Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2025. Vol. 4. No. 2; e-ISSN: 2808-4519

Implementation of GridSearchCV to Find the Best Hyperparameter Combination for Classification Model Algorithm in Predicting Water Potability

Aliyah Kurniasih[1]*, Cantika Nur Previana[2]

[1,2]*Universitas Ary Ginanjar, Indonesia*
aliyah.kurniasih@uag.ac.id [1*], cantika.nur.p@uag.ac.id [2]

Abstract

Drinking water quality is an important factor in public health, so an accurate approach is needed to determine water potability. This research aims to create a water potability prediction model using machine learning methods, with a focus on model accuracy and testing. The dataset used includes various chemical parameters, as well as one radiological and acceptability parameter. In this study, various machine learning algorithms, such as Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression, were applied using GridSearchCV and their performance compared. Models were evaluated using accuracy, precision, recall, F1-score, and confusion matrix metrics, with cross-validation to ensure generalizability. The results showed that the Support Vector Machine algorithm provided the best performance with an accuracy of 70.43%, followed by Random Forest and Logistic Regression with accuracies of 70.12% and 62.20%, respectively. The Support Vector Machine-based model is able to provide reliable predictions and can be used as a tool to support decision-making in water quality management.

Keywords: *Potability, GridSearchCV, SVM, Random Forest, Logistic Regressio*.

1. Introduction

Potability is the measurement and assessment of water quality related to the suitability of water whether the water is suitable for human consumption or not. Potability involves evaluating water on several aspects including chemical aspects, radiological aspects, acceptability aspects, and microbial aspects [1][2].

Chemical aspects such as water pH with a pH value of water that is suitable for consumption, which is ideally between 6.5 and 8.5, because a pH that is too low or too high can cause corrosion and affect the taste of water [3]. Consumable water should not contain harmful heavy metals such as Lead (Pb), Mercury (Hg), Cadmium (Cd), Chromium (Cr), and Arsenic (As) as they can cause serious health problems [4][5]. High levels of nitrate in water can also be harmful, especially to infants and children as it can interfere with the blood's ability to transport oxygen (this condition is called methamoglobinemia or blue baby syndrome) [6]. This also includes other organic chemicals such as Pesticides, Herbicides, and others should not be present in the water at all [7]. Most chemicals that appear in drinking water only become a health problem after years rather than months of exposure, with the exception of nitrate where changes in water quality are usually progressive. Except for substances that are discharged or seep periodically into the flowing surface water or groundwater supply e.g. contaminated landfill sites [1].

The radiological aspect is where water contains radioactive substances that can pose a risk to human health. This risk is smaller than the risk from microorganisms and chemicals [1]. The acceptability aspect is a top priority where water that is suitable for consumption must be clear and colorless, and is related to acceptability in terms of appearance, taste, and smell [8]. The parameter can be the turbidity value or water clarity. High water turbidity can indicate the presence of suspended solid particles that can carry Pathogenic microorganisms [1].

The microbial aspect is related to water that is suitable for human consumption must be free from pathogens such as Bacteria, Viruses, Helminths, and Protozoa that can cause infectious diseases. Then the water must also be free from contamination with Coliform Bacteria which is a group of Bacteria present in the environment and feces of warm-blooded animals. Because the biggest risk to health from the microbial aspect is in water contaminated with human or animal feces. This can be related to inadequate handling of water supply and unsatisfactory management of water distribution [1].

Through water potability where the water is suitable for human consumption or not, which is then also called drinking water. In Indonesia, drinking water is water that has gone through a treatment process or without treatment that meets health requirements and can be drunk directly. Water that qualifies as safe for health if the physical, microbiological, chemical, and radioactive requirements must be met [9].

By 2022, globally at least 1.7 billion people use drinking water sources contaminated with feces. Microbial contamination of drinking water due to fecal contamination poses the greatest risk to drinking water safety. Microbiologically contaminated drinking water can transmit diseases such as diarrhea, cholera, dysentery, typhoid, and polio and is estimated to cause about 505,000 deaths from diarrhea each year [10][11]. Such contaminated water does not meet the main requirements of water whether it is suitable for consumption or not, namely the microbial aspect.

By 2022, 73% of the global population (6 billion people) use safely managed drinking water services, which are on-site, available when needed and free from contamination. This means that the water must meet the acceptability aspect. In 2010, the UN General Assembly expressly recognized the human right to water and sanitation. Everyone has the right to adequate, sustainable, safe, acceptable, physically accessible and affordable water for personal and domestic use [10].

