Comparative Study of Machine Learning Techniques For Water Potability Prediction

Ying Ying Lai
School of Information Technology
Monash University Malaysia
ylai0018@student.monash.edu

***Abstract*—In recent years, Internet Of Things (IoT) has become increasingly popular due to its smart application in various fields. Modern water treatment plant has been using smart sensors for real-time water quality monitoring and prediction to ensure the water is appropriate for its intended purpose. In this paper, the performance of various machine learning (ML) algorithms on the task of predicting water potability are compared and discussed. The techniques chosen including but not limited to Logistic Regression, Random Forest, K-Nearest Neighbours (KNN), Support Vector Machine (SVM) and Artificial Neural Network(ANN). The accuracy of these models are determined using evaluation metrics such as the accuracy, precision, recall and f1 score. The results shown that the SVM model performed the best with the highest accuracy and lesser training time as compared to the more complex classification algorithms. It is able to achieve the accuracy of 98.16%, followed by Random Forest with an accuracy of 97.53%.**

***Index Terms*—water quality prediction, binary classification, SVM, Random Forest, logistic regression, neural network, IoT**

I. Introduction

Water is an essential resource that is needed to sustain life on earth. In recent years, Malaysia is experiencing an increased demand for water as its supply has changed from one of relative abundance to one of scarcity [1]. In fact, this is a global issue that is already affecting every continent. Water is needed for agriculture, energy production, recreation, and manufacturing [2] and a lack of access to clean water can threatens human well-being, economic productivity as well as the ecological communities. To make matters worse, abrupt climate change coupled with massive population growth and rapid urban development further aggravate the issue, pushing governments to actively look for innovative and collaborative solutions to tackle this issue.

In an effort to achieve sustainable water use and management, governments and environmental organizations have been exploring the potential reuse of wastewater post-treatment for different end use. An important aspect in determining the suitability of water usage is the monitoring and prediction of water quality by analysing its chemical, physical and biological properties. The integration of IoT allows these information to be collected over an extended period of time and often the data exist in the form of multivariate time-series datasets [3]. Artificial Intelligence (AI) offers significant opportunities to help improve the classification and prediction of water quality by developing a dependable approach for forecasting water quality as accurately as possible [4].

This study focuses on predicting the potability of water based on 20 water parameters, which includes but not limited to the aluminium, ammonia, copper and selenium content of the water, using various machine learning techniques. The models are fine-tuned to achieve its optimal performance before evaluating their accuracy.

Section II of this paper provides some background of the project topic and literature review of related works. Section III dives into the research methodology, where I will be explaining on the data collection methods, pre-processing steps, and classification methods. Then, the performance of the models will be discussed in section IV, along with their respective test scores and visual representations of the evaluation metrics. Lastly, Section V concludes the paper with a brief summary of this research findings.

II. Background

A. Water Quality Monitoring

Depending on the designated usage, water specifications for different applications possess a set of standards that must be met in order to be considered safe. These guidelines are based on scientific research and epidemiological findings, and as such provide guidance for making risk management decisions related to the protection of public health and the preservation of the environment [5]. For example, water bodies such as rivers, lakes, and streams have specific quality standards that indicate their quality to ensure that the amount of pollutants presence is within the tolerance limit of water species [6]. On the other hand, water must not be too saline nor contain toxic materials to be qualified for irrigation purposes to protect the plant and ecosystems [6]. Hence, continuous monitoring, evaluation and prediction of water quality are necessary.

B. Traditional Water Sampling and Testing

Traditional wastewater quality monitoring methods involve manual collection of water samples at different locations, followed by laboratory analytical techniques in order to characterize the water quality [7]. Besides being labour intensive, laboratory analysis is time consuming and high costs is needed for labour, maintenance and operations of equipment. Considering all water quality parameters is also unrealistic because it is not only expensive and technically difficult but also fails to deal with the variability in water quality [8]. Besides that, this method does not offer real-time water quality detection, hence

responsible authorities will not be able to take immediate action if the water is unsafe for usage.

C. IoT based Water Quality Monitoring System

With the advancement in technology, a real-time, cost-effective water quality monitoring system can be easily deployed. This is achieved by embedding sensors that sent its sensor readings to the server for monitoring and analytical purposes. In recent years, the vision of the Internet of Things (IoT) augmented with advances in software technologies, such as service-oriented architecture (SOA), software as a service (SaaS), cloud computing, and others, has stimulated the development of smart water quality monitoring systems [7].

