Evaluating the Use of Support Vector Machines and Multilayer Perceptrons for the analysis of Water Quality

Karin Falconer-Bailey | 210025601 | INM427 Neural Computing

Abstract — This paper focuses on data originally sourced from Kaggle user Aditya Kadiwa. Through Support Vector Machines (SVMs) and Multilayer Perceptrons (MLPs), a critical analysis of these supervised learning algorithms will be conducted upon a dataset concentrated on identifying the characteristics and trends of water potability. These algorithms have been conducted in the Jupyter Notebook.

Keywords — Support Vector Machines, Multilayer Perceptrons, Deep Learning, Classification

1 PROBLEM STATEMENT

Access to safe drinking water is vital for the evolution of life. While many people believe water accessibility is an issue only experienced by developing nations, according to a report conducted by the World Health Organization (WHO, 2019), “one in three individuals globally do not have access to safe drinking water”. In March 2022, the World Health Organization estimated that 2.2 billion people, worldwide, do not have frequent access to safe drinking water (WHO, 2022). While this study’s dataset does not focus on regional or country-specific water quality, given that clean water is globally unattainable, we can study which factors contribute to water potability and thus, produce a dependable model to predict the quality of water to improve upon current water systems.

Support Vector Machines (SVMs) and Multilayer Perceptrons (MLPs) are the two supervised machine learning methods that will be used to analyse and investigate water quality trends. The justification for the selection of these algorithms, as opposed to others, will be discussed separately once we have gained familiarity with the components of the dataset.

2 DATASET

While the collation of the dataset is of a single origin, obtained from Kaggle, the CSV file presents various components attributing to the quality of data. Before any pre-processing or training, the data is consistent with 3276 observations over ten columns. These variables include pH, hardness, solids, chloramines, sulfates, conductivity, organic carbon, trihalomethanes, turbidity, and of course potability. The dataset is mostly inclusive of measured quantitative data with the potability observation serving as a dummy variable; zero for not potable and one for potable. Given the nature of the study, potability will serve as the response variable and the remaining variables as the predictors.

3 SUPERVISED LEARNING

Supervised learning algorithms tend to deal with two forms of machine learning problems: classification and regression (Muller and Guido, 2016). To manage the problems experienced within classification tasks, we aim “to predict a class label” while in regression problems, we aim “to predict a continuous number” (Muller and Guido, 2016, pg.25-26).

Multilayer Perceptrons (MLPs)

Multilayer Perceptrons, commonly referred to as Feedforward Artificial Neural networks, is a neural network consisting of “three layers; an input layer, an output layer, and a layer in between, not connected directly to the input or the output” (Beale and Jackson, 1990, pg. 67). This form of deep learning algorithm is used for “tabular datasets, classification prediction problems, and regression prediction problems” (Brownlee, 2019).

Table 01. Advantages and disadvantages of MLPs

Advantages	Disadvantages
· For classification and regression problems · Simple and non-time consuming to design · Applicable for large datasets · Capable of learning non-linear models · Used on a variety of dataset formations	· Has a sensitivity to feature scaling · Requires hyperparameter tuning · Random weights can cause incorrect performance accuracies · Model training can be critical

Adapted from: Chawla (2021) & Akkaya and Çolakoğlu (2019, pg.25)

Support Vector Machines (SVMs)

Support Vector Machines are classed as one of the popular supervised learning algorithms to this day. Like MLPs, Support Vector Machines can train classification and regression models. To grasp the concept of SVMs, we must refer to the term *hyperplane*. In machine learning, a hyperplane is a boundary that acts as a tool in supervised learning algorithms that creates a separation between the different classes within a dataset.

