How to Cite:
SAHRAOUI, A., & SOUAKRI, R. (2025). Predicting water potability using machine learning classification methods. *International Journal of Economic Perspectives*, *19*(12), 19–28. Retrieved from https://ijeponline.org/index.php/journal/article/view/1236

Predicting water potability using machine learning classification methods

SAHRAOUI Abdelaziz
University of ABBES Laghrour Khenchela, Algeria
Email: sahraoui.abdelaziz@univ-khenchela.dz

SOUAKRI Roufaida
University of Shahid Mustapha Benboulaid Batna 2, Algeria
Email: roufaida.souakri@univ-batna2.dz

***Abstract*---**Ensuring access to safe drinking water remains a critical global challenge, making the ability to assess water potability both efficiently and accurately increasingly important. This study explores the use of supervised machine learning techniques to predict the potability of water based on a set of physicochemical attributes, including pH, hardness, turbidity, dissolved solids, and various chemical concentrations. Two classification models are examined: logistic regression, which provides a linear and interpretable decision boundary, and Support Vector Machines (SVM), which offer greater flexibility by capturing complex, non-linear relationships within the data. A systematic evaluation framework is employed to compare both models using key performance indicators such as accuracy, recall, precision, and F1-score. These metrics allow for a comprehensive understanding of each model's strengths, limitations, and capacity to generalize to unseen samples. The analysis aims not only to identify the more effective algorithm but also to highlight the potential of machine learning as a reliable tool for environmental monitoring and water quality assessment. The findings contribute to ongoing efforts to automate and enhance potability prediction, thereby supporting informed decision-making in public health and resource management.

***Keywords*---**Water Potability, Machine Learning, lassification, Logistic Regression, Support Vector Machines (SVM).

1. INTRODUCTION

Access to clean and safe drinking water is essential to human health, economic development, and environmental sustainability. However, despite global progress,

***Corresponding author**: SAHRAOUI, A., Email: sahraoui.abdelaziz@univ-khenchela.dz*
Submitted: 09 May 2025, Revised: 21 July 2025, Accepted: 27 November 2025

water contamination remains a widespread issue, particularly in developing regions where rapid urbanization, inadequate waste management, and industrial pollution place increasing pressure on water resources. Traditional laboratory-based methods for evaluating water quality, although reliable, often involve lengthy procedures, specialized equipment, and significant financial resources. These constraints limit the frequency and accessibility of testing, especially in remote or underserved areas. As a result, there is a growing need for faster, more cost-effective, and scalable approaches to assess water potability.

Recent advances in data science have opened new possibilities for environmental monitoring. Machine learning, in particular, has emerged as a powerful tool capable of detecting complex patterns within large datasets. By leveraging physicochemical indicatorssuch as pH, hardness, turbidity, dissolved solids, sulfate levels, and other measurable water characteristicsmachine learning models can infer the likelihood that a water sample is fit for human consumption. Such predictive systems have the potential to complement traditional laboratory analyses, providing early warnings, guiding field testing, and optimizing resource allocation for water treatment operations.

Among the wide range of supervised learning algorithms, logistic regression and Support Vector Machines (SVM) are both recognized for their classification capabilities. Logistic regression offers a simple, interpretable, and computationally efficient baseline, making it suitable for initial assessments and practical deployment. In contrast, SVM models are designed to handle high-dimensional datasets and capture non-linear relationships through the use of kernel functions, often leading to improved predictive performance in complex classification scenarios.

The aim of this study is to investigate the effectiveness of logistic regression and SVM in predicting water potability from physicochemical data. By comparing these models across key performance metricsincluding accuracy, precision, recall, and F1-score, the analysis seeks to identify the most reliable approach for this specific classification task. Beyond model comparison, this work highlights the broader potential of machine learning to support public health efforts, strengthen water quality monitoring systems, and contribute to global initiatives aimed at ensuring universal access to safe drinking water.

2. Model Background and Theoretical

2.1. Logistic Regression

Logistic regression is a statistical model used to analyze the relationship between a binary dependent variable and a set of independent variables. It predicts the probability that a given event will occur (class 1) or not occur (class 0) based on explanatory features. As part of the family of generalized linear models, logistic regression employs the logisticor sigmoidfunction as its link function, transforming a linear combination of inputs into a probability between 0 and 1.

Model fitting involves estimating the coefficients β\betaβ in a way that maximizes the likelihood of the observed data. In other words, the goal is to find the

parameter values that make the observed outcomes most probable under the model. This estimation is typically performed using the Maximum Likelihood Estimation (MLE) method.