D. Machine Learning in Water Quality Prediction

Based on Zhu et al. [8], the integration of sensors in water treatment plant has multiple benefits such as continuous water quality monitoring and prediction, pollutant source tracking, and pollutant concentration estimation, water resource allocation and water treatment technology optimization. These sensors generate a high volume of multivariate data. Manual water quality analysis are no longer preferred due to human's inability to handle such complex data. Besides being more error-prone, this process is both cost- and time-consuming.

These limitations are overcome by machine-learning based models, which is widely known to increase productivity and efficiency in many fields. It's ability to correctly interpret external data, to learn from such data, and to use those information to achieve specific goals and task through flexible adaptation [9] allows the model to achieve better performance and accuracy in evaluation and analysis [10].

Machine learning can be divided into two main classes, which is supervised and unsupervised learning [11]. In the case of water quality prediction, we will be looking into supervised learning algorithms since all data in the dataset are labeled. The aim is to study the relationships or correlation between the paired labels and generate a predictive model. The model is then used to predict the result based on a set of input data. Data classification and regression apply the concept of supervised learning to generate a model, while utilising algorithms such as naive Bayes, logistic regression, decision tree, random forest, support vector machine, k-nearest neighbour and support vector machine [8].

Unlike manual water sampling, ML models are able to consistently provide more accurate evaluation results despite the complexity of the dataset. The feasibility and effectiveness of these models demonstrates its potential to be used as a tool in water quality evaluation [8] since the evaluation process often involves large multivariate time-series dataset. Multiple parameters will need to be considered when determining water quality because a slight change in the water parameters level can potentially affect the suitability of the water for a certain usage. Lastly, the use of ML models over manual water sampling and analysis also has the benefits of reducing labour cost, less error-prone and more time-efficient.

E. Related Works

In recent years, many research has been carried out to study how different types of machine learning (ML) classification algorithms can affect the accuracy of the model in water quality prediction. Zai C. et. al [12] found that Decision Tree Algorithm performed decently well in evaluating the water quality index. It has a R-squared error R^2 score of 72% and Root Mean Square Error (RMSE) score of 0.41, which denotes excellent performances in the proposed models for forecasting IWQ showing to other models that have been proposed in the literature.

Nasir N. et. al [4] investigated the performance of various machine learning algorithms on predicting the water quality index (WQI) using seven different water parameters. The classifiers that are implemented include Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), CATBoost, XGBoost, and Multilayer Perceptron (MLP).Results show that the CATBoost model has the highest accuracy of 94.51%. The accuracy of all classifier reached 100% when implemented along with stacking ensemble models.

A similar study is carried out by Poudel D. [13] to study water potability using a dataset with different water parameters. The research compares the efficiency of Logistic Regression (LR), K-Nearest Neighbours (KNN) and Random Forest (RF) and Artificial Neural Network (ANN) in predicting dataset that are statistically imputed.

Zhou J. et. al [3] proposed a model that uses IGRA for feature selection and LSTM for water quality prediction. IGRA calculates the similarity and proximity by relative area change ratio, whereas LSTM is faster at converging to the optimal solution, especially when dealing with time sequence prediction problems. This model takes full advantage of the multivariate correlation and time sequence of water quality information, and is ideal for multi-class classification problem.

Another research by Aldhyani et. al [6] studied the performance of various machine learning algorithms on different water types such as surface water, wastewater and drinking water. The performances are evaluated based on the recognition rate, training time and robustness, and it is revealed that ANN has the highest recognition rate and and SVM is very robust to noise. These methods are widely applied in large dimension to dynamically monitor the quality and safety of drinking water in real-time since they have the best training performances for predicting the potability of water among the other algorithms [14].

III. Methodology

The aim of this paper is to study the accuracy of 9 different binary classification algorithms in predicting the potability of water based on 20 water chemical properties. The models were implemented using Logistic Regression (LR), Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Extreme Gradient Boosting (XGBoost), and the Multi-layer Perceptron (MLP) algorithms. The following sections include

discussions on the data pre-processing steps, classification techniques and evaluation metrics used to compare the performance of the models.

