to delete them, but to fill them with median values. There is no specific reason for this decision, except to avoid the influence of the magnitude of outlier values that can affect the distribution of the data. Another step we took was to change the data type according to the nature of each feature.

After ensuring that all data is complete, clean, and appropriate, the next step is to look at the distribution of data with the help of boxplot visualization.

Figure 1. Data Distribution with Boxplot

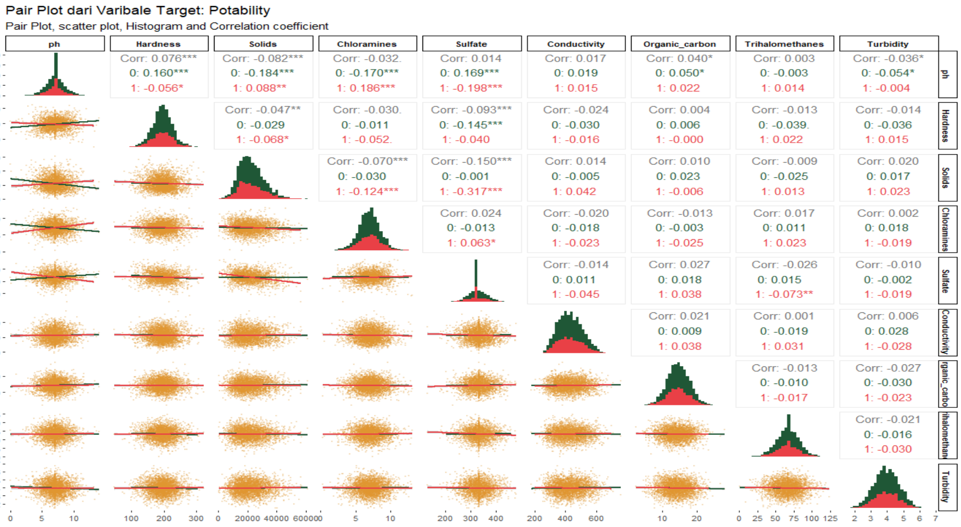


It can be seen that whether the water quality mentioned is potable or not, it has a relatively similar distribution. The most highlightable thing in this graph is the outliers, where almost all features have them. To maximize the modeling process, the step taken is to standardize the values on each feature with the scale function.

b. Correlation

An equally important data exploration is to see how strong the relationship between each feature is. Although correlation is not causality, relationships/patterns that "happen" to be the same can indicate their influence on the model to be used.

Figure 2. Correlation between Features in Data



It can be seen that there is no significant or strong correlation across all features. The negative or positive sign is a description of the relationship that occurs in the feature.

c. Cluster

The purpose of clustering is to group data that has similar patterns in its values. Basically, this process is done for unclassified data. Even though there is already a classification in the data, there is no harm in doing this step. The reason is that clustering can simultaneously validate the classification that is already available in the dataset, in addition to meeting the assessment criteria. With the Hierarchical Clustering model, the next step is to determine the best method to use in finding clusters. Four methods were tested, namely Average, Single, Complete, and Ward. The result:

Table 2. Result Method Test

Methods	Result
Average	0.7894
Single	0.6941
Complete	0.8829
Ward	0.9685

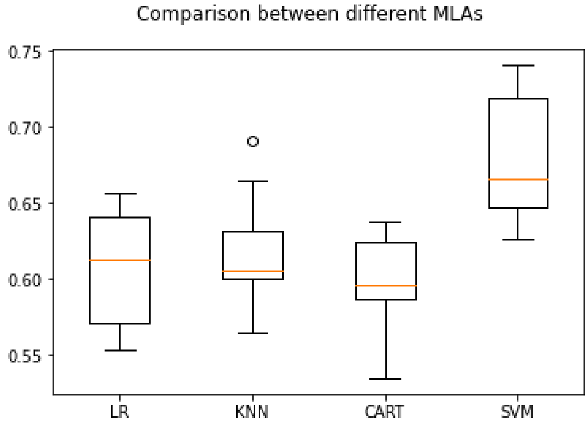
The result that is close to one will be selected as the method used in the three processes of finding the best number of clusters (k optimal). The three processes conducted, such as the Elbow Method, Gap Statistic, and Silhouette Method resulted in the conclusion that the best number of clusters is two clusters. For Elbow Method and Gap Statistic, the conclusion is based on the most significant decrease while Silhouette Method is seen from the most significant increase (see appendix 3). In other words, the number of classifications of two listed in the dataset corresponds to the search for the best cluster based on the machine learning process of the values in the dataset.

