Classification of Water Quality using Logistic Regression and K-Nearest Neighbors

Darren Kang Wan Chee
Department of Statistics
Institut Teknologi Sepuluh Nopember
e-mail: 5003201184@student.its.ac.id

***Abstrak*—Water is an important aspect in our daily life. Regardless, we need to ensure that the water that we use or consume is safe. In this paper we aim to predict the quality of water by using classification methods such as Logistic Regression and K-Nearest Neighbors. The two methods are compared through a series of classification evaluation metrics. The result is that K-Nearest Neighbors perform far better all-around in classification of water quality than Logistic Regression.**

***Keywords*—Classification, Logistic Regression, K-Nearest Neighbors**

I. INTRODUCTION

WATER is very important for us humans as we could not live without water. However, not every water is safe for people to use or consume. Growing concerns over pollution and environmental degradation have a large impact on water quality. Hence, water quality assessment becomes necessary in safeguarding the health of ecosystems and ensuring the availability of safe drinking water.

This journal aims to conduct a classification of water quality focusing on the application of two widely used methods: Logistic Regression (LR) and K-Nearest Neighbors (KNN). Logistic Regression is a robust statistical method and particularly suited for binary classification problems, making it an ideal candidate for discerning between safe drinking water and non-safe drinking water. On the other hand, K-Nearest Neighbors is a non-parametric algorithm which has gained popularity for its simplicity and versatility in handling multi-class classification tasks.

The primary objective of this research is to compare the Logistic Regression and K-Nearest Neighbors method in classifying water quality based on parameters such as chemical composition and microbiological content. We aim to train and evaluate these algorithms to discern patterns and relationships within data and enhance our ability to predict water quality classes more accurately.

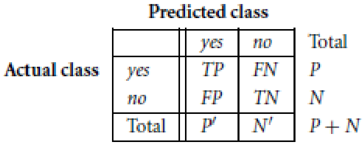
II. LITERATURE REVIEW

A. *Classification*

Classification is a form of data analysis that extracts models describing important data classes. Classification is a two-step process, consisting of a learning step, where a classification model is constructed and classification step, where the model is used to predicting the class.[1]

B. *Confusion Matrix*

Confusion matrix is a useful tool for analyzing how well the classifier can regonize tuple of different classes. Confusion matrix is usually in a form of a table of at least size *m* by *m*. Where *m* is the number of given class. [1]



Predicted class

		yes	*no*	Total
Actual class	*yes*	*TP*	*FN*	*P*
	no	*FP*	*TN*	*N*
	Total	*P'*	*N'*	*P + N*

Figure II.1 General Form of Confusion Matrix

Where *TP* is True Positive, referring to the positive tuples that were correctly labeled by the classifier. *TN* is True Negative, referring to the negative tuples that were correctly labeled by the classifier. *FP* is False Positive, referring to negative tuples that were incorrectly labeled as positive. *FN* is False Negative, referring to positive tuples that were mislabeled as negative.[1]

C. *Accuracy*

Accuracy is one of the evaluating metrics for classifier. Accuracy is the percentage of test set tuples that are correctly classified by the classifier.[1]

$$accuracy = \frac{TP+TN}{P+N} \tag{1}$$

D. *Precision and Recall*

Precision can be thought of a measure of exactness, what percentage of tuples labeled as positive are actually such. Recall is a measure of completeness, what percentage of positive tuples are labeled as such. [1]

$$precision = \frac{TP}{TP+FP} \tag{2}$$

$$recall = \frac{TP}{TP+FN} = \frac{TP}{P} \tag{3}$$

E. *F_1 Score*

F score is the harmonic mean of precision and recall. It gives equal weight to precison and recall. F score is also an alternative to use precisoni and recall as F score combine them into a single measure.

$$F = \frac{2 \times precision \times recall}{precision + recall} \qquad (4)$$

F. K-Nearest Neighbors

K-Nearest Neighbors is a classification algorithm that uses the data directly for classification without building a model. Hence, no details of model construction need to be considered and the only adjustable parameter in the model is k, the number of nearest neighbors to include in the estimate of class.

The value of $P(y|x)$ is calculated simple as the ratio of members of class y among the k nearest neighbors of x. By varying k, the model can be made flexible.