Based on this, the need for water that is suitable for consumption or not is very important and cannot be ignored. To obtain a category of water that is safe for health to drink, a water quality test (water potability) that fulfills several aspects is needed, and this will be very helpful if it is guaranteed to the personal level. Testing these aspects requires a lot of time and money. Therefore, a system is needed that can automatically predict whether the water is suitable for drinking or not. To build the system automatically, machine learning system design is required. One part of machine learning system design is building machine learning modeling from data. Machine learning technology has been used in creating models to predict water potability, including using the Backpropagation Neural Network algorithm [12], Random Forest [13], Extreme Learning Machine [14], and K-Nearest Neighbor [15].

In this study, the authors used the GridSearchCV method during model training to find the best hyperparameter combination to help optimize model results in machine learning algorithms such as Logistic Regression, Random Forest and Support Vector Machine. Then the model is evaluated using the machine learning model evaluation metrix, namely accuracy, precision, recall, f1-score, and confusion matrix to measure the overall performance of the model in determining the results of accurately predicting the feasibility of drinking water later. The results of the best model in this study will be tested using new data and produce predictive results whether water that has these parameter values is suitable for human consumption or not.

2. Related Work

Logistic Regression (LR) is an algorithm that utilizes logistic functions to create modeling in finding patterns of data relationships between model input variables and model targets by producing probability values between 0 and 1 [16]. Therefore, logistic regression algorithm is suitable for machine learning modeling that has a target model to predict the probability of occurrence of an event or category with two target class labels. Equation (1) of logistic regression, where P is the probability of an instance, a is a constant or bias, bX is a coefficient that describes the relationship between attribute X and the probability of an instance [17].

$$P = \frac{e^{a+bX}}{1+e^{a+bX}} \quad (1)$$

Support Vector Machine (SVM) works on the basic principle of a linear classifier where classification cases that have linear data can be well separated. However, with the kernel function, SVM is successfully developed so that it can solve non-linear data problems by transforming data from lower dimensions into a higher dimensional space. In high dimensional space, the hyperplane can maximize the distance (margin) between data classes better, the best hyperplane is located in the middle between two support vectors [18].

Equation (2) of the linear kernel, it is proven that if the value of C is low, it will produce a low margin error value, but it can widen the margin value and ignore points that are close to the decision boundary, and vice versa. So in the linear kernel function, the value of C can affect the margin and the location of the hyperplane [19]. The linear kernel only works by mapping the data into the same feature dimension, meaning that it does not perform any non-linear transformations on the data, the data is only linearly separated in the original feature dimension [20]. Then Equation (3) polynomial kernel, has a degree value to control the flexibility of the classification results, the higher the degree value will allow more flexible decision boundaries. However, a degree value that is too high can cause overfitting [19]. Polynomial kernel in handling data patterns using hard or soft margins, suitable for use on training data that has been normalized [20]. Furthermore, Equation (4) RBF (Radial Basis Function) kernel is also called Gaussian kernel [21], where the gamma parameter is very influential on the performance of the kernel in regulating the kernel distribution, if the value of the gamma parameter is too high, the exponential behavior can become linear and can result in loss of non-linear functions [20].

$$K(x_i, x) = \left(x_i^T, x\right) \quad (2)$$

$$K(x_i, x) = \left(\gamma\left(x_i^T x\right) + r\right)^d \quad (3)$$

$$K(x_i, x) = exp\frac{-\left|\left|x_i^T - x\right|\right|^2}{2\sigma^2} \quad (4)$$

Random Forest (RF) in reducing the risk of overfitting during the training process and producing accurate models in predicting is to create a model by combining the results of several decision trees into one tree [16]. Random forest works by randomly generating features for each node [22].

3. Research Method

This research consists of several stages of research methods, as presented in Fig. 1 below. These stages start from dataset exploration, data visualization, data preprocessing, defining hyperparameters used in each algorithm, training models using GridSearchCV and cross validation, model evaluation and comparison, and finally model testing.