Logistic regression is widely used across many fields, including medicine, economics, marketing, and the social sciences. It is commonly applied to binary classification tasks such as predicting whether a patient is ill or healthy, whether a customer will make a purchase, or whether an individual is likely to vote for a particular candidate.

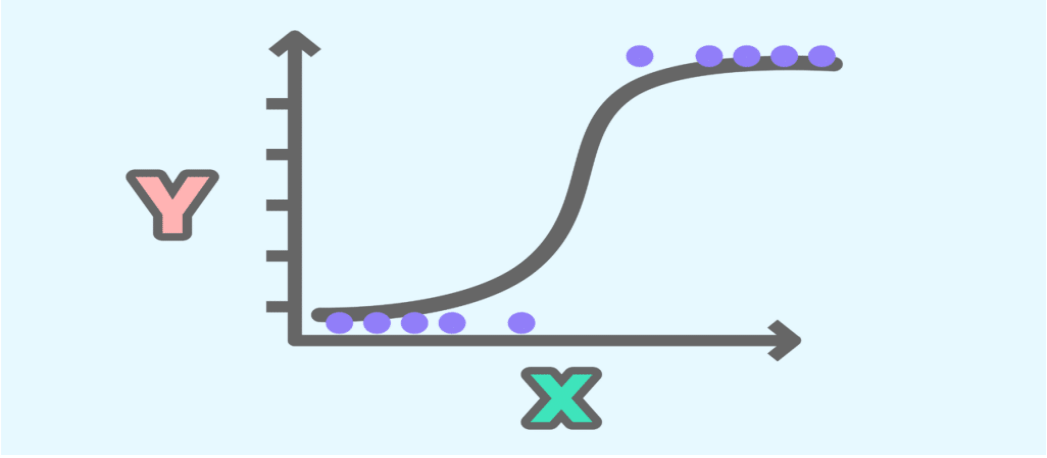


Figure 1: Logistic regression graph

The fundamental principle of logistic regression is based on using the sigmoid function to model the probability that the dependent variable takes the value 1. The sigmoid function is defined as follow :

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$

In this equation, β0betais the intercept, β1, β2, ..., βn are the coefficients associated with the independent variables X1, X2, ..., Xnand e is the base of the natural logarithm. The sigmoid function transforms the linear combination of independent variables into a probability value between 0 and 1.

2.2. Support Vector Machines (SVM)

Support Vector Machines (SVM) are a set of supervised learning algorithms widely used for classification and regression tasks. The core idea of SVM is to find the optimal hyperplane that best separates data points of different classes in a high-dimensional feature space. The optimal hyperplane is defined as the one that maximizes the margin, which is the distance between the hyperplane and the nearest data points from each class, known as support vectors. By focusing on these critical points, SVM achieves robust generalization even with complex datasets.

SVM can handle both linearly and non-linearly separable data. For linearly separable cases, the algorithm identifies a straight hyperplane that divides the classes. For non-linear scenarios, SVM employs kernel functions (such as polynomial, radial basis function, or sigmoid kernels) to map the data into a higher-dimensional space where a linear separation becomes possible. This flexibility makes SVM particularly effective for capturing complex patterns and relationships within the data.

SVM has been successfully applied in numerous fields, including image recognition, bioinformatics, text classification, and environmental monitoring. In the context of water potability prediction, SVM can effectively classify water samples based on physicochemical attributes, even when the relationship between features and potability is non-linear or highly complex.

