The Indonesian Journal of Computer Science
www.ijcs.net
Volume 13, Issue 4, August 2024
https://doi.org/10.33022/ijcs.v13i4.3884

Predictive Analytics for Water Safety: Data Mining and Supervised Learning in Potability Classification

Nanda Aulia Sofiah[1], Fanny Olivia[2], Muhammad Ihsan Jambak[3]
09031182126017@student.unsri.ac.id[1], 09031182126011@unsri.ac.id[2], jambak@unsri.ac.id[3]
[1,2] Information System, Faculty of Computer Science, Sriwijaya University, Indonesia
[3] Informatics Management, Faculty of Computer Science, Sriwijaya University, Indonesia

Article Information

Received : 30 Mar 2024
Revised : 22 May 2024
Accepted : 30 Jun 2024

Keywords

Classification, Data Mining, Supervised Machine Learning, Water Potability

Abstract

Water is crucial for survival, especially for consumption, yet its quality is under threat due to human-caused pollution. Contaminated water poses serious health risks, including the transfer of diseases transmitted by water. Therefore, assessing water quality is critical for ensuring its safety for consumption. Data mining and supervised machine learning algorithms can help classify water potability, revealing hidden patterns and correlations between water parameters. This study evaluates the effectiveness of K-Nearest Neighbors (KNN), Naïve Bayes, Support Vector Machine (SVM), and Neural Network methods in categorizing a water quality dataset. The evaluation is aimed at selecting the most accurate procedure, as indicated by the highest accuracy rate. Results show that Neural Network exceeds KNN (81%), Naïve Bayes (63%), and SVM (73%), with a 85% accuracy rate.

A. Introduction

Water is one of the most crucial elements on earth. Without water, any living thing, especially humans will be struggling in order to maintain their survival. These days, water is an essential component for human life, as they need it for various activities, including consumption [1]. It is unimaginable how human condition would be if there was not any clean water supply to help them carry on their activities. However, not all water is suitable for drinking, thus it becomes necessary to determine the quality of water that is suitable for consumption.

As the time goes by, plenty of water has been contaminated by pollution from the rapid development of constructions and factory activities, which has become the main reason for declining water quality [2]. Many chemical elements have dissolved into the water, making not all water safe for consumption [2]. In addition to chemical and microbiological contaminants, physical aspects such as turbidity and the color of water should also be considered when examining the suitability of water.

Water is a medium that easily transmits diseases, even over long distances. These polluted water may lead to various waterborne diseases including cancer, disruption, endocrine, developmental problems, reproductive issues, and cardiovascular [3]. Poor water quality has negative impacts and poses a significant threat to public's health, hence it is necessary to assess water quality in order to know whether it is safe to be consumed or not.

Therefore, if people do not know how to evaluate the quality of water, then there will undoubtedly be a lot of issues. Still, the majority of the water potability procedure is carried out manually in the lab, which may be challenging and time-consuming when dealing with large amounts of data. Data mining is therefore required to help with the categorization of water potability. The rapid and precise analysis of vast volumes of data using data mining facilitates the process of determining the potability of water, hence enhancing its efficacy and efficiency in water quality management [4].

When there's a pressing need to swiftly and properly analyse vast amounts of water quality data—like in the case of natural natural disasters, industrial pollution situations, or population growth—data mining becomes essential for classifying water potability. Today's manual laboratory procedures are ineffective owing to the requirement to handle massive volumes of data; data mining is a crucial answer since it makes it possible to quickly identify patterns and abnormalities in the data [5]. Since they offer the data required to guarantee safe water supplies, implement preventative actions, and successfully maintain public health, the categorization findings are highly valuable to health authorities, managers of water resources, and the general public.

Water potability can be classified using a data mining approach and supervised machine learning algorithms. The application of data mining techniques to the problem of water potability classification has numerous significant advantages. In particular, data mining techniques enable the processing of huge and complex water quality data, which is sometimes impossible to examine manually. Data mining tools, with their ability to reveal hidden patterns and complicated correlations between diverse water parameters, are useful in discovering crucial elements influencing water potability. Data mining can analyze massive data patterns and then learn them to produce usable information, whereas a supervised machine learning system can

predict the possibility that the water is safe to consume. So we'll find out what the most important factors are in identifying which air is safe to drink and which is not. The application of supervised machine learning for water potability classification has the potential to improve decision-making efficiency and accuracy in the environmental and public health sectors.