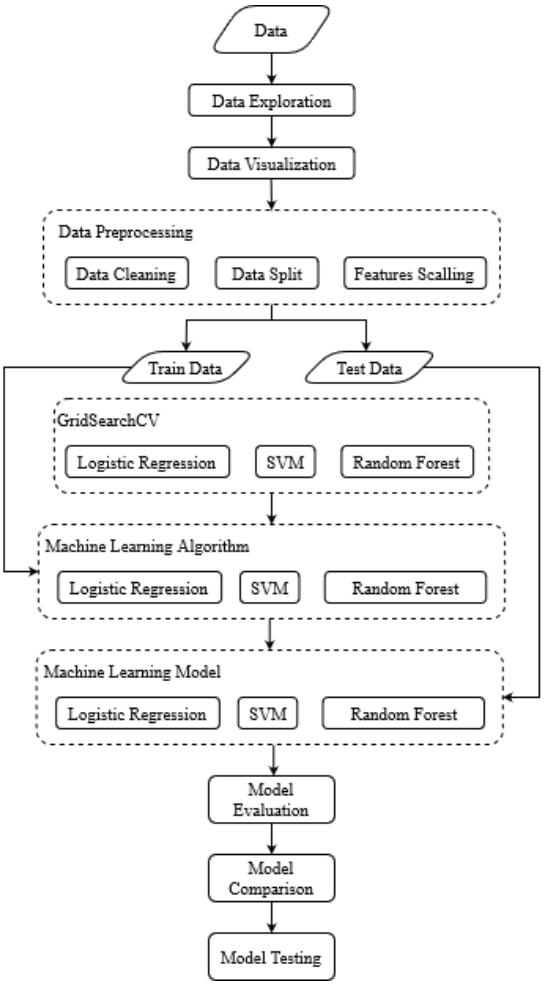


Fig. 1: Research Method

3.1. Data

The data used to build machine learning modeling is sourced from a public dataset on the Kaggle website titled Water Quality and Potability [23].

3.2. Data Exploration

Data exploration is a process where researchers search for initial datasets before any data processing is carried out or datasets that have just been obtained from dataset sources with the aim of understanding the data and the problems that exist in the data, so that later they can really carry out the right data preprocessing process according to the needs of existing data problems. The process carried out at this stage is checking data dimensions, checking feature names, checking the data type of each feature, checking data content, checking duplicate data, checking the amount of missing data from each attribute and the total missing data, and unique data from the target feature model.

3.3. Data Visualization

At the data visualization stage, namely checking the distribution of data based on the target class model with a bar chart and pie chart diagram, where the bar chart is done with the aim of knowing the distribution of data visually from each target class label to make it easier to understand, and the pie chart is done with the aim of knowing the distribution of data in the form of a percentage of the amount of data from each target class.

3.4. Data Preprocessing

Data preprocessing is carried out with the aim of dealing with problems that exist in the dataset so that the data is ready to be used for machine learning modeling training [24]. Data preprocessing includes cleaning the data, namely replacing missing values in each attribute with the median value of each attribute. The median value is used because it is robust to outlier data. Then divide the data into model input data and model output data. Furthermore, model input data and model output data are divided into train data and test data with a percentage

ratio of 90% train data and 10% test data from the total data, where in this process using random state 42. Train data is used for model training, and test data is used for model evaluation.

Furthermore, the model input data in the train data and test data is carried out a feature scalling process, namely changing the value scale of each attribute to a value that has a range of values that is not too extreme high or low. This method is carried out with the aim of being able to help get optimal and better model results in machine learning algorithms. This process is carried out using the Standard Scaler function in Equation (5), where μ is the average of the training data or 0 if the with_mean = False parameter, and σ is the standard deviation of the training data or 1 if with_std = False, standard scaler has with_mean and with_std by default True. The standard scaler function on the train data is to normalize the data where each feature has a mean range of 0 and a standard deviation of 1. The standard scaler on the test data is to use the parameters obtained from the fit on the training data. The standard scaler function works by centering and scaling the data independently of each feature by calculating the relevant statistics on the data in the training data, then the average and standard deviation are stored for use in subsequent data using transforms. However, the standard scaler is sensitive to outlier data so that each feature can scale differently from each other in the presence of outlier data. This means that the standard scaler functions to normalize the data by standardizing the data from each feature by removing the mean and scaling to the unit variance [25].

$$X_{standard} = \frac{X-\mu}{\sigma} \quad (5)$$

3.5. GridSearchCV

GridSearchCV method is a technique used in machine learning to find the hyperparameter value with the best combination during model training. GridSearchCV is used to determine the optimal hyperparameter combination based on the evaluation metrics used. This method works by trying all combinations of hyperparameter values that have been defined and evaluating the model performance for each combination [16]. Table 1 is a hyperparameter along with variations in hyperparamter values used in this study, namely in the Logistic Regression, Support Vector Machine and Random Forest algorithms.

Table 1: Hyperparameter Tuning GridSearchCV

Algorithm	Hyperparameter		Values
Logistic Regression	penalty		L1 dan L2
	C		1, 10 dan 100
Support Vector Machine	Kernel RBF	Gamma	1, 0.1, dan 0.01
		C	1, 10, dan 100
	Kernel Linear	C	1 dan 10
	Kernel Polynomial	Degree	4, 6 dan 8
		C	1 dan 10
Random Forest	n_estimators		100, 200, dan 300
	max_depth		None, 10, 20 dan 30
	min_samples_split		2, 5, dan 10

3.6. Model Training

Model training is the process of training a model using algorithms and hyperparameters along with variations in hyperparameter values that have been defined, trained using training data that has gone through the data preprocessing process. Where in this process the model will look for patterns in the data so that it can produce a rule that the model can later use in predicting new data to be able to determine the output automatically based on the data it inputs. In this research, model training is carried out using *machine learning* algorithms, namely the *Logistic Regression* algorithm, *Support Vector Machine* and *Random Forest*. The training process uses *GridSearchCV* to determine the best hyperparameter combination and *cross validation* 10. The results of the training model with the best hyperparameter combination will then be used to evaluate the model.