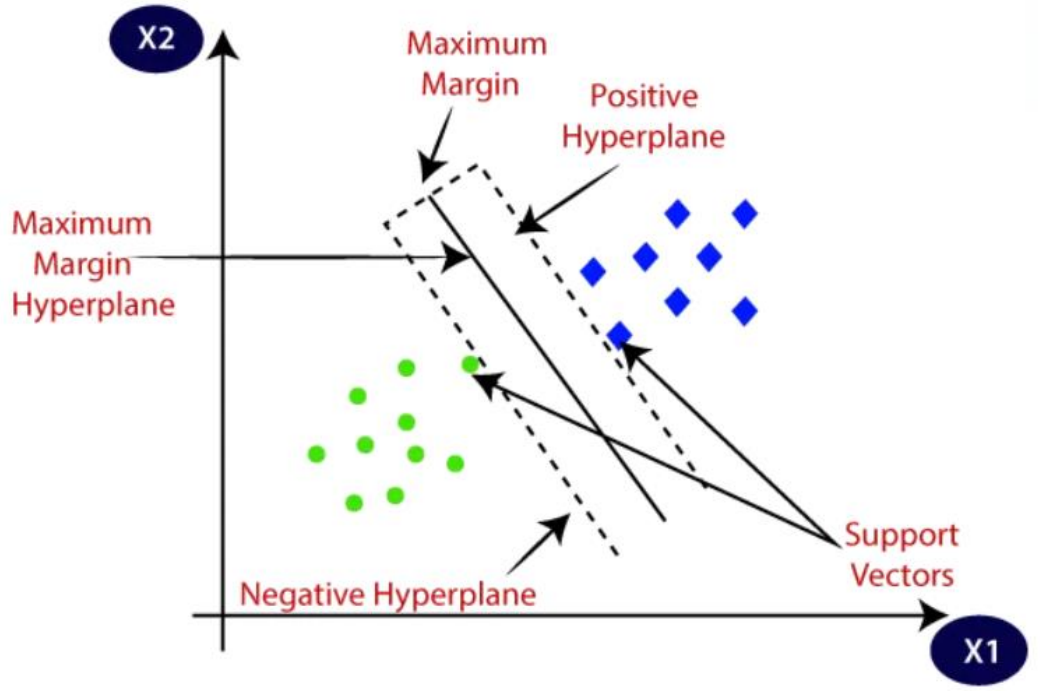


Figure 2: SVM graph

In Support Vector Machines, the central concept is the **hyperplane**, which is a decision boundary that separates data points of different classes in the feature space. For a dataset with two features, the hyperplane is a line; for three features, it becomes a plane; and in higher dimensions, it is referred to generally as a hyperplane. The goal of SVM is not just to find any separating hyperplane, but the **optimal hyperplane** that maximizes the **margin**.

The **margin** is defined as the distance between the hyperplane and the nearest data points from each class, which are called **support vectors**. Maximizing the margin is crucial because a larger margin reduces the model's generalization error, making it more robust to unseen data. Essentially, the SVM algorithm selects the hyperplane that provides the **largest possible separation** between the

classes, ensuring the model is less sensitive to noise and small variations in the dataset.

Mathematically, if the hyperplane is defined as:

$$w^T x + b = 0$$

Were

- w is the **weight vector**, which is perpendicular (normal) to the hyperplane.
- x is the **feature vector**, representing a point in the input space.
- b is the **bias** (or offset), which shifts the hyperplane away from the origin.

2.3. Theoretical Comparison Between Logistic Regression and SVM:

Both Logistic Regression (LR) and Support Vector Machines (SVM) are supervised learning algorithms used for classification, but they differ fundamentally in their approach, assumptions, and strengths.

Table 1: Theoretical Differences Between Logistic Regression and Support Vector Machines

Feature	Logistic Regression (LR)	Support Vector Machines (SVM)
Type of model	Probabilistic, based on estimating probabilities using the logistic (sigmoid) function.	Geometric, based on finding the optimal hyperplane that maximizes the margin between classes.
Decision boundary	Linear in the feature space (unless manually adding polynomial or interaction terms).	Can be linear or non-linear depending on the kernel used; capable of mapping data to higher-dimensional spaces.
Output	Provides probabilities for class membership (values between 0 and 1).	Provides a hard classification; can also produce probabilistic estimates with additional calibration.
Handling non-linear data	Requires feature engineering (polynomials, interaction terms) to model non-linear relationships.	Naturally handles non-linear relationships through kernel functions (e.g., RBF, polynomial).
Interpretability	High: coefficients directly indicate the influence of each feature on the outcome.	Moderate to low: the decision boundary is determined by support vectors, which may not give a simple interpretation of feature importance.
Optimization	Uses Maximum Likelihood Estimation (MLE) to minimize the logistic loss.	Solves a convex optimization problem to maximize the margin between classes (hinge loss).

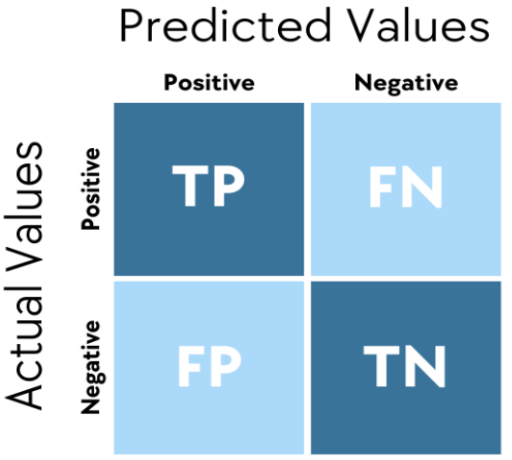
2.4. Evaluation Metrics for Classification Models

To assess the performance of machine learning models for water potability prediction, several standard evaluation metrics are employed. These metrics

provide a comprehensive view of how well a model classifies water samples as potable or non-potable.

