Predicting Water Potability with Support Vector Machines and Multi-Layer Perceptron's

Tharun Guduguntla

Abstract

This paper aims to determine the best model suited for the binary classification task of predicting the portability of the water. Here two models were compared against each other in a similar methodology of choosing the parameters. The two models were SVM (Support Vector Machines) and MLP (Multi-Layer Perceptron). Various methodologies like Cross validation, early stopping etc has been performed to choose the best model. With the help of Confusion Matrix and the ROC curves, the performance has been compared. The outcome has resulted that the SVM performed well in this type of analysis.

1. Introduction

After air, water is the most essential need for life on earth. The quality of water can be described with its physical, chemical and biological characteristics. Water can be classified as potable, palatable, polluted and infected water. Out of these the potable water is those which are firstly safe to drink, secondly pleasant to taste and not usable for any domestic purposes. Water quality can be determined by considering various factors such as physical parameters like turbidity, colour, odour, conductivity etc, chemical parameters such as pH, chloride, fluoride, hardness etc, biological parameters such as viruses, bacteria, algae etc.

Determining the water quality by considering various parameters can determine the health conditions of many regions. Also, it gained the economic benefit for the regions as the cost for the health care superseded the costs of water treatment. We aim to achieve an accurate prediction in determining the potability of water by applying two neural network algorithms. When determining the classification problem, SVM performs best due to its decision boundary ability even in n-Dimensional space.

2. Dataset

The data for this analysis has been fetched from Kaggle (kadiwal, 2021). Water quality dataset consists of 10 variables out of which potability (last feature) is our target variable with 3276 observations taken from different water bodies. The feature variables considered in this analysis are pH value, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes and Turbidity as shown in Figure 1.

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0
5	5.584087	188.313324	28748.687739	7.544869	326.678363	280.467916	8.399735	54.917862	2.559708	0
6	10.223862	248.071735	28749.716544	7.513408	393.663396	283.651634	13.789695	84.603556	2.672989	0
7	8.635849	203.361523	13672.091764	4.563009	303.309771	474.607645	12.363817	62.798309	4.401425	0
8	NaN	118.988579	14285.583854	7.804174	268.646941	389.375566	12.706049	53.928846	3.595017	0
9	11.180284	227.231469	25484.508491	9.077200	404.041635	563.885481	17.927806	71.976601	4.370562	0
10	7.360640	165.520797	32452.614409	7.550701	326.624353	425.383419	15.586810	78.740016	3.662292	0

Figure 1: Dataset view

With the first glance on the dataset, we can observe that every column has numerical datatype of float except the target variable. In this dataset, target variable is a binary data type which has 0 for non-potable water and 1 for potable water. The target variable is balanced with 1998

records of 0 and 1278 records of 1 as shown in Figure 2. There could be a discussion to perform SMOTE technique to balance the target variable, but the values aren't highly imbalanced. Hence analysis is performed with the original dataset.