A. Dataset Selection

The dataset chosen provides information on the safety of water for drinking purposes. It is a public dataset that can be downloaded from Kaggle website [15]. It contains a total of 7998 rows of water sample and the potability of each sample is determined by 20 continuous features. The classification feature is potability with values of either 0 or 1, where 0 denotes non-potable, and 1 denotes potable. The features and their respective datatype are shown in Table I.

Table I
WATER PARAMETERS IN DATASET (RAW)

Parameter	Datatype	Parameter	Datatype
aluminium	float64	viruses	float64
ammonia	object	lead	float64
arsenic	float64	nitrates	float64
barium	float64	nitrites	float64
cadmium	float64	mercury	float64
chloramine	float64	perchlorate	float64
chromium	float64	radium	float64
copper	float64	selenium	float64
fluoride	float64	silver	float64
bacteria	float64	uranium	float64

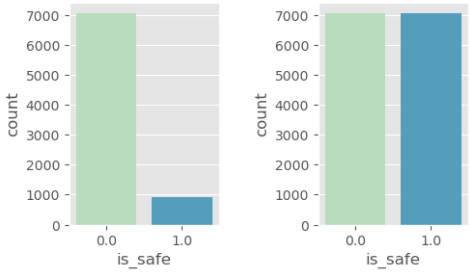
B. Data Pre-processing

Real-world data often contains missing values, outliers and misleading data, and sometimes there is no guarantee to the reliability and credibility of the data sources. These issues may drastically affect the performance and accuracy of the machine learning model since the training outcome depends heavily on the dataset [16]. Data pre-processing removes irrelevant entries by cleaning (i.e., removing outliers) and labeling the data [4]. Below are the steps taken during pre-processing:

1) Categorical Variable: The column for the water parameter, *ammonia*, and potability, *is_safe*, have a datatype of object. Since the dataset contains information regarding the potability of water samples and their corresponding features values, the column for potability should be stored as either 0 or 1, and the *ammonia* column is a continuous variable that stores a range of numeric. Therefore, both the column were casted as *float64*.

2) Missing Values: In this dataset, only 3 out of 7998 water samples contain *Null* values, which is less than 0.01% of the entire dataset. Therefore, I decided to remove the instances of water samples instead since removing them will not cause a significant change in the training result.

3) Imbalanced dataset: Figures 1a shows that the dataset has a highly imbalanced data distribution, where the number of water samples for potable water is significantly lower than that of the non-potable water. To overcome this, oversampling technique is applied to the minority class to generate new samples, aiming to achieve a well-balanced class distribution [17]. The synthetic minority oversampling technique (SMOTE) algorithm is chosen to perform oversampling due to its ability to prevent over-fitting by creating samples from the interpolation between the existing samples and their neighbors [4].



(a) Before oversampling (b) After oversampling

Figure 1. Potability Class Distribution

4) Data-Split: The dataset is divided into training and testing sets with a proportion of 70% and 30% respectively, resulting in 9917 training data and 4251 testing data.

5) Normalisation: Data Normalisation is done by scaling and translating each feature in the training set individually to a range between the minimum and maximum values [18]. The transformation is given by:

$$\begin{aligned} X\_std =& (X - X.min(axis = 0))/ \\ & (X.max(axis = 0) - X.min(axis = 0)) \\ X\_scaled =& X\_std * (max - min) + min \end{aligned}$$

C. Classification

A brief description of each classification technique is presented in the section below.

1) Naive Bayes: A probabilistic classification technique used to determine is a data point falls into a particular class. It is based on the Bayes Theorem, which is used to calculate conditional probability. Many Naive Bayes are available, however, the Gaussian Naive Bayes is used in this study and the likelihood of each feature is assumed to be:

$$P(x_i|y) = \frac{1}{\sqrt{(2\pi\sigma_y^2}\exp(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2})}$$

2) Logistic Regression: A standard classification classifier that uses the probabilistic statistics of the data to predict binomial outcomes. The standard logistic function is an S-shaped curve, which is given by the equation:

$$\frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}$$

3) Decision Tree: Decision Tree is used for both classification and regression. It first learn to generate a decision tree from a set of training samples that have been classified, and the tree obtained is for used as a model to classify other unclassified data [19]. A decision tree starts with internal nodes representing a test of each features, moving down branches which represents the output, and the leaves nodes contain