3.7. Model Evaluation and Comparison

Model evaluation is the process of seeing the overall performance of the model after model training, the model is evaluated using *test data*, namely data that is completely unrecognized during the training process with the aim that the model is not biased. Model performance metrics used include *accuracy* (6), *recall* (7), *precision* (8), and *f1-score* (9) functions [26]. In the formula, it is known that there are TP parameters which in this study are interpreted as true 'not potable', TN is true 'potable' FP is false 'not potable', and FN is false 'potable'. Furthermore, the evaluation results of each model are compared by looking at the results of the highest accuracy value of the model, then the model can be said to be the best in this study.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$recall = \frac{TP}{TP+FN} \quad (7)$$

$$precision = \frac{TP}{TP+FP} \quad (8)$$

$$f1-score = \frac{2\times precision\times recall}{precision+recall} \quad (9)$$

3.8. Model Testing

Model testing can also be said to be a deployment model, which is the process of testing the model by providing data input on each feature of the input model for further automatic predictions related to whether a water quality with the values of these features is suitable for consumption or not. Deployment of the model is made based on a web base locally using *streamlit*. This process is carried out only on the best model generated from this research, by generating the probability and percentage values of each class and its prediction results.

4. Result and Discussion

4.1. Data Exploration Results

Based on the results of checking at the data exploration stage, there are 3,276 data and 10 *features* including *features* 'ph', 'Hardness', 'Solids', 'Chloramines', 'Sulfate', 'Conductivity', 'Organic_carbon', 'Trihalomethanes', 'Turbidity', 'Potability'. There are 9 *features* with *float* data type and 1 *feature* with *int* data type, where the data content of each *feature* is in accordance with the data type. Then there are indications of *missing values* with a total of 1,434 input *missing values* in the *features* 'ph', 'Sulfate', 'Trihalomethanes', but there are no indications of duplicate data. Details of the data exploration results along with a description of each feature and aspect type are presented in Table 2 below.

Table 2: Data Exploration Results

No	Features	Data Type	Missing Values	Values of Sample of Data	Description	Aspect Type
1	Ph	Float	491	9.092223456290965	Water ph level	Chemical
2	Hardness	Float	0	181.10150923612525	Size of mineral content	Chemical
3	Solids	Float	0	17978.98633892625	Total density dissolved in water	Chemical
4	Chloramines	Float	0	6.546599974207941	Chloramine concentration in water	Chemical
5	Sulfate	Float	781	310.13573752420444	Sulfate concentraion in water	Chemical
6	Conductivity	Float	0	398.4108133818447	Electrical conductivity of ater	Radiology
7	Organic Carbon	Float	0	11.558279443446397	Organic carbon in water	Chemical
8	Trihalomethanes	Float	162	31.997992727424737	Concentration of trihalomethanes in water	Chemical
9	Turbidity	Float	0	4.075075425430034	Turbidity level, a measure of water clarity	Acceptability
10	Potability	Int	0	0	Target variable; indicates the water's suitability with values of 1 (potable) and 0 (not potable).	Potable water eligibility

4.2. Data Visualization Results

Furthermore, Fig. 2 is a *bar chart* of the distribution of the total data based on the model's target class label on the 'Potability' *feature*, and it is known that there are a total of 1,998 data with the target class ‘not potable’ (0) and 1,278 data with the target class ‘potable’ (1).

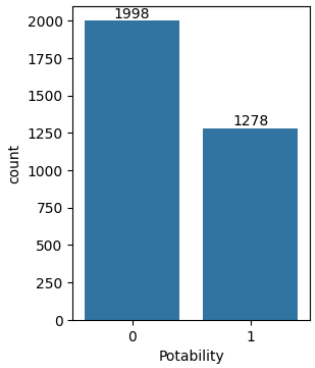


Fig. 2: Data Distribution Based on Model Target Class

Fig. 3 is the percentage of the amount of data based on the model's target class label from the total data, it is known that there are 60.99% for the ‘not potable’ (0) class and 39.01% for the ‘potable’ (1) class.