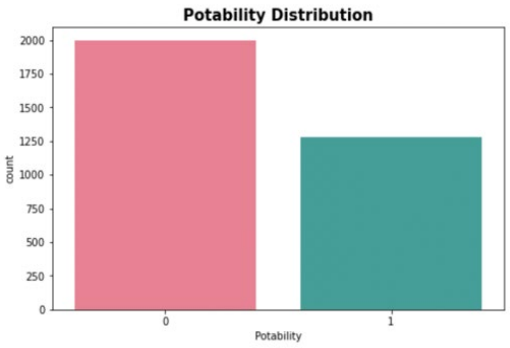

Figure 2: Target variable balance

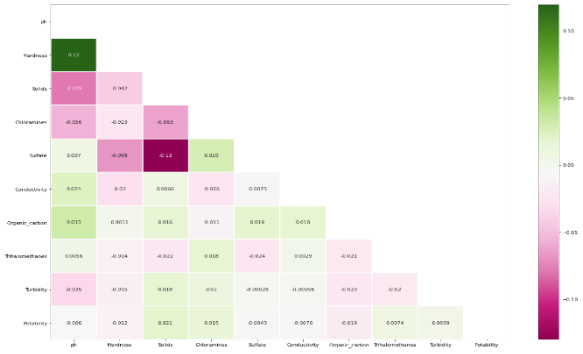

Figure 3: Correlation Heatmap

This dataset has null values in pH, Sulfate and Trihalomethanes. After finding the distribution of columns with null values, it is observed that the data points are properly skewed. Hence replacing the missing values with mean would be the best option. In order to check the outliers, different scatter plots were visualized and found outliers in many features. Hence with the help of Z-score method, outlier was removed. Z-score for more than 3 are considered outliers as these values aren't three standard deviations away from the mean. To check the correlation between the dependent variables a correlation heatmap is drawn as shown in Figure 3. From the figure we can observe that the potability is partially dependent on most of the features and reducing the feature dimension will not yield huge improvement in the results.

3. Neural Network Models

The two models used for the analysis are SVM (Support Vector Machines) and MLP (Multi-Layer Perceptron).

3.1. Support Vector Machines (SVM)

Support Vector Machines works on the principle of classifying the given data points by finding a hyperplane in an N-dimensional space. It works on a target to obtain the best hyperplane which has the highest margin. The procedure to find the maximum distance between the data points helps in increasing the confidence of classification for the new data samples. Simple SVM works best in finding the linear hyperplane between binary classes. When it comes to more classes, it uses Kernel trick to form a decision boundary. SVM kernel function takes the low dimensional space into a higher dimensional space but applying few complex data transformations (Gandhi, 2018).

Pros:

1. Most effective with higher dimensional problems
2. Memory efficient algorithm as it utilizes a subset of data points to determine the decision boundary (Sunil, 2017).
3. An effective model when the features count is higher than the data samples.
4. It is bound to reduce the influence from the outliers

Cons:

1. When used on very large datasets, the performance isn't good as the training time is huge.
2. If the data has higher noise, then target classes will be overlapping which reduces the power of prediction in SVM (McGregor, 2020).
3. It is highly inclined for over-fitting.

3.2. Multi-Layer Perceptron

It is a deep, artificial neural network which has more than one perceptron. A perceptron has the ability to result the output by creating a linear combination with the help of weights from the input values. MLP consists of an input layer to accept the signal, an output layer which creates a decision on the input and a chosen number of hidden layers which has the greatest computation power throughout the model. It is a feed forward algorithm which combines few initial weights with the input with respect to the activation function and is propagated through other layers in sequence. On the next step it has the property of back propagation where weights are adjusted in order to minimize the cost function (Bento, 2021).

Pros:

1. This method can be applied to non-linear scenarios as well
2. The prediction of the output will be quick after training is completed and can be used for real time data
3. The accuracy will not vary a lot even with the sliced set of data.

Cons:

1. More hyper parameters such as hidden layers, iterations and neurons makes it complicated to tune.
2. With the initial weights assignment, there arises a problem of local minima resulting in multiple validation scores.
3. The quality of the training plays a crucial role in the performance of the model.

4. Hypothesis Statement

With the size of dataset in consideration, few references (Wilson, 2020) suggest that the performance of SVM beats the best estimation of MLP. Local approximation plays a crucial role in SVM while global approximation is needed for MLP. The training time is expected to be high for SVM due to its property of transforming the data to multi-dimensional space with the help of kernel function to define the decision boundary. Nonetheless, both models' performance shouldn't have huge differences as per the data properties and the complexity of the models.

5. Training and Evaluation Methodology (5%)

5.1. SVM

The data has been divided into 85:15 ratio for training and testing respectively. Further, training data is divided into training and validation data in the ratio of 80:20. While training the SVM model, all the steps for analysis was performed on the scaled data as unit variance data will become a standard for all the features. The parameters are passed into our model are fine-tuned by both the search methods which are RandomizedSearchCV and GridSearchCV. With an aim to obtain the best parameters for our model, various parameters were used in these search methods and GridSearchCV will give absolute results than the other when the

parameter choices are optimal (Torino, 2020) Cross Validation has been applied at the beginning to check our model if it can generalize the results for different sets of data points. It also has an advantage of increasing the accuracy (Shulga, 2018). All the parameter optimizations techniques were cross validated in order to eliminate the overfitting problem. Due to the size of the dataset, we are more prone to over fitting. Hence a good K-fold cross validation can minimize this issue.