In this study, we will explore the classification by implementing supervised machine learning techniques such as K-Nearest Neighbors, Naive Bayes, Support Vector Machine, and Neural Network. This research uses data that we collected from the open-source platform Kaggle to perform water potability classification. We intend to evaluate the effectiveness and performance of numerous algorithms, based on a range of chemical and physical characteristics to measure the accuracy of each model. These methods that involve data analysis and predictive modeling, these techniques provide a reliable solution for accurately categorizing water based on its potability. Through this exploration, we are hoping to contribute to the expanding body of knowledge about water quality assessment and the use of machine learning in environmental science.

B. Research Method

We address these challenges in using the Cross-Industry Standard for Data Mining (CRISP-DM) method to asess water quality data from the open-source platform Kaggle, which consist of six phases of process. Those six stages are business understanding, data understanding, data preparation, modeling, evaluation, and deployment as shown in the Figure 1. CRISP-DM, compared to other data mining methodologies, is more comprehensive and well-documented. Each stage is well organized and has defined definitions, making it easy to implement even for beginners [6]. This methodology is often chosen as a guide for managing data mining projects to ensure the process is structured and efficient.

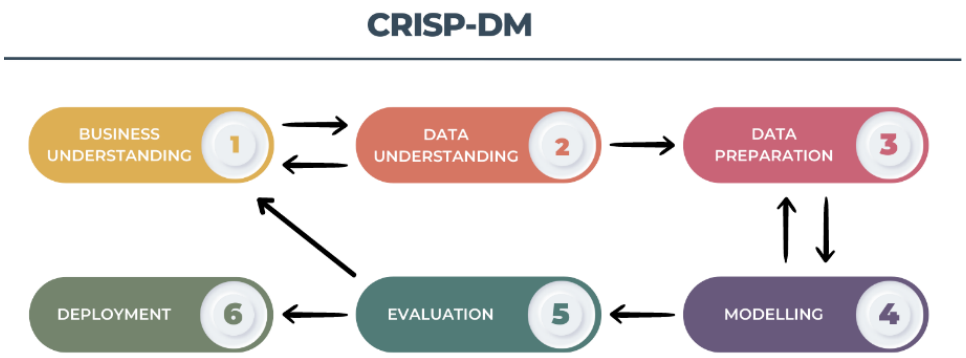


Figure 1. CRISP – DM Method

2.1. Business Understanding

This initial phase requires us to analyze the situation of a business in order to gain an overview of available and required resources [7]. At this phase, it is critical to grasp the value of using water characteristic data to determine the classification of water feasibility and predict whether the water is safe to use or not. In this case, we need to assess the water quality for safety measurement, avoiding any unqualified water for consumption.

2.2. Data Understanding

Data Understanding is required in order to meet the primary goal of this study. Understanding data involves steps to prepare data, evaluate data needs, and includes data collection. There is a close connection between business understanding and data understanding where data collection is carried out at this stage. Understanding the needs and desires of the business will guide data selection, including the data sources and collection methods used.

2.2.1. Data Variable

There are a total of 3276 data points, with 9 columnincluding pH, hardness, solids, chloramines, sulfate, conductivity, trihalomethanes, turbidity, organic carbon and potability as shown in the Table 1.

Table 1. Water Quality Metrics

No	Identity	*Type*	*Explanation*
1	pH	Float	A metric to determine the acid-base balance of water.
2	Hardness	Float	The ability of water stimulated by Calcium and Magnesium to precipitate soap.
3	Solids (total dissolved solids)	Float	The water with high TDS value indicates that water is highly mineralized.
4	Chloramines	Float	Chloramines are typically produced by combining chlorine with ammonia during the treatment of drinking water.
5	Sulfate	Float	A chemical compound composed of sulfur and oxygen atoms (SO4) that occurs naturally in minerals, soil, water, and living beings.
6	Conductivity	Float	Pure water is not a weak conductor of electricity; instead, it serves as an effective insulator.
7	Organic carbon	Float	Total Organic Carbon (TOC) in source waters develops from both decaying natural organic matter (NOM) and manmade sources.
8	Trihalomethanes	Float	Chemicals that might be present in water treated with chlorine
9	Turbidity	Float	A method of assessing the luminous properties of water to gauge the quality of wastewater discharge in relation to colloidal particles
10	Potability	Integer	Indicates whether or not the water is safe for human consumption, with 1 which implies potable and 0 not potable.