Table II
SAMPLE DATASET

aluminium	ammonia	arsenic	barium	cadmium	chloramine	chromium	copper	flouride	bacteria	viruses
1.65	9.08	0.04	2.85	0.01	0.35	0.83	0.17	0.05	0.2	0.0
2.32	21.16	0.01	3.31	0.0	5.28	0.68	0.66	0.9	0.65	0.65
1.01	14.02	0.04	0.58	0.01	4.24	0.53	0.02	0.99	0.05	0.0
1.36	11.33	0.04	2.96	0.0	7.23	0.03	1.66	1.08	0.71	0.71
0.92	24.33	0.03	0.2	0.01	2.67	0.69	0.57	0.61	0.13	0.0

lead	nitrates	nitrites	mercury	perchlorate	radium	selenium	silver	uranium	is_safe
0.05	16.08	1.13	0.01	37.75	6.78	0.08	0.34	0.02	1.0
0.1	2.01	1.93	0.0	32.26	3.21	0.08	0.27	0.05	1.0
0.08	14.16	1.11	0.01	50.28	7.07	0.07	0.44	0.01	0.0
0.02	1.41	1.29	0.0	9.12	1.72	0.02	0.45	0.05	1.0
0.12	6.74	1.11	0.0	16.9	2.41	0.02	0.06	0.02	1.0

the resulting decisions. Due to the presence of variance, the downside to this method is that they tend to cause overfitting. Besides that, decision tree can become costly especially when the data set is large [20].

4) Random Forest Classifier: The random forest is a classification algorithm that aims to overcome multiple weaknesses of decision trees, such as (1) Prevent overfitting by reducing variance (2) Less sensitive to outliers during training (3) Does not require pre-preprocessing and pruning [20]. Unlike Decision Trees, Random Forest makes a conclusion by taking the average of all the decision trees, which produces predictions with high accuracy.

5) K Nearest Neighbours (KNN): KNN assumes similar data points appear in close proximity. It divides the data into classes based on the closeness between data points by calculating the distance between them. The distance is calculated using Euclidean distance and the equation is given by:

$$D(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + ... + (a_n - b_n)^2}$$

6) Support Vector Machine (SVM): SVM is a popular and robust technique for classification and regression that is widely used. SVM constructs a hyperplane that best separates the data into two or more classes. Essentially, the algorithm tries to find the optimal separation lines and achieve maximum separation of the classes. The larger the distance of the training data point from the hyperplane, the more likely that the data point is correctly classified. Hence, the goal is to have a hyperplane with the largest distance to the training data points.

7) Extreme gradient boosting (XGBoost): XGBoost improves the model generalization capabilities of Gradient Boosting by using L1 and L2 regularization to prevent overfitting. XGBoost is used to solve classification problems by combining many shallow decision trees, which results in high prediction accuracy and minimizes the objective function [21]. It is parallelizable across clusters, which minimizes the training time.

8) Multi-layer Perceptron (MLP): One of the most widely used artificial neural network (ANN) algorithms for classification and regression. There are at least three or more layers in MLP, which consist of an input layer receiving inputs from the environment, an output layer generating the network's response, and one or more hidden layers [22].

MLP is composed of multiple layers of neurons, which are connected between layers via weights. The training process includes adjusting the weight on each node of each layer [23]. The goal is to find an optimal set of weights that is capable of producing accurate output based on the input given. Due to this reason, a large number of training sets and training cycles is preferable.

D. Evaluation Metrics

1) Log Loss: The Log Loss, also known as Cross Entropy Loss, is used to evaluate the performance of classifier using a probability value between 0 and 1. The lower the value of log loss, the higher the accuracy of the model. Hence, a lower value is desirable. The equation for calculating log loss can be given by:

$$L_{log}(y,p) = -(ylog(p) + (1-y)log(1-p))$$

2) Confusion Matrix: In binary classification, confusion matrix provides the count of true negatives (tn), false negatives (fn), true positives (tp) and false positives (fp), which allow us to calculate the accuracy (1), precision (2), recall (3) and F1-score (4). Equations given by:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (1)$$

$$Precision = \frac{tp}{tp + fp} \quad (2)$$

$$Recall = \frac{tp}{tp + fn} \quad (3)$$

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

a) Accuracy: The percentage of correct predictions out of all the predictions.

b) Precision: The percentage truly positive predictions out of all the predicted positive predictions.

c) Recall: The percentage of truly positive predictions out of all the actual positive predictions (True Positive Rate)

d) F1-score: A better reflection of the overall accuracy by taking into account both the false positive and false negatives.