The advantage of K-NN is that the neighbors can provide an explanation for the classification result. The drawback of K-NN is in the calculation of the case neighborhood. It is not clear how to, other than by trial and error, define a metric in such a way that the relative importance of data components is reflected in the metric. [2]

G. Logistic Regression

Logistic regression consists of a model characterized by a dependent variable, binary, or polytomous, nominal, or ordinal. The general form of logistic regression can be expressed by equation (5).

$$P(Y) = \frac{e^{\beta_0 + \beta_i x_i}}{1 + e^{\beta_0 + \beta_i x_i}} \qquad (5)$$

Where $P(Y)$ is the probability of success when the predictive variable is x; x_i is the set of independent variables; β_i is the set of coefficients of the independent variables; $i = 1,2, \dots, n$ is the number of observations.

Logistic regression is often used to obtain the odds ratio when there is more than 1 independent variable, which indicates the impact of each variable on the odds ratios, the coefficient of independent variable represents the estimated increase in log odds of the response variable per unit increase in the value of the independent variable. [3]

III. RESEARCH METHODOLO.GY

A. Data Source

The data used for this analysis is a dataset about water quality which is obtained from Kaggle website.

The variables of the dataset are:

Table III.1 Variables of Dataset

Aluminum	Copper	Perchlorate
Ammonia	Fluoride	Radium
Arsenic	Bacteria	Selenium
Barium	Viruses	Silver
Cadmium	Lead	Uranium
Chloramine	Nitrates	Nitrites
Chromium	Mercury	Is_safe (0 – not safe, 1 – safe)

B. Data Structure

The dataset consists of 21 columns and 7999 rows. The variables used for analysis is is_safe as the response variable and the rest of the variables as the predictor variables.

The data structure is:

Aluminium	Copper	Ammonia	...	Is_safe
$X_{1,1}$	$X_{2,1}$	$X_{3,1}$	...	Y_1
$X_{1,2}$	$X_{2,2}$	$X_{3,2}$	...	Y_2
...	...	...	...	...
$X_{1,n}$	$X_{2,n}$	$X_{3,n}$	...	Y_n

C. Analysis Steps

1) Logistic Regression

The step for doing logistic regression is as follows:

1. Import data and library.
2. Conduct data cleaning, including removing NA values and converting the data into the appropriate type of data. Calculate the descriptive statistics of the data.
3. Plot the correlation plot of the data.
4. Checking the balance of response predictor.
5. Split the data into training set and testing set.
6. Build the General Linear Model with binomial family and check the hoslem test value.
7. Removing the predictors variable that are not significant and rebuilding the model.
8. Evaluate the model by building the confusion matrix and calculate the value of accuracy, precision recall and f1 value of the model.

2) K-Nearest Neighbors

The steps for doing K-Nearest Neighbors classification is as follows:

1. Import data and modules.
2. Conduct data cleaning, including removing NA values and converting the data into the appropriate type of data. Calculate the descriptive statistics of the data.
3. Plot the correlation plot of the data.
4. Checking the balance of the response variable.
5. Conduct resampling to balance the data.
6. Split the data into training set and testing set.
7. Implementing the K-NN algorithm using KNeighborsClassifier.
8. Evaluate the model by building the confusion matrix and calculate the value of accuracy, precision recall and f1 value of the model.
9. Enhance the model by standardizing the predictor variables and conducting hyperparameter tuning.

IV. RESULT AND DISCUSSION

A. Descriptive Statistics

First, we are checking whether the data has missing value. We found that only variable ammonia and is_safe has missing value. We proceeded to remove them from the observation. Next, we can proceed to display the descriptive statistics of all variables.

Table IV.1 Descriptive Statistics All Variables

Variable	Min	Median	Mean	Std	Max
Aluminum	0	0.07	0.666	1.27	5.05
Ammonia	-0.08	14.1	14.3	8.88	29.8
Arsenic	0	0.05	0.161	0.253	1.05
Barium	0	1.19	1.57	1.22	4.94
Cadmium	0	0.04	0.0428	0.036	0.13
Chloramine	0	0.53	2.18	2.57	8.68
Chromium	0	0.09	0.247	0.271	0.9
Copper	0	0.75	0.806	0.654	2
Fluoride	0	0.77	0.772	0.435	1.5
Bacteria	0	0.22	0.32	0.329	1
Viruses	0	0.008	0.329	0.378	1
Lead	0	0.102	0.0994	0.0582	0.2
Nitrates	0	9.93	9.82	5.54	19.8
Nitrites	0	1.42	1.33	0.573	2.93
Mercury	0	0.005	0.00519	0.00297	0.01
Perchlorate	0	7.74	16.5	17.7	60
Radium	0	2.41	2.92	2.32	7.99
Selenium	0	0.05	0.0497	0.0288	0.1
Silver	0	0.08	0.148	0.144	0.5
Uranium	0	0.05	0.0447	0.0269	0.09
Is_safe	0	0	0.114	0.318	1

Next, we can continue to check the correlation between the variables to ensure that there are no multicollinearities occurring.