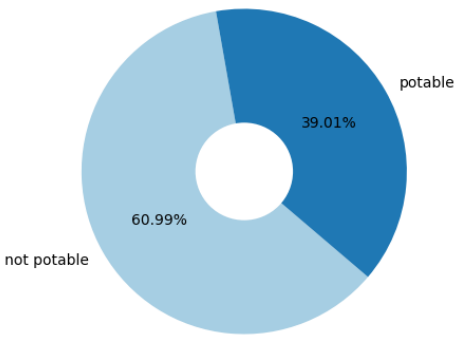


Fig. 3: Percentage of Data Based on the Target Class

4.3. Preprocessing Data Results

Data preprocessing is carried out by three methods, namely first, data cleaning is carried out to fill in the missing value with the median value. The second method is data division by dividing the data into data for model input and model target first, which then the model input

data and model target data are further divided into train data and test data. In this study, the input model consists of features 'ph', 'Hardness', 'Solids', 'Chloramines', 'Sulfate', 'Conductivity', 'Organic_carbon', 'Trihalomethanes', 'Turbidity', and the model target, namely the 'Potability' feature. The results of data division are presented in Table 3, 90% of the training data is 2,948 data, and 10% of the test data is 328 data. Where the training data has a total of 1,794 data with a target class of 'not potable' class (0), and 1,154 data with a target class of 'potable' class (1). Then the test data selected a total of 204 data of the 'not potable' class (0) and a total of 124 data of the 'potable' class (1). The results of the data division are presented in Table 3 below. The last step of data preprocessing is feature scalling which is carried out only on model input data, including training data and test data. Where in the input of the training data model feature scalling is carried out with fit_transform, then on the test data only with transforms, and the results of feature scalling on the training data are stored for further use at the model testing stage (model deployment).

Table 3: Data Split Results

Data Split	Amount	Amount of Label
Train	2.948	Not Potable = 1.794 Potable = 1.154
Test	328	Not Potable = 204 Potable = 124

4.4. GridSearchCV Results

Table 4 shows the best combination of hyperparameters of each algorithm used at the time of model training. The GridSearchCV method succeeded in finding the best combination of hyperparameters, namely the L2 penalty and the value of C=1 in the logistic regression algorithm. Then the Support vector machine produces a combination in the form of an RBF (Radial Basic Function) kernel, gamma value = 0.1 and C value = 1. Furthermore, the random forest algorithm produces the best combination at max_depth=None, min_samples_split=2, and n_estimators=300. Furthermore, the model with the best combination of hyperparameter results will be used at the model evaluation and model testing stages.

Table 4: Hyperparameter Tuning Result

Algorithm	Hyperparameter	Values	Best Score for Training
Logistic Regression	Penalty C	L2 1	60,85%
Support Vector Machine	Kernel Gamma C	RBF 0,1 1	67,40%
Random Forest	max_depth min_samples_split n_estimators	None 2 300	68,35%

4.5. Model Evaluation Results

Table 5 is the results of the model evaluation with the metric performance model accuracy, precision, recall and f1-score. Where the value of the evaluation results from each machine learning algorithm uses the best combination of hyperparameters obtained from the model training process using GridSearcCV. The best model was found in the support vector machine algorithm with an accuracy of 70.43%, which means that 70.43% of the model can make correct predictions, which is the ratio between the number of correct predictions and the total number of predictions. It then results in a recall value of 70%, which means that 70% of the model can read instances that can be correctly classified as a specific class. Furthermore, the SVM model produces a precision value of 71%, which means that 71% of the model successfully predicts the 'not potable' class correctly. SVM also produces an f1-score of 67%, which means that 67% of the model can describe a weighted average comparison of precision and recall.

Table 5: Evaluation Model Result

Algorithm	Accuracy	Recall	Precision	F1-score
Logistic Regression	62,20%	62%	76%	48%
Support Vector Machine	70,43%	70%	71%	67%
Random Forest	70,12%	70%	70%	67%

Fig. 4 is the table confusion matrix from the results of the evaluation of the support vector machine model. The confusion matrix provides more detailed details about the prediction of the correct and false classifications for each class. Based on the table, of the 204 test data with the 'not potable' class, only 190 data were correctly classified as 'not potable' (TP), and 83 data (FP) failed to be classified. Then out of a total of 124 test data with the 'potable' class, there are a total of 41 data that have been correctly classified as 'potable' (TN), and only 14 data are incorrectly classified as 'potable' (FN) class.