Table III
RESULTS OF THE PROPOSED CLASSIFIERS

	Naive Bayes	Logistic Regression	Decision Tree	Random Forest	KNN	SVM	XGBoost	MLP
TP	1680	1680	2077	2114	2143	2095	2129	2082
TN	1651	1665	1984	2018	1503	1934	2034	1980
FP	448	434	115	81	596	165	65	119
FN	472	472	75	38	9	57	23	70
Log Loss	7.4750	7.3612	1.5437	0.9669	4.9157	1.8037	**0.7150**	1.5356
Accuracy	0.7836	0.7869	0.9553	0.9720	0.8577	0.9478	**0.9793**	0.9555
Precision	0.7895	0.7947	0.9475	0.9631	0.7824	0.9270	**0.9704**	0.9459
Recall	0.7807	0.7807	0.9651	0.9823	**0.9958**	0.9735	0.9893	0.9675
F1-score	0.7850	0.7876	0.9563	0.9726	0.8763	0.9497	**0.9798**	0.9566

IV. RESULT AND DISCUSSION

The models are compared using their F1-score and log loss since F1-score takes into the account both precision and recall (ie. the false positives and false negatives), and log loss determines how close the predictions are to the actual value. Accuracy will not be a good representation of the performance since the original dataset is imbalanced. Besides that, recall is not reliable as well since the score is affected by oversampling the original dataset. This is shown by the recall score of KNN, which is the highest among all models, even though it is evident that it does not perform as well as the other models. Recall is equivalent to the total positive rate, which means that it is able to identity water that are potable 99.58% of the time. However, this value may not be reliable since the minority class (ie. the number of potable water samples) is oversampled, which may results in biased data points. KNN algorithm tend to group data points in close proximity into a single class, which increases the percentage of correct positive predictions.

A. Model Evaluation

All models are implemented using the classification algorithms provided in the Scikit-learn machine learning library. All hyper parameters were unmodified and left at default. The trained models are then evaluated using the testing dataset. The log loss, accuracy, precision, recall and F1-score is shown in Table III.

XGBoost has the best performance by having the highest F1-score of 0.9798 and lowest log loss of 0.7150. Random Forest performs almost as well as XGBoost with a F1-score of 0.9726, and a log loss of 0.9669. This is followed by MLP, Decision Tree and SVM, with close F1-score of 0.9566, 0.9563 and 0.9497, and log loss of 1.5356, 1.5437 and 1.8037 respectively. The remaining 3 models (KNN, Logistic Regression and Naive Bayes) did not perform as well compared to the other models, having log loss between 4.5 to 7.5 and f1-score between 0.78 to 0.88.

B. Parameter Tuning

A randomized search on hyper parameters with 10-fold cross validation is used to select the best parameter for training. This is to ensure the final model is optimised for the best performance. The section below present the value of the tuned hyper parameters for each model.

Table IV
TUNED HYPER PARAMETERS FOR LOGISTIC REGRESSION

Parameters	Values
C	10
class_weight	None
dual	False
fit_intercept	True
intercept_scaling	1
l1_ratio	None
max_iter	100
multi_class	auto
n_jobs	None
penalty	l1
random_state	None
solver	liblinear
tol	0.0001
verbose	0
warm_start	False

Table V
TUNED HYPER PARAMETERS FOR NAIVE BAYES

Parameters	Values
priors	None
var_smoothing	0.005335599

Table VI
TUNED HYPER PARAMETERS FOR DECISION TREE

Parameters	Values
ccp_alpha	0
class_weight	None
criterion	log_loss
max_depth	None
max_features	None
max_leaf_nodes	None
min_impurity_decrease	0
min_samples_leaf	1
min_samples_split	2
min_weight_fraction	0
random_state	None
splitter	best

Table VII
TUNED HYPER PARAMETERS FOR RANDOM FOREST

Parameters	Values
bootstrap	False
ccp_alpha	0
class_weight	None
criterion	entropy
max_depth	None
max_features	sqrt
max_leaf_nodes	None
min_impurity_decrease	0
min_samples_leaf	1
min_samples_split	2
min_weight_fraction	0
n_estimators	100
n_jobs	None
oob_score	False
random_state	None
verbose	0
warm_start	False