5. Experiment, Result, and Discussion

a. Experiment and Result

Before applying the methods used in classification, the dataset is first splitted with 80% of the data to be trained and 20% for the test data. The four models are also cross validated with k and a random state of 10 (k=10; random state=10).

Figure 3. Boxplot Scoring of Each Models



On average in cross validation, the best accuracy is owned by Support Vector Machine (SVM). Empat model yang telah dilakukan cross validation tersebut diaplikasikan kembali dengan turunan-turunannya.

Table 3. Evaluation of Each Model Tested

Model used	Train Accuracy	Test Accuracy	Precision	Recall	AUC
SVC	0.7359	0.6860	0.7021	0.2705	0.6013

RFClassifier	1.0000	0.6677	0.5985	0.3238	0.5976
NuSVC	0.8645	0.6616	0.5632	0.4016	0.6086
LRCV	0.6053	0.6280	0.0000	0.0000	0.5000
LinearSVC	0.6053	0.6280	0.0000	0.0000	0.5000
KNNClassifier	0.7653	0.6311	0.5056	0.3730	0.5785
DecisionTreeClassifier	1.0000	0.5762	0.4346	0.4631	0.5532

For each evaluation value of the models used, we also perform an ROC comparison that explains the results of all possible True Positive (TP) and False Positive (FP) values contained in the data. The model is said to be good if it has a higher TP value and a lower FP value. Hasilnya model Support Vector Classifier terlihat lebih bagus dibandingkan model lainnya (see appendix 4).

b. Discussion (Future Engineering)

As explained earlier from model testing, the best model results are owned by the Support Vector Classifier. This model looks better in all aspects, such as Accuracy, Precision, AUC, and ROC. If you want to see the model predict actual data from the learning process of data from the total population, the accuracy score can be a reference to be applied.

However, overall the tested model has a tendency of overfitting, where one way to see it is from the accuracy train value which is higher than the accuracy test value. Therefore, it is important to do several things to avoid this, such as training the model by adding more data, performing cross validation, selecting the features needed, regularization, and adding basic knowledge according to the problem to be studied. The lack of knowledge that we have, of course, this final project certainly has shortcomings, one of which is overfitting the model mentioned earlier. In addition, further testing can be done by increasing the accuracy of each model or increasing other scores that can maximize the model to be said to be good. Overcoming class imbalance in the data and hyperparameter search are two of the many things that can improve the performance of the selected model.

6. Conclusion

Back to the analysis objective that we set, that the purpose of this test is to find the best model that can ensure the water is suitable for drinking or not. Based on the test results, we found two models that can be chosen by companies or countries that want to make sure the water is suitable for drinking or not. Support Vector Classifier can be the model used with an alternative model that can also be used is Random Forest.

This test will be useful if we refer to the fact that the clean water crisis is still happening in some countries. Water, especially clean water, is one of the important components in life, even overcoming various diseases. Therefore, making sure it is feasible or not needs to be done. This prediction can ensure that an area has water content that can be said to be clean until it is safe to drink, which is an important point in overcoming the crisis. The results of this machine learning can be accessed by end-users from various stakeholders as long as they know the content of the water tested and then matched with the features learned by the machine (model). However, it is necessary to improve the model to make the predicted data more accurate, which will also affect the prediction of water conditions.

Link code: Ramshall/IntroML_Water_Potability (github.com)

Link youtube: https://youtu.be/-MLI0LrugJY

References

Bagaimana Cara Meningkatkan Akurasi Model Regresi Logistik Di Scikit Python? (2022). Fmihm. Retrieved December 20, 2022, from https://id.fmihm.org/373992-how-to-increase-the-model-KJHOWE

Datasans. (2019, March 17). *Memahami ROC dan AUC*. Medium. Retrieved December 20, 2022, from https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397

Dutta, B. (2022, February 1). *6 Types of Classifiers in Machine Learning*. Analytics Steps. Retrieved December 20, 2022, from https://www.analyticssteps.com/blogs/types-classifiers-machine-learning

Frost, J. (2022). *Overfitting Regression Models: Problems, Detection, and Avoidance - Statistics By Jim*. Statistics by Jim. Retrieved December 20, 2022, from https://statisticsbyjim.com/regression/overfitting-regression-models/