Table 02. Advantages and disadvantages of SVMs

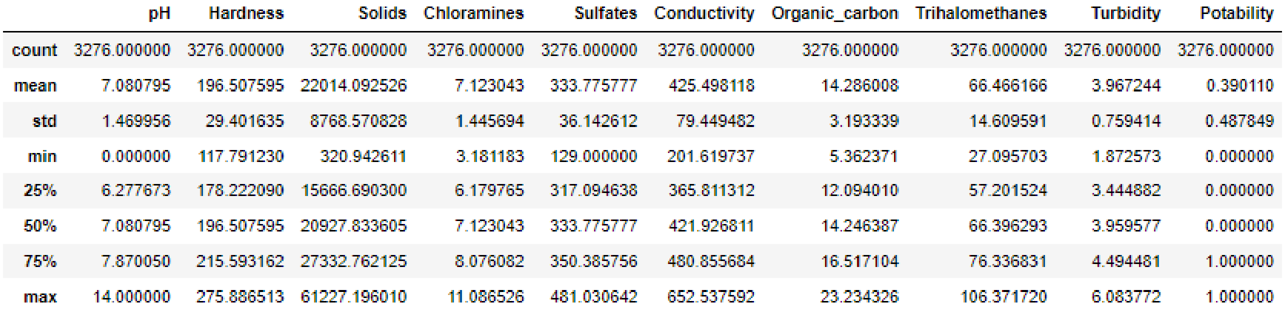
Advantages	Disadvantages
· For classification and regression problems · Aims to identify the differentiation between classes · Effective in high-dimensional spaces	· Training can be timely if the dataset is large · If noise is present, SVMs do not tend to perform well · Results can be difficult to interpret

Adapted from: Sarker (2021, pg. 6) & Akkaya and Çolakoğlu (2019, pg.24)

4 Exploratory data analysis (EDA)

After cleaning the data, by removing column inconsistencies by replacing NaN values with the corresponding column’s mean, the following stage of the data process was to conduct exploratory data analysis. This first began with examining the descriptive statistics of the water quality dataset.

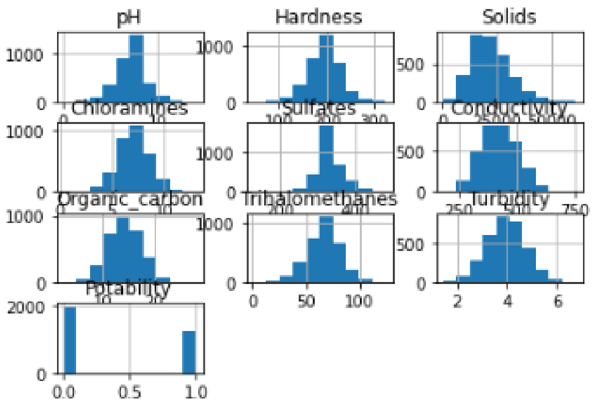
Figure 01. Descriptive statistics of water quality



	pH	Hardness	Solids	Chloramines	Sulfates	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	3276.000000	3276.000000	3276.000000	3276.000000	3276.000000	3276.000000	3276.000000	3276.000000	3276.000000	3276.000000
mean	7.080795	196.507595	22014.092526	7.123043	333.775777	425.498118	14.286008	66.466166	3.967244	0.390110
std	1.469956	29.401635	8768.570828	1.445694	36.142612	79.449482	3.193339	14.609591	0.759414	0.487849
min	0.000000	117.791230	320.942611	3.181183	129.000000	201.619737	5.362371	27.095703	1.872573	0.000000
25%	6.277673	178.222090	15666.690300	6.179765	317.094638	365.811312	12.094010	57.201524	3.444882	0.000000
50%	7.080795	196.507595	20927.833605	7.123043	333.775777	421.926811	14.246387	66.396293	3.959577	0.000000
75%	7.870050	215.593162	27332.762125	8.076082	350.385756	480.855684	16.517104	76.336831	4.494481	1.000000
max	14.000000	275.886513	61227.196010	11.086526	481.030642	652.537592	23.234326	106.371720	6.083772	1.000000

In figure 01, we observe the descriptive statistics of the dataset. Here we can examine each observation's central tendencies, and thus their distribution. From the *count* row, we can see that all attributes each contain 3276 recorded values. Both pH and potability possess the lowest retrieved value of 0, while the solids contain the maximum value of 61227.19. The standard deviation of each variable is minimal, excluding solids, conductivity, and hardness. This means that the dataset does not highly deviate from the mean and thus, does not consist of a great extent of variability.