Table VIII
TUNED HYPER PARAMETERS FOR KNN

Parameters	Values
algorithm	auto
leaf_size	30
metric	manhattan
metric_params	None
n_jobs	None
n_neighbors	30
p	2
weights	uniform

Table IX
TUNED HYPER PARAMETERS FOR SVM

Parameters	Values
C	50
break_ties	False
cache_size	200
class_weight	None
coef0	0
decision_function_shape	ovr
degree	3
gamma	scale
kernel	rbf
max_iter	-1
probability	False
random_state	None
shrinking	True
tol	0.001
verbose	False

C. Make Predictions

The final performance results were obtained by predicting the outcomes of the test set using the optimised version of the models. With optimization, it is evident that all algorithms are able to improve their performance. XGBoost still has the best performance with log loss of 0.6500, which is 0.065 lower than the unoptimised version. Random Forest is able to maintain its high accuracy in prediction. The optimized SVM is able to perform better than MLP and Decision Tree by having a log loss of 1.1456, which is 31% lower, whereas the F1-score increases from 0.9497 to 0.9681. The scores of the other evaluation metrics are shown in Table XII.

Table X
TUNED HYPER PARAMETERS FOR XGBOOST

Parameters	Values	Parameters	Values
base_score	0.5	max_leaves	0
booster	gbtree	min_child_weight	1
callbacks	None	missing	nan
colsample_bylevel	1	monotone_constraints	()
colsample_bynode	1	n_estimators	1000
colsample_bytree	1	n_jobs	0
early_stopping_rounds	None	num_parallel_tree	1
enable_categorical	False	objective	binary:logistic
eta	0.2	predictor	auto
eval_metric	logloss	random_state	0
gamma	0	reg_alpha	0
gpu_id	-1	reg_lambda	1
grow_policy	depthwise	sampling_method	uniform
importance_type	None	scale_pos_weight	1
learning_rate	0.2	subsample	1
max_bin	256	tree_method	exact
max_cat_to_onehot	4	use_label_encoder	False
max_delta_step	0	validate_parameters	1
max_depth	6	verbosity	None

Table XI
TUNED HYPER PARAMETERS FOR MLP

Parameters	Values	Parameters	Values
activation	relu	momentum	0.9
alpha	0.001	n_iter_no_change	10
batch_size	auto	nesterovs_momentum	True
beta_1	0.9	power_t	0.5
beta_2	0.999	random_state	None
early_stopping	False	shuffle	True
epsilon	1e-08	solver	adam
hidden_layer_sizes	(100,)	tol	0.0001
learning_rate	constant	validation_fraction	0.1
learning_rate_init	0.001	verbose	False
max_fun	15000	warm_start	False
max_iter	500		

V. CONCLUSION

This study compared the performance of various binary classification techniques in predicting the potability of water based on 20 features. The classifiers implemented are Naive Bayes, Logistic Regression, Decision Trees, Random Forest, K Nearest Neighbours, SVM, XGBoost and MLP. According to the estimates, The results suggest that XGBoost overall has the best performance in terms of log loss and F1-score, followed by Random Forest Classifier. It is also shown that SVM has the potential to achieve high accuracy after optimising the hyper parameters, even outperforming the Multi-layer Perceptron (MLP), an artificial neural network. In conclusion, Decision Tree, Random Forest, SVM, XGBoost and MLP are reliable machine learning algorithms that are able to predict water potability with high accuracy.

Table XII
PERFORMANCE OF OPTIMISED CLASSIFIERS

	Naive Bayes	Logistic Regression	Decision Tree	Random Forest	KNN	SVM	XGBoost	MLP
TP	1701	1701	2090	2119	2138	2141	2132	2112
TN	1634	1663	1980	2029	1472	1969	2039	1990
FP	465	436	119	70	627	130	60	109
FN	451	451	62	33	14	11	20	40
Log Loss	7.4425	7.2068	1.4706	0.8369	5.2082	1.1456	**0.6500**	1.2106
Accuracy	0.7845	0.7913	0.9574	0.9758	0.8492	0.9668	**0.9812**	0.9649
Precision	0.7853	0.7960	0.9461	0.9680	0.7732	0.9428	**0.9726**	0.9509
Recall	0.7904	0.7904	0.9712	0.9847	0.9935	**0.9949**	0.9907	0.9814
F1-score	0.7879	0.7932	0.9585	0.9763	0.8696	0.9681	**0.9816**	0.9659

REFERENCES

[1] Ferdoushi Ahmed and Chamhuri Siwar. "Concepts, Dimensions and Elements of Water Security". In: *Pakistan Journal of Nutrition* 13.5 (May 2014), pp. 281–286. DOI: 10.3923/pjn.2014.281.286.