2.2.2. Data Distribution

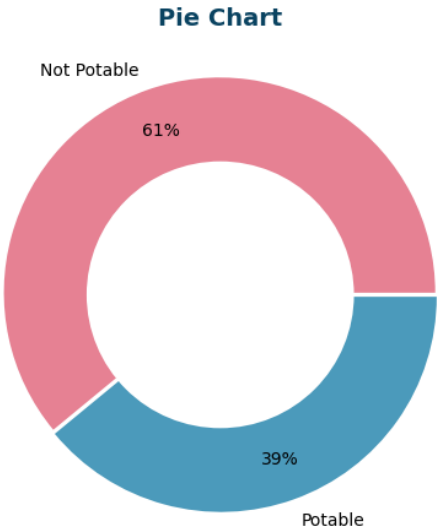


Figure 2. Potability Distribution

Figure 2 shows the data distribution of the potability of water, with 61% of the data which is 1998 rows of data is about non-potable water. The rest of the dataset, around 31% or 1278 rows of data is about potable water

2.2.3. Correlation Matrix

Figure 3 below shows the result of the correlation between each of the variable. The highest correlation reached 0.08 between pH and hardness. Meanwhile, the correlation between sulfate and solids with -0.15 correlation.

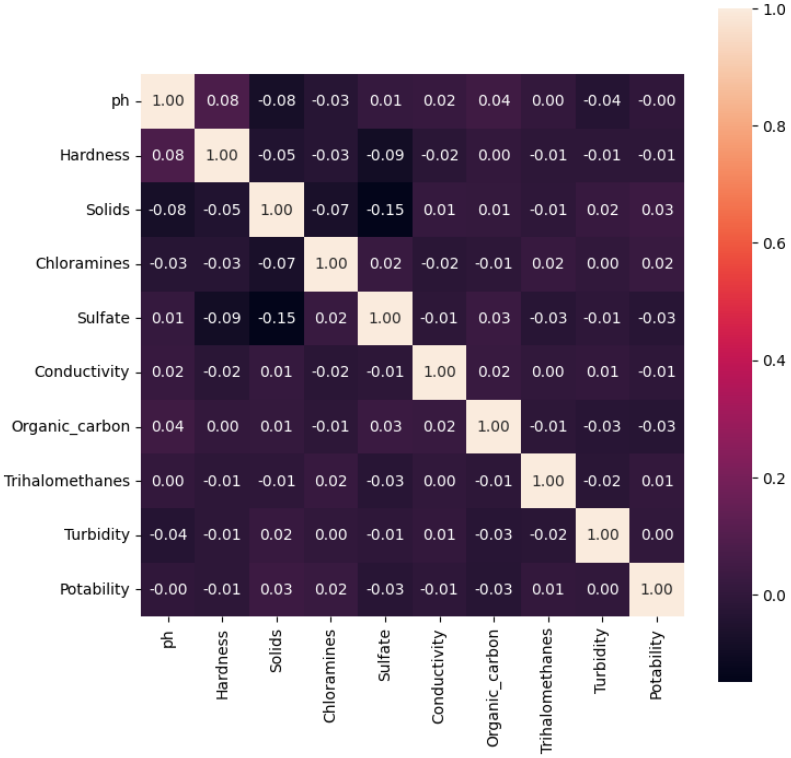


Figure 3. Correlation Matrix

2.3. Data Preparation

The raw dataset must be prepared and cleaned before being processed into the model. Data preparation is the stage in which data problems are fixed before the data is used in the modeling stage, resulting in high-quality modeling. Data preparation begins after the data has been successfully collected and requires the steps of identification, selection, cleaning, and arrangement in the desired format. Once the dataset is selected, the next step is to evaluate the possibility of doubtful cases, missing data, or ambiguous information [8].