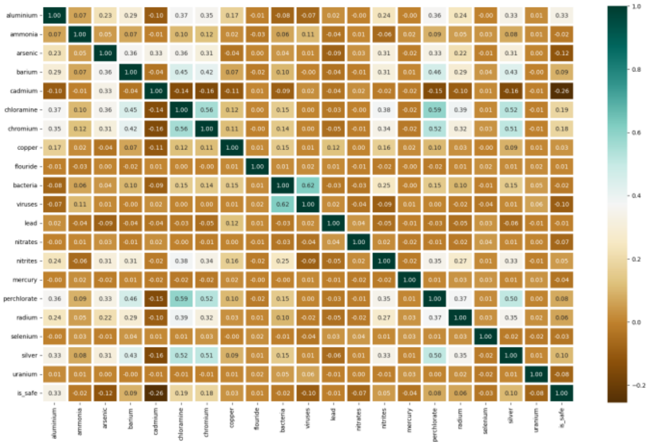


Figure IV.1 Correlation Plot All Variables

From the correlation plot we can see that most of the variables have low correlation with another, indicating that there are no multicollinearities in between variables.

B. *Logistic Regression Classification*

The logistic regression model is built based on the General Linear Model with binomial family. We are fitting is_safe variable as the response variable and the rest as predictor variables. After that we remove the non-significant variables if there are any and proceed to check the Hosmer-Lemeshow GOF test.

Table IV.2 Coefficients of Model 1

Variables	Z-value	P-value
Intercept	2.175	0.029598
Aluminum	19.429	< 2e-16
Ammonia	-4.384	1.17e-05
Arsenic	-8.643	< 2e-16
Barium	2.376	0.017484
Cadmium	-10.806	< 2e-16
Chloramine	8.522	< 2e-16
Chromium	6.123	9.2e-10
Copper	-4.255	2.09e-05
Fluoride	1.651	0.098833
Bacteria	2.734	0.006264
Viruses	-5.750	8.92e-09
Lead	-1.532	0.125540
Nitrates	-5.805	6.45e-09
Nitrites	-2.251	0.024390
Mercury	-2.810	0.00496
Perchlorate	-7.735	1.03e-14
Radium	-2.835	0.004576
Selenium	-3.114	0.001844
Silver	-3.776	0.000159
Uranium	-6.926	4.34e-12

From the table above we can see that fluoride and lead are not significant, the resulted p-value of the 2 variables is both greater than alpha value of 0.05. We can proceed to build a new model without the 2 variables.

Table IV.3 Coefficients of Model 2

Variables	Z-value	P-value
Intercept	2.488	0.012833
Aluminum	19.376	< 2e-16
Ammonia	-4.403	1.07e-05
Arsenic	-8.566	< 2e-16
Barium	2.324	0.020148
Cadmium	-10.774	< 2e-16
Chloramine	8.498	< 2e-16
Chromium	6.133	8.63e-10
Copper	-4.396	1.10e-05
Bacteria	2.794	0.005212
Viruses	-5.761	8.36e-09
Nitrates	-5.943	2.79e-09
Nitrites	-2.270	0.023184
Mercury	-2.759	0.005802
Perchlorate	-7.813	5.58e-15
Radium	-2.755	0.005871
Selenium	-3.055	0.00227
Silver	-3.648	0.000265
Uranium	-6.912	4.79e-12

From this second model we can see that all the variables are already significant. Indicated by the p-value is smaller than alpha value of 0.05. Next, we can proceed to check the Hosmer-Lemeshow Test value of both models.

Hosmer-Lemeshow GOF Test has the following hypothesis:

H_0: The model is appropriate.

H_1: The model is not appropriate.