Figure 02. Feature distribution



Furthering from the last point made above, an almost Gaussian distribution appears present in the adjacent figure. Though, upon conducting Jarque-Bera's test for normality, at a level of significance of α= 0.05, the null hypothesis was rejected for hardness, chloramines, conductivity, organic carbon, trihalomethanes, and turbidity concluding that these observations do not have a normal distribution. Outliers can prove useful in the construction of a machine learning model, therefore, only extremely skewed observations were removed to prevent statistical errors and reduced levels of accuracy.

Results of Pearson's correlation coefficients functioned as a filter method for feature selection. Through correlation, we can eliminate the chances of our model coinciding with the assumptions of multiple linear regression, specifically multicollinearity. Instead, we can save time by reducing the number of training models, and in its place, we generate a model using features that solely correlate with potability.

From the correlation coefficients, it was evident that the variables hold a strong correlation neither with each other nor with the response variable, potability. While this study aims to address a classification problem, the absence of correlation does not pose a significant concern as it would in regression analysis. Assessing the data for multicollinearity is necessary to avoid issues when interpreting model results.

5 HYPOTHESIS

Given the capabilities of Support Vector Machines and the reputation that it precedes, we should expect the SVM model to outperform the MLP model.

6 METHODOLOGY AND EXPERIMENTAL RESULTS

The test and train methodology divides the data into an 80:20 ratios for training and testing, based on the 3276 observations in the water quality dataset. The Pareto principle, devised by economist Vilfredo Pareto, acknowledges the potential of the 80:20 rule by insuring the collection of the model's best results. By partitioning the data between training, validation, and testing, we can ensure that the model provides reliable results against unseen data.

Feature scaling was implemented using the approach of standardisation, after the models were separated into testing and training partitions. After observing and attempting to generate a Gaussian distribution, it was decided that this method of feature scaling was appropriate, rather than normalisation, due to the data's normal distribution.

MLP

On the hidden layer of our neural network, a ReLu activation function was employed, while on the outer layer, a softmax activation function was used, resulting in a better computational result and assuring that our network's neurons were not simultaneously active. It is likely that within the final model solely a ReLu activation will be applied, as implementing both functions produce a trade-off. For the simplicity of the model, the parameters were assigned to specific values: a maximum of 30 epochs, a learning rate of 0.01, a dropout of 0.5 and no method of cross validation was used. SMOTE was used in the training model, however upon seeing the results, the model was found to be underperforming.

It can be observed that the initial MLP model was built with a simple architecture. As a result, it was possible to determine how each parameter contributed to each model's performance, avoiding the insertion of any factors that had no substantial impact on the model's conclusion.

SVM

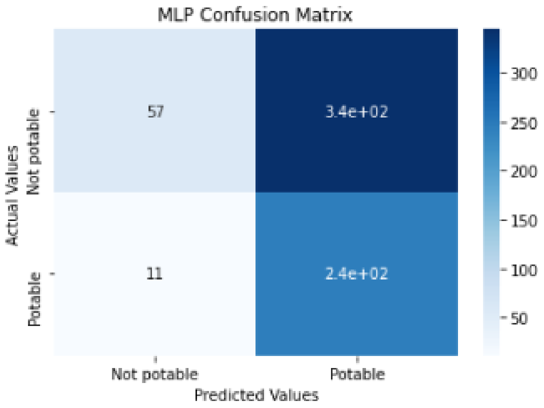
Given the limitations of the parameters within Support Vector Machines, the approach of a simple model architecture was too applied for this algorithm. A linear kernel was used and again, SMOTE caused the SVM model to underperform and thus, will be removed from the final structure of both testing models.

7 Results

In the final models of the Multilayer Perceptron and the Support Vector Machines, both models underwent extensive hyperparameter tuning.