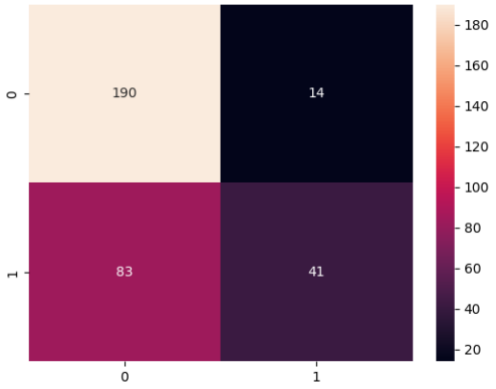


Fig. 4: Confusion Matrix SVM

The results of model training that have passed the model evaluation stage of each algorithm are stored, where the best model, namely the SVM algorithm, will be reused at the model testing stage (model deployment).

4.6. Model Test Results

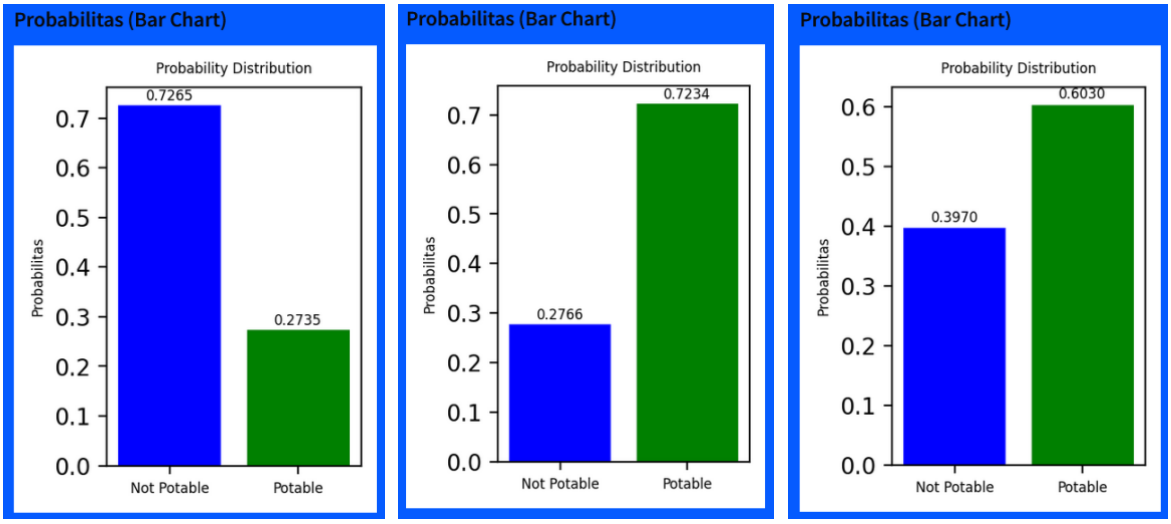
In the testing phase of the model, a web-based system was built using Streamlit locally. The best model with SVM and Feature Scalling algorithms has been stored in the deployment of the Streamlit program, which is then tested with 3 different data inputs. In the first and second Fig. 5 (a) and (b), the input data for testing is taken from the data in the dataset, where the actual value of the prediction results is known, namely 'not potable' for the first input, and 'potable' for the second input. Then the third data input Fig 5 (c), the input for testing is used a random number according to the author's input at the time of testing, where in this data input with the value for each feature used in the estimate is below the value in the second input data, the purpose is to make the data easy to analyze whether the prediction results are correct or false, with the reference to the data value used not far from the range of values with the data, namely the value of the data that has the actual prediction 'potable'.

(a) First Data Input (b) Second Data Input (c) Third Data Input

Fig. 5: Input Data

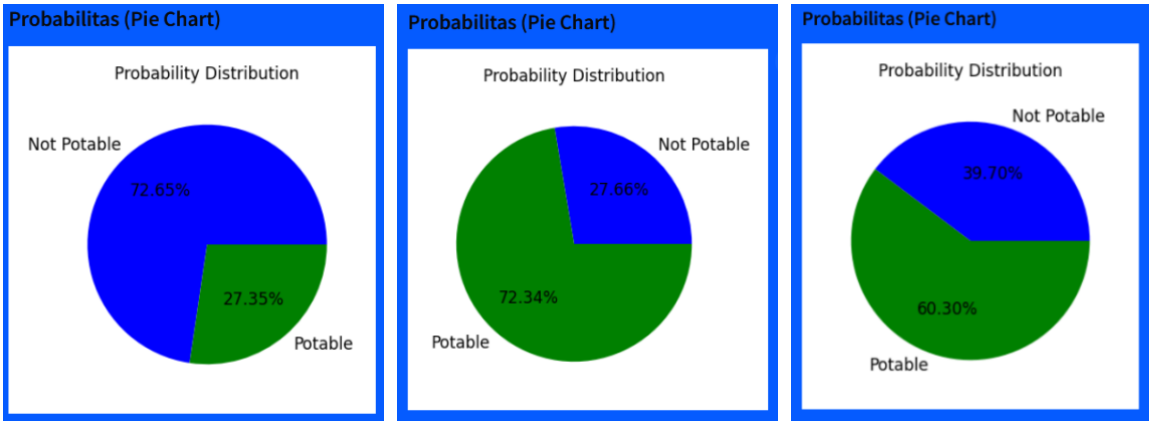
Fig. 6 (a) is the prediction result for the first input data, resulting in a prediction with a class of 'not potable' with a probability value of 0.7265 or 72.65% Fig. 7 (a), the result of this prediction has been in accordance with the actual value of 'not potable'. Then on Fig. 6 (b) is the prediction result for the second input data, the model produces a prediction with a 'potable' class which has a probability value of 0.7234 or 72.34% Fig. 7 (b), where the result of this prediction has been in accordance with the actual value of 'potable'. Furthermore, in the third data input, the model finds a prediction result with a class 'potable' with a probability value of 0.6030 or 60.30% Fig. 7 (c), where this