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   ph               3276 non-null   float64
 1   Hardness         3276 non-null   float64
 2   Solids           3276 non-null   float64
 3   Chloramines      3276 non-null   float64
 4   Sulfate          3276 non-null   float64
 5   Conductivity     3276 non-null   float64
 6   Organic_carbon   3276 non-null   float64
 7   Trihalomethanes  3276 non-null   float64
 8   Turbidity        3276 non-null   float64
 9   Potability       3276 non-null   int64  
dtypes: float64(9), int64(1)
memory usage: 256.1 KB
```

Figure 4. Data Information

Based on the results from Figure 4, we can understand the information that this data has the type of float and integer, and does not have any missing values, which means it is already clean and ready to be processed by the model. If the data has already been cleaned, we have to split it. For this study, 60% of the data was allocated to training, 20% to validation, and the remaining 20% to testing as shown in the Figure 5 below. We also set the oversample to false for this research so that the originality data is kept.

```
[10]  1 train, valid, test = np.split(df.sample(frac=1), [int(0.6*len(df)), int(0.8*len(df))])
```

```
[11]  1 def scale_dataset(dataframe, oversample=False):
      2   X = df[df.columns[:-1]].values
      3   y = df[df.columns[-1]].values
      4
      5   scaler = StandardScaler()
      6   X = scaler.fit_transform(X)
      7
      8   if oversample:
      9     ros = RandomOverSampler()
     10     X, y = ros.fit_resample(X, y)
     11
     12   data = np.hstack((X, np.reshape(y, (-1, 1))))
     13
     14   return data, X, y
```

```
[12]  1 train, X_train, y_train = scale_dataset(train, oversample=False)
      2 valid, X_valid, y_valid = scale_dataset(valid, oversample=False)
      3 test, X_test, y_test = scale_dataset(test, oversample=False)
```

Figure 5. Splitting Data

2.4 Modeling

Modeling is the process that occurs when the necessary parameters with ideal values are determined, along with the data mining techniques to be applied, the tools and algorithms chosen. The process of developing a prediction model, which categorizes water quality according to nine characteristics from a water feasibility test, is called data modeling. Tools from Google Colab will help with the development of the data model with the algorithms of KNN, Naive Bayes, Support Vector Machine, and Neural Network as shown in Table 2 below.

Table 2. Program Code for Each Model

Modeling	
Model 1 : K-Nearest Neighbor	Model 2 : Naïve Baiyes

`[11] 1 from sklearn.neighbors import KNeighborsClassifier`
`2 from sklearn.metrics import classification_report`
 	
`[12] 1 knn_model = KNeighborsClassifier(n_neighbors=3)`
`2 knn_model.fit(X_train, y_train)`
`3 y_pred = knn_model.predict(X_test)`