MLP

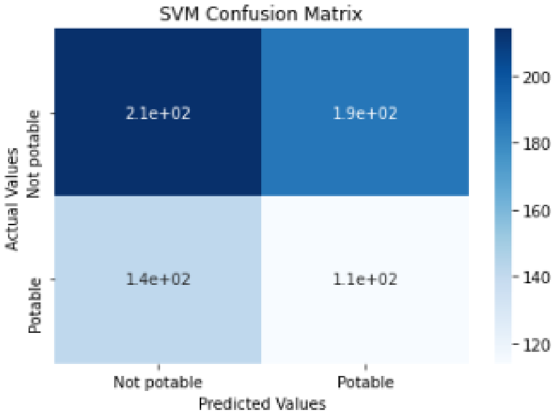
Figure 03. MLP confusion matrix



For the final Multilayer Perceptron model, various hyperparameters were implemented. We then provided with an output of the most optimal parameters for the testing of the model. It was concluded that the optimal hyperparameters were 40 epochs, a learning rate of 0.1, a batch size of 200, random grid cross-validation of 3, the momentum of 1.0, and ReLu and cross-entropy loss as activation functions. In the testing of the final MLP model, the testing dataset proved to be more accurate than when in training, providing 61 per cent of accuracy during testing and an accuracy of 45.88 per cent during training.

SVM

Figure 04. SVM confusion matrix



In the final SVM model, the following parameters were included: an RBF kernel, a random state of zero and k-fold cross-validation of 10. Despite fewer parameters existing in the SVM model, the Support Vector Machine model outperformed the Multilayer Perceptron in both training and testing. The SVM exhibited an accuracy of 50 per cent during training and improved accuracy of 67.29 per cent during training.

8 CONCLUSION

It was hypothesised that the Support Vector Machine would outperform the Multilayer Perceptron model. While the accuracy results of each model are closer to being an average than great accuracy, to an extent, two models have been developed to display and recognise the difference between potable water and non-potable water.

To improve the study, I would reduce the removal of the data's outliers. Perhaps feature scaling through normalisation would improve the performance of both models. Specifically, within the Multilayer Perceptron model, I would include more hyperparameters and conduct experimental results using the backpropagation algorithm or the forward propagation algorithm. This would potentially too provide my MLP model with greater accuracy. The addition of further hyperparameters for the Support Vector Machines would be applied, to improve the model's results and to exhibit a more complex architecture.

9 References

1. Akkaya, B. and Çolakoğlu, N. (2019) Comparison of Multi-class Classification Algorithms on Early Diagnosis of Heart Disease, Recent Advances in Data Science and Business Analytics. Available at: https://www.researchgate.net/publication/338950098_Comparison_of_Multi-class_Classification_Algorithms_On_Early_Diagnosis_of_Heart_Disease, pg.24-25 (Accessed 19th April 2022)
2. Albon, C. (2018) Python Machine Learning Cookbook: Practical Solutions from Pre-processing to Deep Learning. O'Reilly Media: Sebastopol, United States of America
3. Beale, R. and Jackson, T. (1990) Neural Computing: An Introduction. Institute of Physics Publishing: Bristol, United Kingdom
4. Brownlee, J. (2019) When to Use MLP, CNN, and RNN Neural Networks, *Machine Learning Mastery.* Available at: https://machinelearningmastery.com/when-to-use-mlp-cnn-and-rnn-neutral-networks/ (Accessed 18th April 2022)
5. Diwal, A. (2021) Water Potability, *Kaggle.* Available at: https://www.kaggle.com/datasets/adityakadiwal/water-potability (Accessed 13 April 2022)
6. Elogeel, A. (2010) Multilayer Perceptron, *Word Press.* Available at: https://elogeel.wordpress.com/2010/05/05/multilayer-perceptron/ (Accessed 18th April 2022)
7. Muller, A. and Guido, S. (2016) Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media: Sebastopol, United States of America, pg. 25-26.
8. Sarker, H. (2021) Machine Learning: Algorithms, Real-World Applications, and Research Directions, SN Computer Sci. 2(160). Doi: 10.1007/s42979-021-00592-x, pg.6.
9. WHO (2019) 1 in 3 people globally do not have access to safe drinking water. Available at: https://www.who.int/news/item/18-06-2019-1-in-3-people-globally-do-not-have-access-to-safe-drinking-water-unicef-who (Accessed 15th April 2022)
10. WHO (2022) World Water Day 2022: Groundwater, invisible but vital to health. Available at: https://www.who.int/news-room/feature-stories/detail/world-water-day-2022-groundwater-invisible-but-vital-to-health (Accessed 15th April 2022)