Table IV.4 Hosmer-Lemeshow GOF Test

Model	X-squared	P-value
Model 1	49.637	4.799e-08
Model 2	40.156	2.996e-06

From Hosmer-Lemeshow GOF Test, we can see that both model is still not appropriate because of the p-value less than the alpha value of 0.05. However, model 2 is considered the better model since the p-value is closer to the alpha value and is greater than the p-value of model 1.

We proceed to plot the confusion matrix and the model evaluating metrics for both model 1 and model 2.

Table IV.5 Confusion Matrix of Model 1

	New 0	New 1
0	1423	8
1	128	40

Table IV.6 Model 1 Evaluating Metrics

	Accuracy	Precision	Recall	F1 Score
New 0	0.915	0.9175	0.9944	0.9544
New 1	0.915	0.8333	0.2381	0.3704

Table IV.7 Confusion Matrix of Model 2

	New 0	New 1
0	1424	7
1	128	40

Table IV.8 Model 2 Evaluating Metrics

	Accuracy	Precision	Recall	F1 Score
New 0	0.9156	0.9175	0.9951	0.9547
New 1	0.9156	0.8511	0.2381	0.3721

Table IV.9 AUC Score of Both Model

	AUC Score
Model 1	0.6163
Model 2	0.6166

From both models evaluating metrics, we can see that model 2 performs slightly better than model 1 but doesn't exist a significant difference from both models.

C. *K-Nearest Neighbors Classification*

We start off doing K-NN classification the same steps as doing a normal logistic regression classification. The difference is that we have imbalance response variable handling when proceeding with K-NN classification.

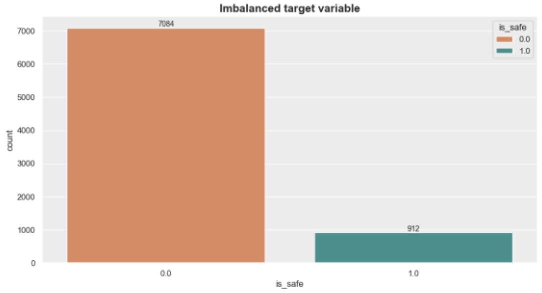


Figure IV.2 Imbalance Response Variable

As we see from the graph, the response variable is dominated by category 0 (not safe). We proceed to balance the response variable using resampling. The sampling method used is Random Over Sampling

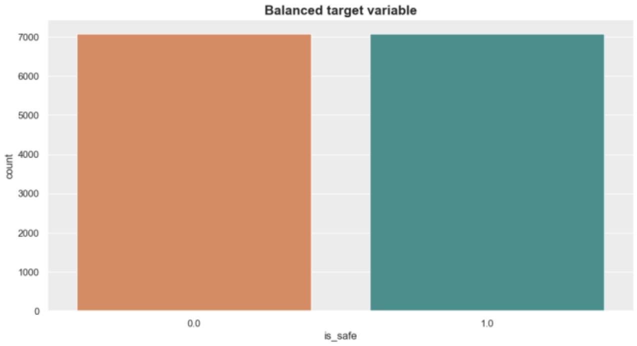


Figure IV.3 Balanced Response Variable

After ensuring that the response variable is balanced, we proceed to split the data into a training set and testing set. Then we continue to build the K-NN classifier, plot the confusion matrix and the evaluation metrics.

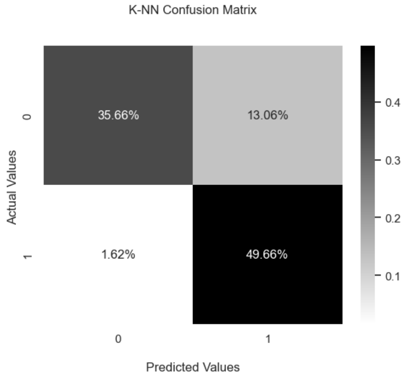


Figure IV.4 Confusion Matrix of K-NN

Table IV.10 Evaluation Metrics of K-NN Classifier

	Accuracy	Precision	Recall	F1 Score
0	0.853	0.96	0.73	0.83
1	0.853	0.79	0.97	0.87

Next, we can continue to further enhance the metrics of the classifier by rescaling and reducing the data and hyperparameter tuning of the K-NN Classifier algorithm.