5.2. MLP

In the process of MLP modelling, the first step is to decide the structure of the network. The basic structure as discussed above has input layer, hidden layers and output layer. For this analysis two hidden layer model is built and used. The reason behind this choice is a deep neural network with 2 hidden layers will be a good start on a dataset of this size and has two target classes (Ranjan, 2019). Various methodologies such as scaling, early stopping and cross validation has been performed and its impact is considered for further modelling. All other parameters needed for this model are searched either by RandomizedSearchCV or with GridSearchCV.

6. Choice of Parameters

6.1. SVM

The parameters that I considered for this analysis in SVM are the choice of kernel, C, gamma and degree. This choice is made as per the general terms that these factors influence the performance of the model at the best. C parameter has the ability to define a penalty for misclassified points. It means that the smaller C values can result in lower penalty which results in larger margin for the decision boundary and many misclassifications are allowed. The parameter of gamma is dependent on the influence of distance between data points. With a lower gamma value, it increases the radius of the group fitting more points than usual. On the other hand, larger gamma value considers only the points which are very close under a class. This defines that the higher value of gamma leads to overfitting problem (Yildirim, 2020). A kernel functions works on the principle of converting the data points into a higher dimensional space resulting in taking a decision on non-linear type of points. The type of Kernel considered are the Linear function, Polynomial Function, Sigmoid Function, RBF (Radial Basis Function). Linear function checks if the data is linearly separable while polynomial function is representative of the feature vector in the training sample in a polynomial feature space. Sigmoid function which works as a two-layer perceptron model where an activation function like this will determine the output. RBF works by adding radial base to the data for dimensional transformation. The degree for the polynomial function can liaise the decision boundary.

6.2. MLP

For MLP, the choice of parameters is bulk, and I have chosen to pick few important of them. They are learning rate, Optimizer, Batch Size, Epoch count, Hidden Node Size and dropout percentage. While searching for the best parameters, 3-fold cross validation is performed to get better parameter choices. With the final choices in place, the best model is built. Even different callbacks were tested and early stopping has been used. Various Activation function choices were explored while and the one which is widely used is considered to be best.

7. Analysis and Critical Evaluation

7.1. SVM

The base model with no parameter selection on the scaled data has good accuracy level of 66.54%. This is because of the default parameters in the SVM model which has proven to perform well with binary classification. While the use of cross validation hasn't altered much of the evaluation methodology. However, the choice of parameters for the RandomizedSearchCV and GridSearchCV has varied results. With the primary method, the parameter values with all the kernels in consideration made the model to converge at a longer period than expected. Hence, I have divided the kernels and then performed the optimization. Most of the kernels have almost equal performance on the data with the provided set of other parameters. The lower c value between 1 and 2 were resulting in good prediction. However False Positive rate of this model is very high for all the parameters. This is because of the penalty given by the Box constant C for incorrect classification. This can be reduced either by balancing the data samples or by assigning weights initial to the classes. However, GridSearchCV has resulted in parameters which has increased the performance as polynomial function with the degree of 4 was able to train the data at its best than any other combination. It was observed that the training of the data is good with other parameters but the performance on the test set is horrible. The problem of overfitting hasn't eliminated completely even with the cross-validation method. The Area Under Curve hasn't changed much within the analysis as both RBF kernel from default and the polynomial function from optimization has similar approach as shown in Figure 5.

7.2. MLP

When building the initial model, I started with 100 hidden layer nodes as a start and built the network with two hidden layers. It turned out that these values performed well with the data. Few callback options such as Early stopping has the ability to prevent overfitting. This has increased our performance and simultaneously cross validation of 10 folds has resulted in boost for the performance. Both the search methods as above has given good results in determining the learning rete, hidden size of the network, optimizer to use and dropout method. GridSearchCV has been the best performer in choosing the values which suits this model as it iterates through every combination of the parameter and it takes the longest time to produce results. On the contrary to SVM performance, MLP model has very a smaller number of True Negatives as shown in Figure 4. This might be due to the weights assigned to the nodes within the hidden layer. Huge chances that the back propagation for very low number of samples has resulted in misclassification. With only slight increase in the performance has dramatically increased the value of Area Under the Curve due to good number of True Negatives. The performance on the test data set has varied as MLP has acquired good accuracy score than the SVM. But it got defeated with the value of Area Under the curve.