2.4.1. Confusion Matrix

The confusion matrix is a tabular representation of model predictions versus actual class labels:



Where the four basic outcomes are:

- **True Positives (TP):** Cases where the model correctly predicts the positive class.
- **True Negatives (TN):** Cases where the model correctly predicts the negative class.
- **False Positives (FP):** Cases where the model incorrectly predicts the positive class, while the actual class is negative. This is also called a Type I error.
- **False Negatives (FN):** Cases where the model incorrectly predicts the negative class, while the actual class is positive. This is also called a Type II error.

These four outcomes form the confusion matrix, which provides a detailed view of the model's performance and is the foundation for calculating metrics such as accuracy, precision, recall, and F1-score.

2.4.2. Accuracy

Accuracy measures the overall correctness of a model by calculating the proportion of correctly predicted instances out of all predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

It gives a general sense of performance but can be misleading if the classes are imbalanced.

2.4.3. Precision

Precision measures the correctness of positive predictions. It is the proportion of instances predicted as positive that are actually positive:

$$\text{Precision} = \frac{TP}{TP + FP}$$

High precision means that when the model predicts a positive class, it is usually correct.

2.4.4. Recall (Sensitivity)

Recall measures the model's ability to identify all actual positive instances. It is the proportion of true positives that are correctly detected:

$$\text{Recall} = \frac{TP}{TP + FN}$$

High recall ensures that most actual positive cases are captured by the model.

2.4.5. F1-Score

The F1-score is the harmonic mean of precision and recall:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

It balances precision and recall, making it useful when there is a trade-off between false positives and false negatives.

3. Dataset Description and Preprocessing

3.1. Dataset Description

The dataset used in this study consists of 3,276 rows and 10 columns, containing water quality measurements for 3,276 different water samples. It is sourced from a publicly available dataset designed to assess water potability based on physicochemical measurements. The following table summarizes the main variables included in the dataset:

Table 2: presentation of the variables

Variable	Description
Ph	Indicates the acidity or alkalinity of water (0 – 14)
Hardness	Water hardness (presence of calcium and magnesium)
Solids	Total dissolved solids concentration (mg/L)
Chloramines	Concentration of chloramines used for disinfection
Sulfate	Sulfate concentration (mg/L)
Conductivity	Electrical conductivity of water (µS/cm)
Organic_carbon	Total organiccarbon concentration
Trihalomethanes	Presence of trihalomethanes, by-products of chlorination

Variable	Description
Turbidity	Measure of water turbidity (NTU)
Potability	Target variable: 0 = non-potable, 1 = potable

3.2. Dataset Preprocessing

Before applying machine learning models, the dataset underwent several preprocessing steps to ensure quality and suitability for analysis:

1. **Handling Missing Values:**
   Some features contained missing values, which were imputed using the mean or median of the respective column.
2. **Feature Scaling:**
   Since many features are on different scales (e.g., pH vs. solids), numerical variables were standardized to improve model performance, especially for algorithms sensitive to scale such as SVM.
3. **Encoding the Target Variable:**
   The target variable was converted to a binary format (0 for non-potable, 1 for potable) suitable for classification algorithms**.**
4. **Data Splitting:**
   The dataset was divided into training and testing sets, typically using an 80/20 split, to train the models and evaluate their generalization performance.