GeeksforGeeks. (n.d.). *Difference between K means and Hierarchical Clustering*. GeeksforGeeks. Retrieved December 20, 2022, from https://www-geeksforgeeks-org.cdn.ampproject.org/v/s/www.geeksforgeeks.org/difference-between-k-means-and-hierarchical-clustering/amp/?amp_gsa=1&amp_js_v=a9&usqp=mq331AQKKAFQArABIIACAw%3D%3D#amp_ct=1671688608444&amp_tf=From%20%251%24s&aoh=16716886035694&re

Gupta, S. (2020, November 30). *Classification Models in Machine Learning | Classification Models*. Analytics Vidhya. Retrieved December 20, 2022, from https://www.analyticsvidhya.com/blog/2020/11/popular-classification-models-for-machine-learning/

Itah, A. Y., & Akpan, C. E. (2005). Potability of drinking water in an oil impacted community in southern Nigeria.

K, D. K. (2020, May 31). *The problem of Overfitting in Regression and how to avoid it?* DataDrivenInvestor. Retrieved December 20, 2022, from https://medium.datadriveninvestor.com/the-problem-of-overfitting-in-regression-and-how-to-avoid-it-dac4d49d836f

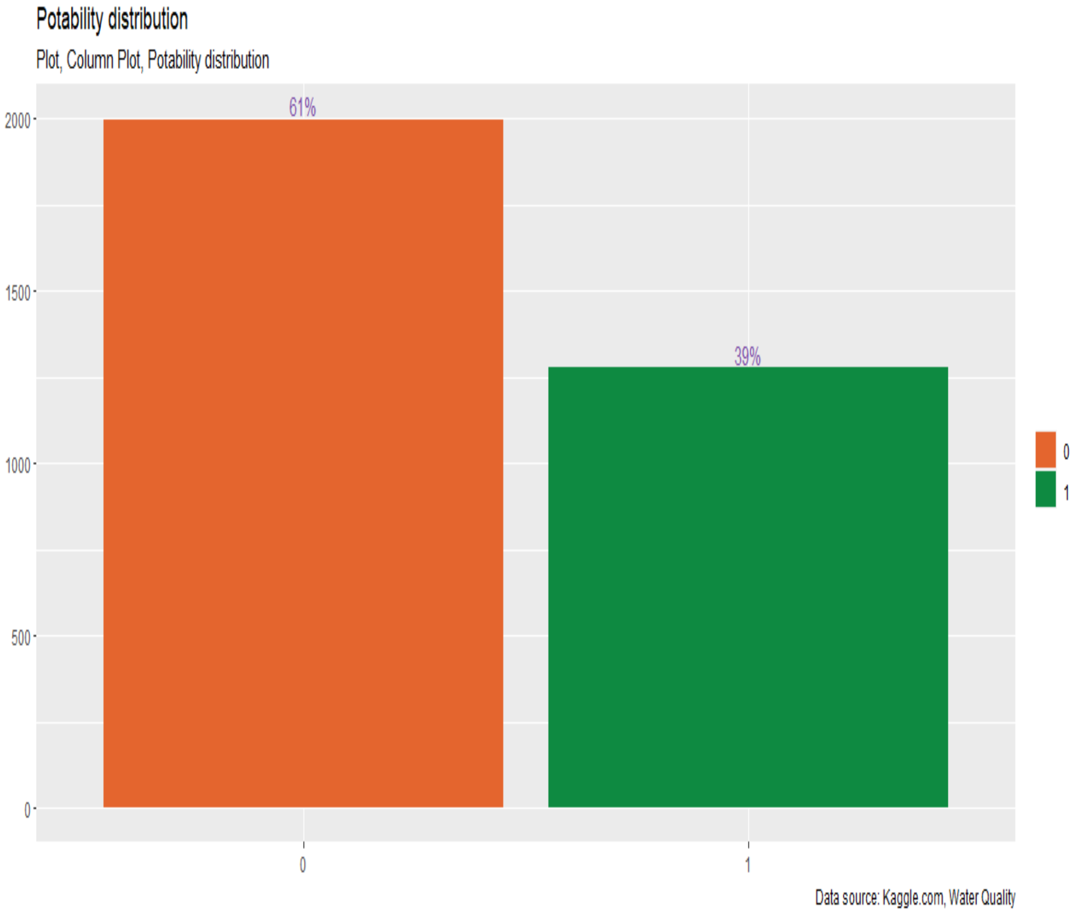
Kohli, S. (2019, November 19). *Understanding a Classification Report For Your Machine Learning Model*. Medium. Retrieved December 22, 2022, from https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397