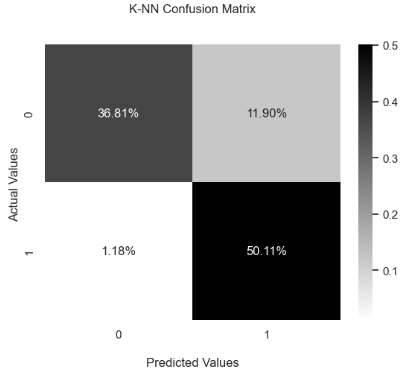


Figure IV.5 Confusion Matrix after Rescale and Reduction

Table IV.11 Evaluation Metrics after Rescale and Reduction

	Accuracy	Precision	Recall	F1 Score
0	0.869	0.97	0.76	0.85
1	0.869	0.81	0.98	0.88

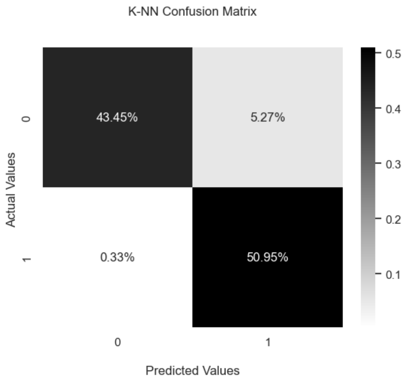


Figure IV.6 Confusion Matrix after Hyperparameter Tuning

Table IV.12 Evaluation Metrics after Hyperparameter Tuning

	Accuracy	**Precision**	**Recall**	**F1 Score**
0	0.944	0.99	0.89	0.94
1	0.944	0.91	0.99	0.95

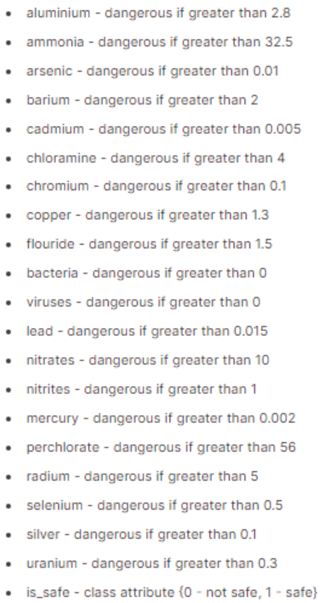
Table IV.13 AUC Score of K-NN Classifier

	AUC Score
K-NN Classification	0.85
K-NN after Rescale & Reduction	0.866
K-NN after Hyperparameter Tuning	0.943

V. CONCLUSION

After seeing both methods of performing a classification task, K-Nearest Neighbors produces a far better classification result than Logistic Regression model. The logistic regression model's result is not optimal because the model that is used is still not appropriate, but the predictor variables are already significant in the model. K-NN on the other hand, does not care about the appropriateness of the model because it is a black-box model.

ATTACHMENT



- aluminium - dangerous if greater than 2.8
- ammonia - dangerous if greater than 32.5
- arsenic - dangerous if greater than 0.01
- barium - dangerous if greater than 2
- cadmium - dangerous if greater than 0.005
- chloramine - dangerous if greater than 4
- chromium - dangerous if greater than 0.1
- copper - dangerous if greater than 1.3
- flouride - dangerous if greater than 1.5
- bacteria - dangerous if greater than 0
- viruses - dangerous if greater than 0
- lead - dangerous if greater than 0.015
- nitrates - dangerous if greater than 10
- nitrites - dangerous if greater than 1
- mercury - dangerous if greater than 0.002
- perchlorate - dangerous if greater than 56
- radium - dangerous if greater than 5
- selenium - dangerous if greater than 0.5
- silver - dangerous if greater than 0.1
- uranium - dangerous if greater than 0.3
- is_safe - class attribute {0 - not safe, 1 - safe}

Figure V.1 Attributes of Variables

BIBLIOGRAPHY

[1] J. Han, M. Kamber, and J. Pei, “Classification: Basic Concepts,” in Data Mining: Concepts and Techniques, Morgan Kaufmann, 2012, pp. 327–386

[2] S. Dreiseitl and L. Ohno-Machado, “Logistic regression and Artificial Neural Network Classification models: A methodology review,” Journal of Biomedical Informatics, vol. 35, no. 5–6, pp. 352–359, 2002. doi:10.1016/s1532-0464(03)00034-0

[3] K. Pereira Teodoro da Silva, A. Kalbusch, and E. Henning, “Detection of unauthorized consumption in water supply systems: A case study using logistic regression,” Utilities Policy, vol. 84, p. 101647, 2023. doi:10.1016/j.jup.2023.101647