These preprocessing steps ensure that the data is clean, normalized, and ready for model training and evaluation.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ph               2785 non-null   float64
 1   Hardness         3276 non-null   float64
 2   Solids           3276 non-null   float64
 3   Chloramines      3276 non-null   float64
 4   Sulfate          2495 non-null   float64
 5   Conductivity     3276 non-null   float64
 6   Organic_carbon   3276 non-null   float64
 7   Trihalomethanes  3114 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
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 2116 entries, 3 to 3271
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ph               2116 non-null   float64
 1   Hardness         2116 non-null   float64
 2   Solids           2116 non-null   float64
 3   Chloramines      2116 non-null   float64
 4   Sulfate          2116 non-null   float64
 5   Conductivity     2116 non-null   float64
 6   Organic_carbon   2116 non-null   float64
 7   Trihalomethanes  2011 non-null   float64
 8   Turbidity        2116 non-null   float64
 9   Potability       2116 non-null   int64
dtypes: float64(9), int64(1)
memory usage: 181.8 KB
```

Figure 3: before and after Data Preprocessing

Three columns in the dataset contained missing values:

- **pH:** 491 missing values
- **Sulfate:** 781 missing values
- **Trihalomethanes:** 162 missing values

These missing values were replaced with the **mean of the respective variable**. This imputation method preserves the overall structure of the dataset while

avoiding the loss of information that would occur if rows with missing values were removed.

4. Results and Discussion

The performance metrics of the two classification models, Logistic Regression and Support Vector Machine (SVM), are summarized in the table below:

Table 3: presentation of the variables

Modèle	Accuracy	Recall	Précision	F1 score
Régression Logistique	0,5990	0.3235	0.37	0.42
Support Vector Machine	0,6981	0.4836	0.8088	0.46

From the results, it is clear that the Support Vector Machine outperforms Logistic Regression across all evaluation metrics.

- **Accuracy:** The SVM achieves an accuracy of approximately 69.8%, compared to 59.9% for Logistic Regression, indicating that SVM classifies more samples correctly overall.
- **Recall:** SVM's recall (48.4%) is notably higher than that of Logistic Regression (32.4%), meaning SVM detects a larger proportion of true positive cases (potable water). This is crucial in minimizing false negatives, which in this context correspond to failing to identify safe water.
- **Precision:** The precision of SVM is significantly higher (80.9%) compared to Logistic Regression (37%), demonstrating that when SVM predicts water as potable, it is more likely to be correct. This reduces the risk of false positives—incorrectly classifying unsafe water as safe.
- **F1 Score:** The F1 score, which balances precision and recall, is also higher for SVM (0.46) than for Logistic Regression (0.42), confirming its better overall classification performance.

These results suggest that the SVM model, with its ability to handle non-linear patterns and complex relationships, is better suited for predicting water potability based on physicochemical parameters. Logistic Regression, while simpler and more interpretable, may struggle to capture the complexities in the data, resulting in lower predictive performance.

5. Conclusion

This study successfully demonstrated the power of machine learning in predicting water potability from physicochemical data, a critical step toward ensuring public

health and safe water access. By comparingLogisticRegression and Support Vector Machines (SVM), we highlighted the clear advantage of SVM in handling complex, non-linear relationships inherent in water quality data.

The superior performance of SVMevident in higher accuracy, precision, recall, and F1 scoresunderscores its potential as a reliable and robust tool for real-world water quality assessment. While Logistic Regression offers interpretability and simplicity, its linear limitations reduce effectiveness in this context.

Looking forward, integrating advanced models and richer datasets could further enhance predictive accuracy, enabling proactive water safety management and environmental monitoring. This research paves the way for smarter, data-driven solutions that empower communities and policymakers to make informed decisions about water resources. In a world where clean water is increasingly precious, leveraging machine learning is not just an opportunityit is a necessity.

6. References

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Cortes, C., &Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

Kowsari, K., Brown, D. E., Heidarysafa, M., Meimandi, K. J., Gerber, M. S., & Barnes, L. E. (2019). Text classification algorithms: A survey. *Information, 10*(4), 150.https://doi.org/10.3390/info10040150

Sahu, A. K., & Kumar, D. (2020). Water quality prediction using machine learning techniques. *Environmental Science and Pollution Research, 27*(7), 7500–7511. https://doi.org/10.1007/s11356-019-06969-4

World Health Organization. (2017). *Guidelines for drinking-water quality* (4th ed.). WHO Press.

Chakraborty, S., & Sinha, S. (2021). Predicting water quality using machine learning: A case study. *Journal of Environmental Management, 281*, 111898. https://doi.org/10.1016/j.jenvman.2020.111898

Pham, T. T., Nguyen, H. T., & Le, T. H. (2020). Comparative analysis of machine learning algorithms for water quality prediction. *Environmental Monitoring and Assessment, 192*(10), 650. https://doi.org/10.1007/s10661-020-08440-3

Tiwari, S., & Mishra, S. (2019). Machine learning approaches for water quality prediction. *Sustainable Water Resources Management, 5*(4), 1689–1701. https://doi.org/10.1007/s40899-019-00367-2

Raj, M., & Prasad, K. (2022). Application of SVM and logistic regression in environmental data analysis. *Environmental Data Science, 4*(2), 45–56.

Li, X., Zhou, Q., & Wang, L. (2018). Assessment of drinking water quality using supervised machine learning algorithms. *Journal of Cleaner Production, 198*, 1232–1242. https://doi.org/10.1016/j.jclepro.2018.07.056