prediction result is in accordance with the assumption based on the value of each feature that is inputted not far from the value with the actual data, namely 'potable'.



(a) First Data Input (b) Second Data Input (c) Third Data Input

Fig. 6: Prediction Result (Bar Chart)



(a) First Data Input (b) Second Data Input (c) Third Data Input

Fig. 7: Prediction Result (Pie Chart)

5. Conclusion

The best model produced on the Support Vector Machine algorithm with RBF (Radial Basic Function) kernel, gamma=0.1, and C=1, overall shows the performance of the model that can successfully capture non-linear data patterns well, and the model is robust enough to noise. The relatively small gamma value for the RBF kernel makes the model more focused on global patterns, meaning that each data point has a greater influence on the patterns formed, the model will be able to separate complex data without overfitting, and the model has a wider field of influence for each support vector so that it is able to capture more global relationships in the data. This is evidenced by the model evaluation accuracy value of 70.43%, which is quite close to the accuracy result at the time of model training of 67.40%, where this only has an accuracy difference of 3.03%, meaning that the model has good data generalization. At a relatively small C=1 value on the kernel, RBF allows the model to be able to find larger margins, thus providing tolerance for classification errors in the training data, and helping the model to avoid overfitting. This is also proven at the stage of testing the SVM model on a system built using streamlit, the model is input with unknown data in the training data and test data, and proves that the prediction results from the input data are successfully predicted correctly by having a fairly good probability value.

Suggestions for further research in order to be able to create a model using a dataset that has the 4 aspects criteria in full, where the data used is indeed real data that occurs in the field. And can try to build models using machine learning algorithms and other methods. This research opens up opportunities for further development by incorporating spatial and temporal data to improve the accuracy of the model in various environmental conditions.

References

[1] W. H. O. (WHO), “Guidelines for drinking-water quality: fourth edition incorporating the first and second addenda.” [Online]. Available: https://www.who.int/publications/i/item/9789240045064.

[2] U. S. E. P. A. (EPA), “National Primary Drinking Water Regulations.” [Online]. Available: https://www.epa.gov/ground-water-and-drinking-water/national-primary-drinking-water-regulations.

[3] S. Handayani, Sudarti, and Yushardi, “Analisis Kualitas Air Minum Berdasarkan Kadar PH Air Mineral dan Rebusan Sebagai Sumber Energi Terbarukan,” *Opt. J. Pendidik. Fis.*, vol. 7, no. 2, pp. 385–395, 2023.

[4] T. T. Irianti, Kuswandi, S. Nuranto, and A. Budiyatni, "Logam Berat & Kesehatan," *Buku Logam Berat Kesehat.*, pp. 1–131, 2017.

[5] M. Zaynab *et al.*, "Health and Environmental Effects of Heavy Metals," *J. King Saud Univ. - Sci.*, vol. 34, no. 1, p. 101653, 2022.

[6] W. S. D. of Health, "Nitrate in Drinking Water," *Washington State Department of Health*. [Online]. Available: https://doh.wa.gov/community-and-environment/drinking-water/contaminants/nitrate.

[7] I. El-Nahhal and Y. El-Nahhal, "Pesticide residues in drinking water, their potential risk to human health and removal options," *J. Environ. Manage.*, vol. 299, no. August, p. 113611, 2021.

[8] I. Paradis, U. Syamsudin, and M. I. Rantau, "Optimalisasi Pelayanan Air Minum Oleh PDAM Tirta Benteng Kota Tangerang," *J. Ilm. Wahana Pendidik.*, vol. 10, no. 8, pp. 491–528, 2024.