Pai, P. (2021, May 7). *Hierarchical clustering explained | by Prasad Pai*. Towards Data Science. Retrieved December 3, 2022, from https://towardsdatascience.com/hierarchical-clustering-explained-e59b13846da8

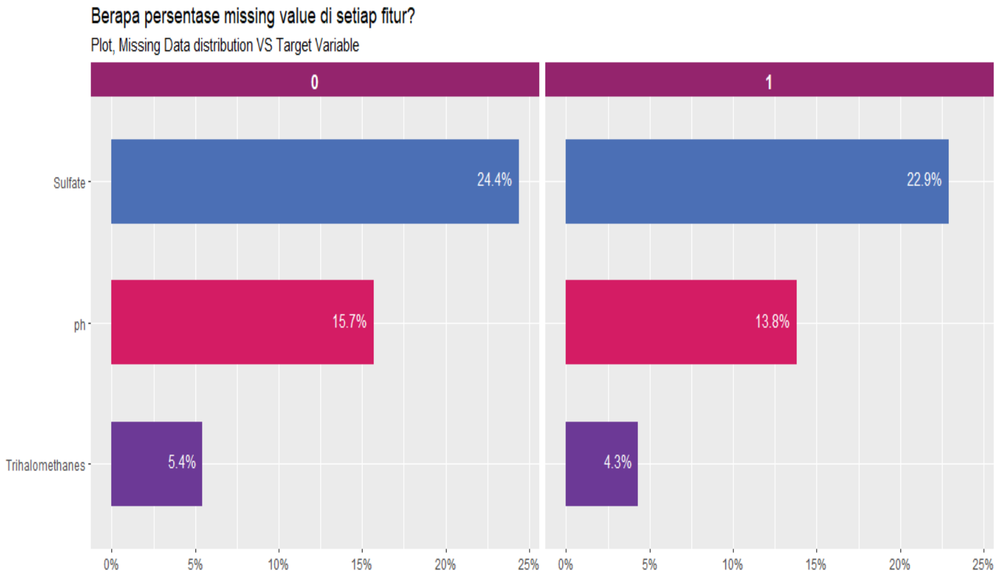
Shekhar, A. (2018, Februari 15). *What Is Feature Engineering for Machine Learning?* Medium. https://medium.com/mindorks/what-is-feature-engineering-for-machine-learning-d8ba3158d97a