	`[15] 1 from sklearn.naive_bayes import GaussianNB`
	
	`[16] 1 nb_model = GaussianNB()`
	`2 nb_model = nb_model.fit(X_train, y_train)`
	`3 y_pred = nb_model.predict(X_test)`

Model 3 : Support Vector Machines (SVM)	Model 4 : Neural Network

`[21] 1 from sklearn.svm import SVC`
 	
`[22] 1 svm_model = SVC()`
`2 svm_model = svm_model.fit(X_train, y_train)`
`3 y_pred = svm_model.predict(X_test)`
	`[29] 1 import tensorflow as tf`
	
	`[30] 1 def plot_history(history):`
	`2 fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(10, 4))`
	`3 ax1.plot(history.history['loss'], label='loss')`
	`4 ax1.plot(history.history['val_loss'], label='val_loss')`
	`5 ax1.set_xlabel('Epoch')`
	`6 ax1.set_ylabel('Binary crossentropy')`
	`7 ax1.grid(True)`
	`8`
	`9 ax2.plot(history.history['accuracy'], label='accuracy')`
	`10 ax2.plot(history.history['val_accuracy'], label='val_accuracy')`
	`11 ax2.set_xlabel('Epoch')`
	`12 ax2.set_ylabel('Accuracy')`
	`13 ax2.grid(True)`
	`14`
	`15 plt.show()`
	
	`[31] 1 def train_model(X_train, y_train, num_nodes, dropout_prob, lr, batch_size, epochs):`
	`2 nn_model = tf.keras.Sequential([`
	`3 tf.keras.layers.Dense(num_nodes, activation='relu', input_shape=(9,)),`
	`4 tf.keras.layers.Dropout(dropout_prob),`
	`5 tf.keras.layers.Dense(num_nodes, activation='relu'),`
	`6 tf.keras.layers.Dropout(dropout_prob),`
	`7 tf.keras.layers.Dense(1, activation='sigmoid')`
	`8 ])`
	`9`
	`10 nn_model.compile(optimizer=tf.keras.optimizers.Adam(lr), loss='binary_crossentropy',`
	`11 metrics=['accuracy'])`
	`12 history = nn_model.fit(`
	`13 X_train, y_train, epochs=epochs, batch_size=batch_size, validation_split=0.2, verbose=0`
	`14 )`
	`15`
	`16 return nn_model, history`
	
	`[43] 1 least_val_loss = float('inf')`
	`2 least_loss_model = None`
	`3 epochs=100`
	`4 for num_nodes in [16, 32, 64]:`
	`5 for dropout_prob in[0, 0.2]:`
	`6 for lr in [0.01, 0.005, 0.001]:`
	`7 for batch_size in [32, 64, 128]:`
	`8 print(f"{num_nodes} nodes, dropout {dropout_prob}, lr {lr}, batch size {batch_size}")`
	`9 model, history = train_model(X_train, y_train, num_nodes, dropout_prob, lr, batch_siz`
	`10 plot_history(history)`
	`11 val_loss = model.evaluate(X_valid, y_valid)[0]`
	`12 if val_loss < least_val_loss:`
	`13 least_val_loss = val_loss`
	`14 least_loss_model = model`

2.5 Evaluation

In the evaluation phase, the data mining process's modeled outcomes from the previous step are thoroughly assessed. The assessment is performed with the intention of adjusting the acquired model to align with the goals defined in the first stage. indicating that it is already clean and ready to be processed by the model.

Based on Table 3 below, not only the accuracy, but we can also see the findings about precision, recall, and f-1 score of each model performance on prediction.

Table 3. Evaluation for Each Model

Accuracy Score

Model 1 : K-Nearest Neighbor