[2] Zuraini Anang et al. "Factors Affecting Water Demand: Macro Evidence in Malaysia". In: *Jurnal Ekonomi Malaysia* 53.1 (2019), pp. 17–25.

[3] Jian Zhou et al. "Water quality prediction method based on IGRA and LSTM". In: *Water* 10.9 (2018), p. 1148.

[4] Nida Nasir et al. "Water quality classification using machine learning algorithms". In: *Journal of Water Process Engineering* 48 (2022), p. 102920.

[5] G Fred Lee, R Anne Jones, and Brooks W Newbry. "Water quality standards and water quality". In: *Journal (Water Pollution Control Federation)* (1982), pp. 1131–1138.

[6] Theyazn HH Aldhyani et al. "Water quality prediction using artificial intelligence algorithms". In: *Applied Bionics and Biomechanics* 2020 (2020).

[7] Ramón Martínez et al. "On the Use of an IoT Integrated System for Water Quality Monitoring and Management in Wastewater Treatment Plants". In: *Water* 12.4 (2020). ISSN: 2073-4441. URL: https://www.mdpi.com/2073-4441/12/4/1096.

[8] Mengyuan Zhu et al. "A review of the application of machine learning in water quality evaluation". In: *Eco-Environment & Health* (2022).

[9] Michael Haenlein and Andreas Kaplan. "A brief history of artificial intelligence: On the past, present, and future of artificial intelligence". In: *California management review* 61.4 (2019), pp. 5–14.

[10] NK Geetha and P Bridjesh. "Overview of machine learning and its adaptability in mechanical engineering". In: *Materials Today: Proceedings* (2020).

[11] Michael W Berry, Azlinah Mohamed, and Bee Wah Yap. *Supervised and unsupervised learning for data science*. Springer, 2019.

[12] Chaimae Zai et al. "Prediction of Water Quality Using Artificial Intelligence (AI) and Statistical Approach". In: *International Conference on Digital Technologies and Applications*. Springer. 2022, pp. 34–42.

[13] Diwash Poudel et al. "Comparison of Machine Learning Algorithms in Statistically Imputed Water Potability Dataset". In: 5 (July 2022), pp. 38–46.

[14] Mohamed Bouamar and Mohamed Ladjal. "Evaluation of the performances of ANN and SVM techniques used in water quality classification". In: *2007 14th IEEE International Conference on Electronics, Circuits and Systems*. IEEE. 2007, pp. 1047–1050.

[15] *Dataset for Water Quality Classification*. URL: https://www.kaggle.com/code/tolgakurtulus/logistic-regression-classification-91-7-acc/data.

[16] Rohan Gawhade et al. "Computerized Data-Preprocessing To Improve Data Quality". In: 2022, pp. 1–6. DOI: 10.1109/ICPC2T53885.2022.9776676.

[17] Georgios Douzas, Fernando Bacao, and Felix Last. "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE". In: *Information Sciences* 465 (2018), pp. 1–20.

[18] *Sklearn.preprocessing.MinMaxScaler*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html.

[19] Ting Lan et al. "A comparative study of decision tree, random forest, and convolutional neural network for spread-F identification". In: *Advances in Space Research* 65.8 (2020), pp. 2052–2061.

[20] Jehad Ali et al. "Random forests and decision trees". In: *International Journal of Computer Science Issues (IJCSI)* 9.5 (2012), p. 272.

[21] Hasriq Izzuan Hasnol Yusri et al. "Water Quality Classification Using SVM And XGBoost Method". In: *2022 IEEE 13th Control and System Graduate Research Colloquium (ICSGRC)*. IEEE. 2022, pp. 231–236.

[22] A Najah et al. "An application of different artificial intelligences techniques for water quality prediction". In: *Int J Phys Sci* 6.22 (2011), pp. 5298–5308.

[23] Saleh Y Abuzir and Yousef S Abuzir. "Machine learning for water quality classification". In: *Water Quality Research Journal* (2022).