[9] Permenkes RI, "Peraturan Menteri Kesehatan Republik Indonesia Nomor 492/Menkes/Per/IV/2010 Tentang Persyaratan Kualitas Air Minum," *Peraturan Mentri Kesehatan Republik Indonesia*. p. MENKES, 2010.

[10] W. H. O. (WHO), "Drinking-water." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/drinking-water.

[11] SDGs, "Summary Progress Update 2021 : SDG 6 — water and sanitation for all," *UN-Water Integr. Monit. Initiat.*, pp. 1–58, 2021.

[12] S. Y. Kurniawan, S. Sanjaya, Y. Vitriani, and I. Afrianty, "Klasifikasi Kelayakan Air Minum dengan Backpropagation Neural Network Berbasis Penanganan Missing Value dan Normalisasi," *J. Inf. Syst. Res.*, vol. 6, no. 1, pp. 87–95, 2024.

[13] K. Abdi, A. Warjaya, I. Muthmainnah, and P. H. Pahutar, "Penerapan Algoritma Random Forest dalam Prediksi Kelayakan Air Minum," *J. Ilmu Komput. dan Inform.*, vol. 3, no. 2, pp. 81–88, 2024.

[14] Y. V. Sari, Z. Muallifah, and A. Fanani, "Klasifikasi Kualitas Air Menggunakan Metode Extreme Learning Machine (ELM)," *J. JUPITER*, vol. 15, no. 2, pp. 983–994, 2023.

[15] F. Malik Namus Akbar, "Metode KNN (K-Nearest Neighbor) untuk Menentukan Kualitas Air," *J. Tekno Kompak*, vol. 18, no. 1, pp. 28–40, 2024.

[16] Achmad Baroqah Pohan, Irmawati, and A. Kurniasih, "Optimization of Classification Algorithm with GridSearchCV and Hyperparameter Tuning for Sentiment Analysis of the Nusantara Capital City," *J. Artif. Intell. Eng. Appl.*, vol. 3, no. 3, pp. 808–814, 2024.

[17] M. P. Pulungan, A. Purnomo, and A. Kurniasih, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Kepribadian MBTI Menggunakan Naive Bayes Classifier," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 7, pp. 1493–1502, 2023.

[18] P. A. Octaviani, Y. Wilandari, and D. Ispiriyanti, "Penerapan Metode Klasifikasi Support Vector Machine (SVM) pada Data Akreditasi Sekolah Dasar (SD) di Kabupaten Magelang," *J. Gaussian*, vol. 3, no. 8, pp. 811–820, 2014.

[19] A. Z. Praghakusma and N. Charibaldi, "Komparasi Fungsi Kernel Metode Support Vector Machine untuk Analisis Sentimen Instagram dan Twitter (Studi Kasus : Komisi Pemberantasan Korupsi)," *JSTIE (Jurnal Sarj. Tek. Inform.*, vol. 9, no. 2, p. 23342, 2021.

[20] S. D. Wahyuni and R. H. Kusumodestoni, "Optimalisasi Algoritma Support Vector Machine (SVM) Dalam Klasifikasi Kejadian Data Stunting," *Bull. Inf. Technol.*, vol. 5, no. 2, pp. 56–64, 2024.

[21] E. Rizqi Mar'atus Sholiihah, I. G. Susrama Mas Diyasa, and E. Yulia Puspaningrum, "Perbandingan Kinerja Kernel Linear Dan Rbf Support Vector Machine Untuk Analisis Sentimen Ulasan Pengguna Kai Access Pada Google Play Store," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 728–733, 2024.

[22] Suci Amaliah, M. Nusrang, and A. Aswi, "Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng," *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 4, no. 3, pp. 121–127, 2022.

[23] L. Tharmalingam, "Water Quality and Potability," *Kaggle Dataset*, 2023. [Online]. Available: https://www.kaggle.com/datasets/uom190346a/water-quality-and-potability.

[24] M. R. Fatturrahman and A. Kurniasih, "Penggunaan Metode NearMiss, SMOTE, dan Naïve Bayes untuk Klasifikasi Gangguan Tidur Berdasarkan Kualitas Tidur dan Gaya Hidup," *Pros. Semin. Nas. Mhs. Bid. Ilmu Komput. dan Apl.*, vol. 4, no. 2, pp. 567–576, 2023.

[25] F. Rachmawati, J. Jaenudin, N. B. Ginting, and P. Laksono, "Machine Learning for the Model Prediction of Final Semester Assessment (FSA) using the Multiple Linear Regression Method," *J. Tek. Inform.*, vol. 17, no. 1, pp. 1–9, 2024.

[26] A. Kurniasih and L. P. Manik, "On the Role of Text Preprocessing in BERT Embedding-based DNNs for Classifying Informal Texts," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 6, pp. 927–934, 2022.