```
[13]  1 print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.81	0.89	0.85	1998
1	0.80	0.68	0.74	1278
accuracy			0.81	3276
macro avg	0.81	0.79	0.79	3276
weighted avg	0.81	0.81	0.81	3276

Model 2 : Naïve Baiyes

```
1 print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.64	0.88	0.74	1998
1	0.56	0.23	0.32	1278
accuracy			0.63	3276
macro avg	0.60	0.56	0.53	3276
weighted avg	0.61	0.63	0.58	3276

Model 3 : Support Vector Machines (SVM)

```
[23]  1 print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.70	0.96	0.81	1998
1	0.85	0.37	0.51	1278
accuracy			0.73	3276
macro avg	0.77	0.66	0.66	3276
weighted avg	0.76	0.73	0.70	3276

Model 4 : Neural Network

```
[46]  1 print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.84	0.94	0.89	1998
1	0.89	0.71	0.79	1278
accuracy			0.85	3276
macro avg	0.86	0.83	0.84	3276
weighted avg	0.86	0.85	0.85	3276

2.6 Deployment

The term deployment refers to the phase where the results of our research are shared. The step that has been taken involves report preparation and presentation of the knowledge obtained from the assessment of the data mining procedure in a format that is easily understood by others.

C. Result and Discussion

3.1 Accuracy Score

Table 4. Accuracy Score of Each Model

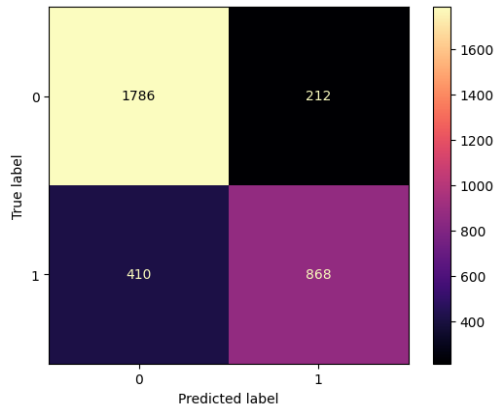
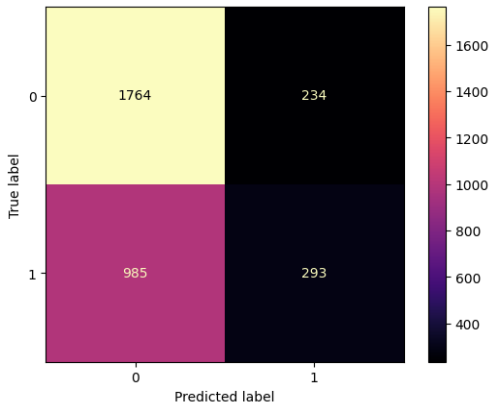
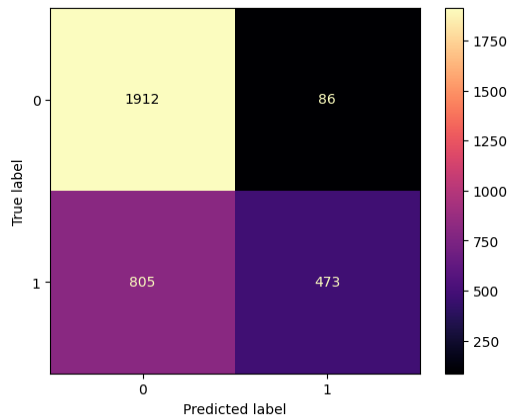
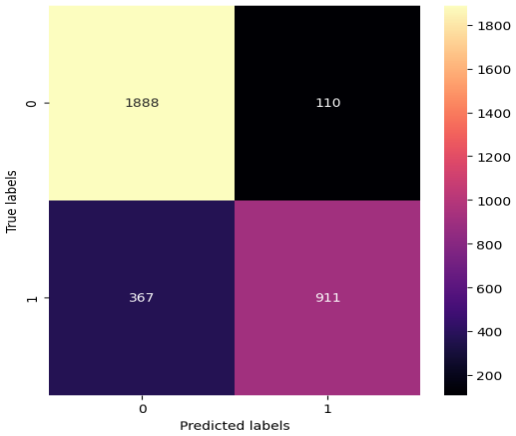
Model	Accuracy
KNN	0.81
Naïve Bayes	0.63
SVM	0.73
Neural Network	0.85

As shown in the Table 4, the classification report provides us the accuracy score for K-Nearest Neighbors 0.81, Naïve Bayes 0.63, Support Vector Machine 0.73, and Neural Network 0.85. It shows that each model has a significant difference of accuracy score despite using the same training, validation, and testing data.

3.2 Confusion Matrix

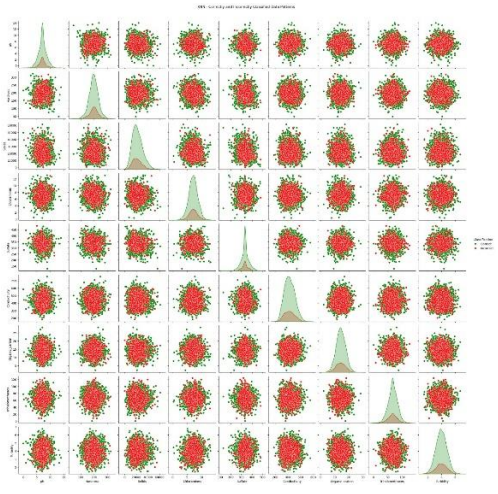
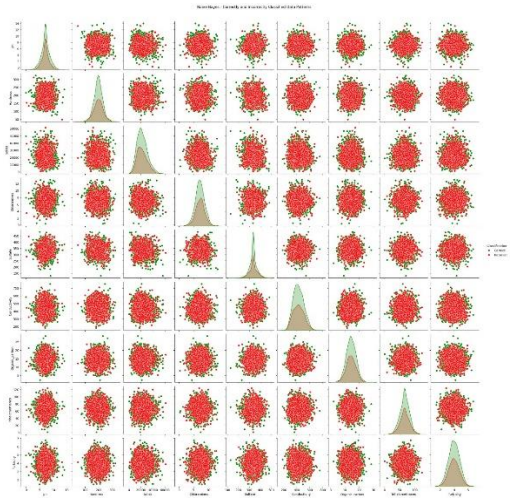
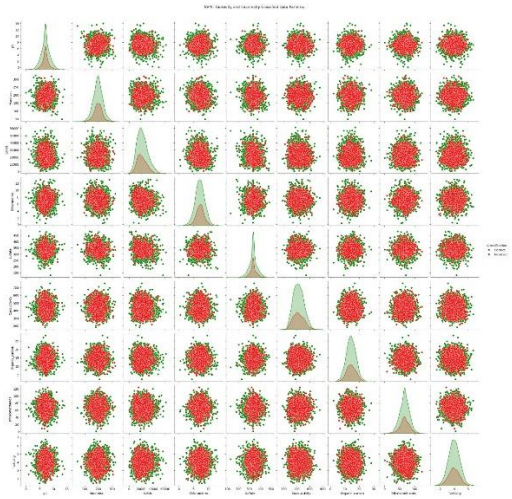
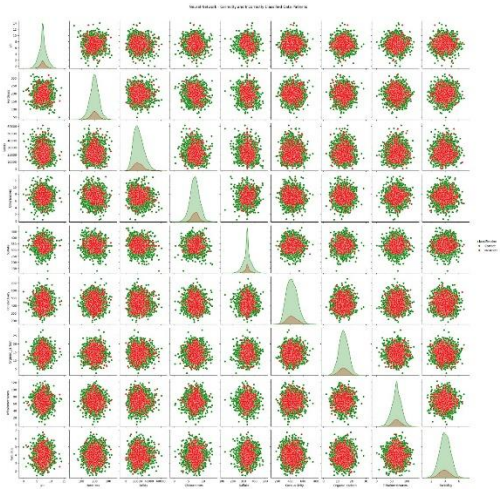
According to Table 5 below, a lighter color indicates a greater number. On the contrary, a deeper tint indicates a lower number. The first model, which is KNN, has a total of 2654 data that correctly classified and 622 data that incorrectly classified. For the second model, which is Naïve Bayes, correctly classified 2057 and incorrectly classified 1219 data. For the last model, SVM correctly classified 2385 and the rest of the 891 data were incorrectly classified. For the last model, which is Neural Network, has the most amount of the correctly classified with 2799 data and 477 data were incorrectly classified.

Table 5. Confusion matrix of each model