Appendix

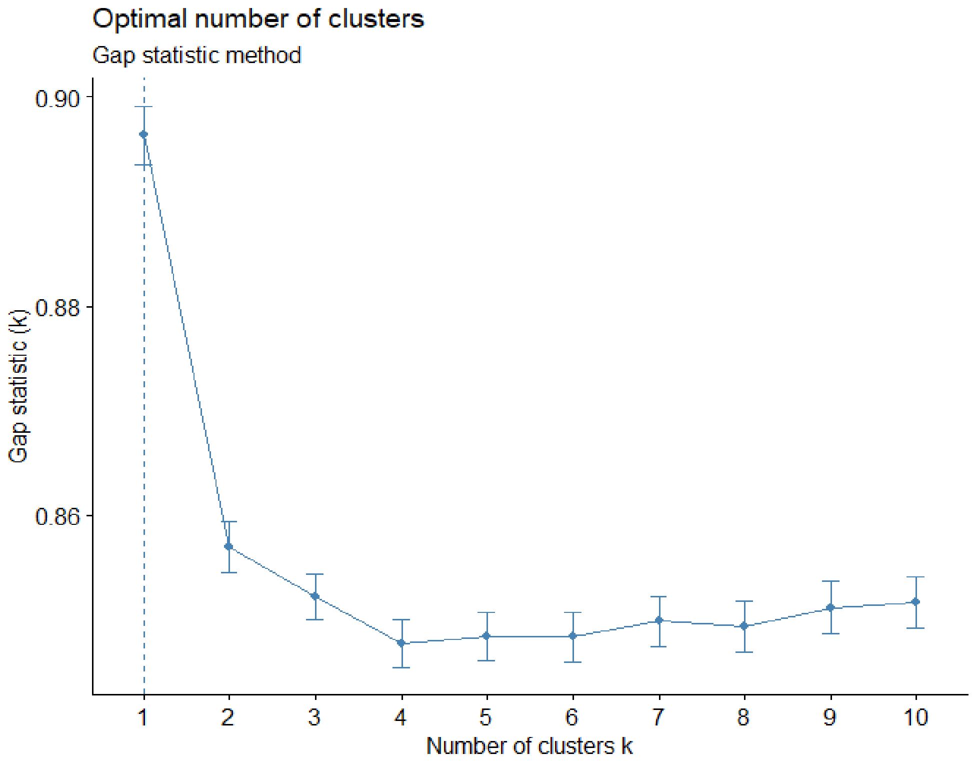
1. Appendix 1

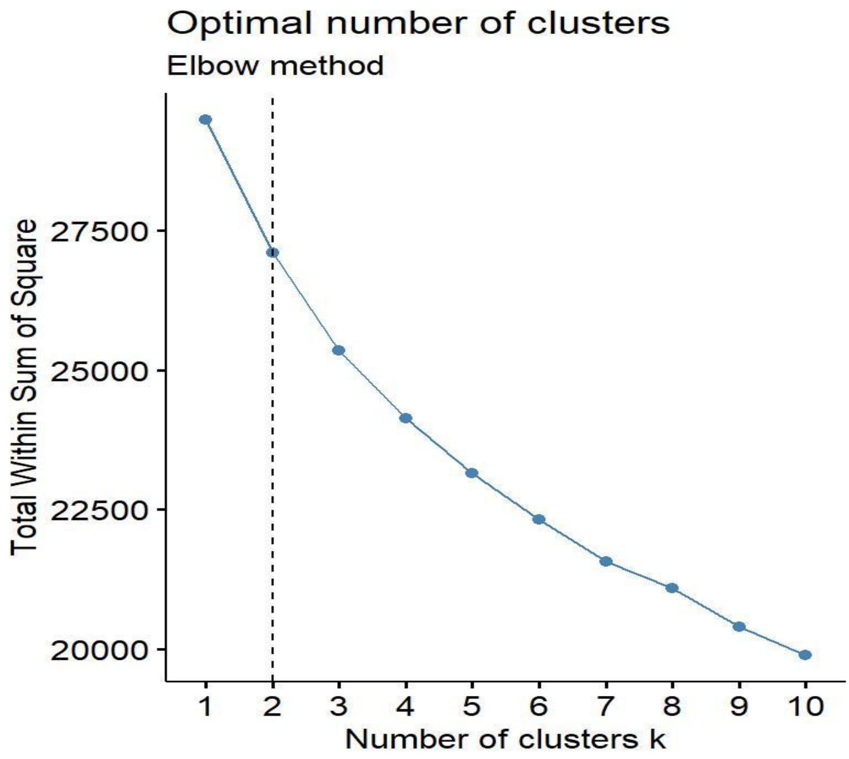


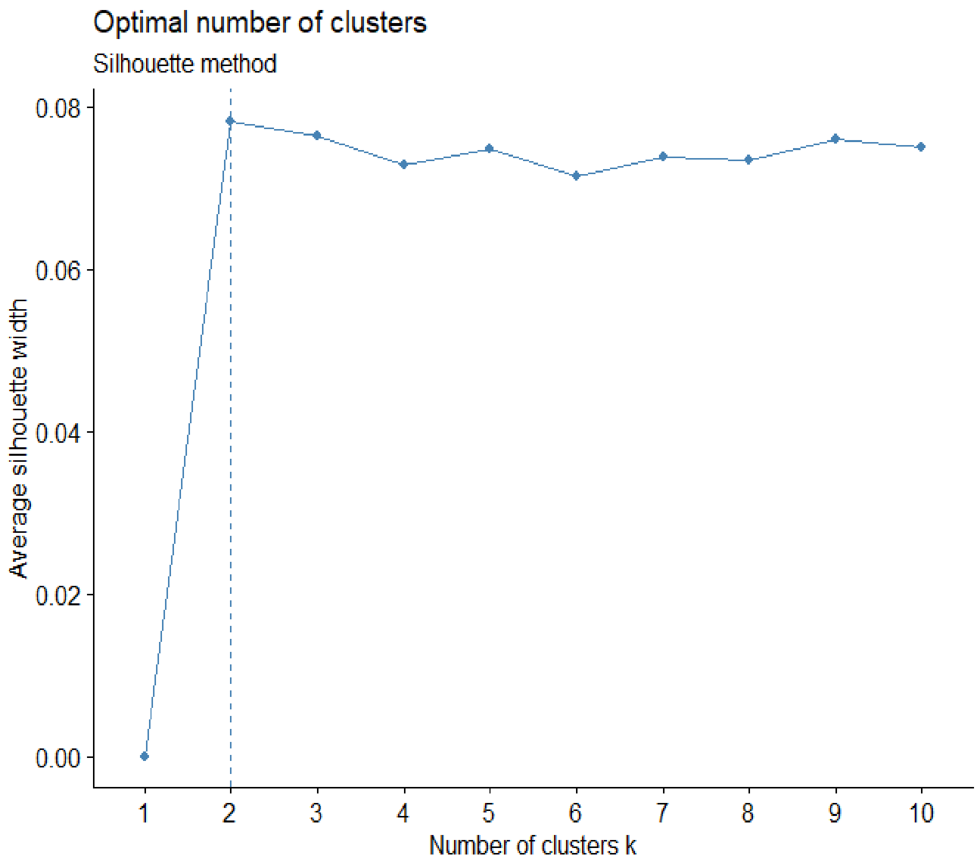
2. Appendix 2


Berapa persentase missing value di setiap fitur?
Plot, Missing Data distribution VS Target Variable
0
1
Sulfate
ph
Trihalomethanes
24.4%
15.7%
5.4%
22.9%
13.8%
4.3%
0%
5%
10%
15%
20%
25%

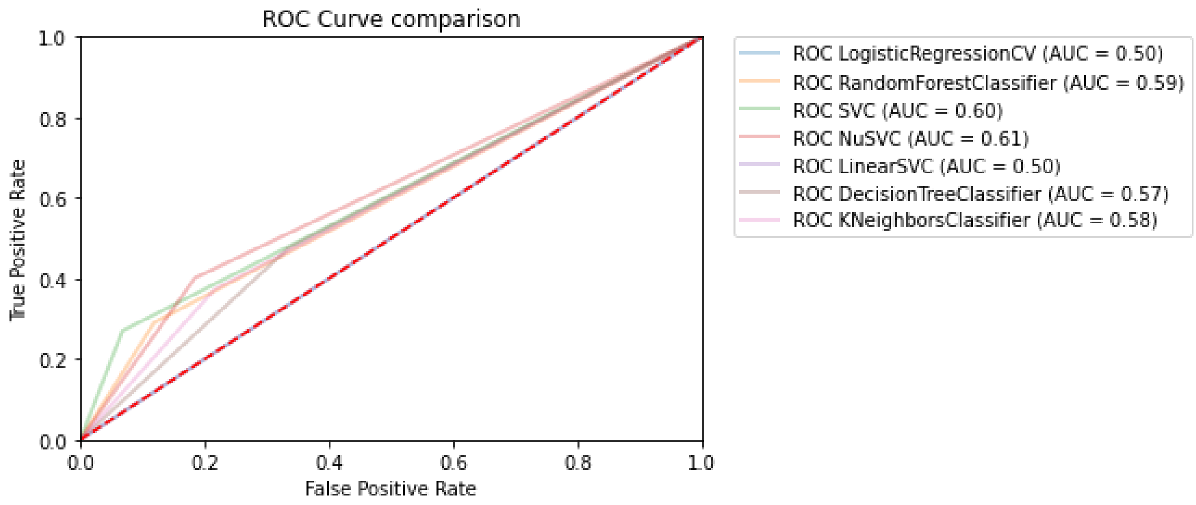
3. Appendix 3


Optimal number of clusters
Gap statistic method
Gap statistic (k)
0.90
0.88
0.86
1
2
3
4
5
6
7
8
9
10
Number of clusters k





4. Appendix f4



Link script

Classification:
https://colab.research.google.com/drive/12LSMm5CZ0I-6ZKmQGX-giavFq0_7bhwe?usp=sharing

Cluster:
https://drive.google.com/file/d/1EOCo_AGd6H-GaLlyjwucvbei-EKZU_M3/view?usp=sharing

EDA:
https://drive.google.com/file/d/1HlBzfZuJVt6iXFyrRaHT8LeBgryylCg8/view?usp=sharing