Confusion Matrix	
Model 1 : K-Nearest Neighbor	Model 2 : Naïve Baiyes
	
Model 3 : Support Vector Machines (SVM)	Model 4 : Neural Network
	

3.3 Data Pattern

Table 6. Data Pattern

Accuracy Score	
Model 1 : K-Nearest Neighbor	Model 2 : Naïve Baiyes
	
Model 3 : Support Vector Machines (SVM)	Model 4 : Neural Network
	

By evaluating the pattern of correctly classified and incorrectly classified for different algorithms, we can determine which algorithm pattern has the greatest accuracy score. Although the changes in classification patterns may not be obvious at first look, they are critical for fine-tuning the algorithm selection process. This comparison research might serve as a starting point for establishing an advanced IT-based water quality testing programme, potentially replacing old manual testing methods used in chemistry labs. Furthermore, the automation and digitalization of water quality monitoring may result in more reliable and repeatable results, minimising human error and allowing for broader, more frequent testing across several sites.

D. Conclusion

To classify water potability, this is done by applying machine learning to enhance the efficiency and accuracy of decision making in the environmental and public health sectors. This machine learning will provide a reliable solution to categorize water accurately. The Cross-Industry Standard for Data Mining (CRISP-DM) method was used as a research stage to assess water quality data.

This research aims to find out the results and compare the level of accuracy of the research methods used such as K-Nearest Neighbors, Naïve Bayes, Support Vector Machine, and Neural Network. Based on classification reports from data mining research results, the method that produces the highest level of accuracy is the Neural Network model, namely 0.85%. Other results show that the K-Nearest Neighbors method has an accuracy rate of 0.81%, the Support Vector Machine method has an accuracy rate of 0.73% and the Naive Bayes method produces the lowest accuracy rate, namely 0.63%. For further research, it is recommended to use other algorithms or hyperparameter tuning to obtain maximum accuracy results. Apart from that, researchers also suggest conducting experiments using more data to classify water potability using several different methods.

E. Acknowledgment

First and foremost, we express our heartfelt gratitude to our supervisor, Mr. Muhammad Ihsan Jambak, for his direction, encouragement, insightful feedback, and assistance with advising the progress of our research and improving its quality. Appreciation is also expressed to the researcher's colleagues, who played critical roles in supporting researchers in achieving this point.

F. References

[1] M. R. Amonovich and N. S. Ahror o'g'li, "Importance of Water For Living Organisms And National Economy, Physical And Chemical Methods Of Wastewater Treatment," *American Journal of Research in Humanities and Social Sciences*, vol. 9, pp. 7–13, 2023.

[2] S. Sahoo and S. Goswami, "Theoretical framework for assessing the economic and environmental impact of water pollution: A detailed study on sustainable development of India," *Journal of Future Sustainability*, vol. 4, no. 1, pp. 23–34, 2024.

[3] P. Babuji, S. Thirumalaisamy, K. Duraisamy, and G. Periyasamy, "Human health risks due to exposure to water pollution: a review. Water 15: 2532." 2023.

[4] S. Suliman, "Implementasi Data Mining Terhadap Prestasi Belajar Mahasiswa Berdasarkan Pergaulan dan Sosial Ekonomi Dengan Algoritma K-Means Clustering," *Jurnal Sistem Informasi dan Sistem Komputer*, vol. 6, no. 1, pp. 1–11, 2021.

[5] C. Zai, "Implementasi Data Mining Sebagai Pengolahan Data," *Jurnal Portal Data*, vol. 2, no. 3, 2022.

[6] A. P. Fadillah, "Penerapan Metode CRISP-DM untuk Prediksi Kelulusan Studi Mahasiswa Menempuh Mata Kuliah (Studi Kasus Universitas XYZ)," *Jurnal Teknik Informatika Dan Sistem Informasi*, vol. 1, no. 3, 2015.

[7] Y. Christian and K. O. Y. R. Qi, "Penerapan K-Means pada Segmentasi Pasar untuk Riset Pemasaran pada Startup Early Stage dengan Menggunakan CRISP-DM," *JURIKOM (Jurnal Riset Komputer)*, vol. 9, no. 4, pp. 966–973, 2022.
[8] N. Mirantika, "Penerapan Algoritma K-Means Clustering Untuk Pengelompokan Penyebaran Covid-19 di Provinsi Jawa Barat," *Nuansa Informatika*, vol. 15, no. 2, pp. 92–98, 2021.
[9] R. L. Bárta *et al.*, "Public water quality in surface and groundwater collection systems in Brazil," *CONTRIBUCIONES A LAS CIENCIAS SOCIALES*, vol. 17, no. 1, pp. 141–161, 2024.
[10] A. Mondal and S. S. Dubey, "Machine Learning-based Water Potability Prediction: Model Evaluation, and Hyperparameter Optimization".
[11] A. Tangkelayuk and E. Mailoa, "The Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes, dan Decision Tree. JATISI (Jurnal Teknik Informatika Dan Sistem Informasi), 9 (2), 1109–1119." 2022.
[12] A. Pambudi, "Penerapan Crisp-Dm Menggunakan Mlr K-Fold Pada Data Saham Pt. Telkom Indonesia (Persero) Tbk (Tlkm)(Studi Kasus: Bursa Efek Indonesia Tahun 2015-2022)," *Jurnal Data Mining dan Sistem Informasi*, vol. 4, no. 1, pp. 1–14, 2023.
[13] I. Budiman, T. Prahasto, and Y. Christyono, "Data Clustering menggunakan metodologi Crisp-DM untuk pengenalan pola proporsi pelaksanaan tridharma," in *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, 2012.