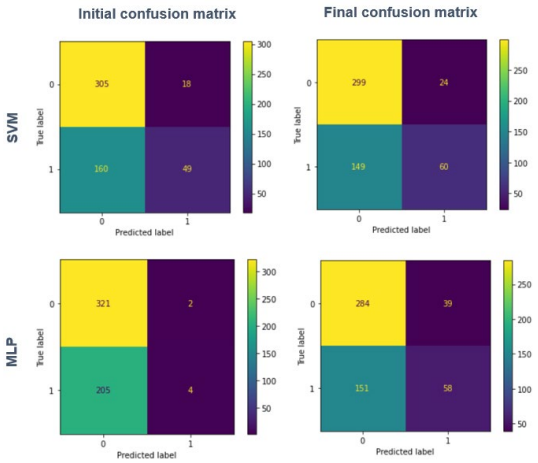

Figure 3: Confusion matrices

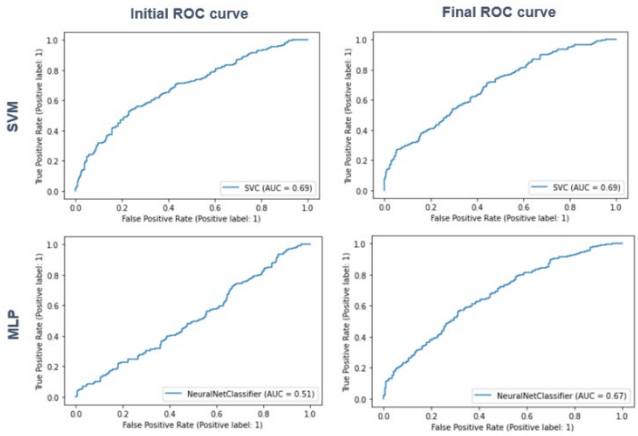

Figure 3: ROC curves

8. Conclusion

With this analysis, it has observed that the SVM model has good performance than the MLP. On the contrary the hyper parameter tuning has significant effect on MLP than on the initial model. SVM has performed well in predicting the True Negatives more accurately than another model. Never had a chance for the model to perform well on the testing set than on the training set. This conclude the limitations in the dataset values and needs to be worked upon to reduce the over fitting. From the hypothesis, it has proven that the SVM performed well. To continue further in the analysis the approaches like more hidden layers and various callback methods can be implemented. SMOTE technique can improve the results with proper balance in the dataset and considering other loss functions and metrics can vary MLP performance. Observing the parameters of SVM with wide range of values can obtain good choices.

9. References

1. Bento, C., 2021. *towardsdatascience.* [Online]
   Available at: https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analysis-cb408ee93141
   [Accessed 21 September 2021].
2. Gandhi, R., 2018. *towardsdatascience.* [Online]
   Available at: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
   [Accessed 7 June 2018].
3. kadiwal, A., 2021. *Kaggle.* [Online]
   Available at: https://www.kaggle.com/datasets/adityakadiwal/water-potability
   [Accessed 2021].
4. McGregor, M., 2020. *FreeCodeCamp.* [Online]
   Available at: https://www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the-support-vector-machine-algorithm-explained-with-code-examples/
   [Accessed 1 July 2020].
5. Ranjan, C., 2019. *Towards Data Science.* [Online]
   Available at: https://towardsdatascience.com/17-rules-of-thumb-for-building-a-neural-network-93356f9930af
   [Accessed 23 July 2019].
6. Shulga, D., 2018. *Towards Data Science.* [Online]
   Available at: https://towardsdatascience.com/5-reasons-why-you-should-use-cross-validation-in-your-data-science-project-8163311a1e79
   [Accessed 27 September 2018].
7. Sunil, 2017. *Analytics Vidhya.* [Online]
   Available at: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/)
   [Accessed 13 September 2017].
8. Torino, B., 2020. *towardsdatascience.* [Online]
   Available at: https://towardsdatascience.com/gridsearchcv-or-randomsearchcv-5aa4acf5348c
   [Accessed 29 November 2020].
9. Wilson, J., 2020. *IT-QA.com.* [Online]
   Available at: https://it-qa.com/is-svm-better-than-mlp/
   [Accessed 7 July 2020].
10. Yildirim, S., 2020. *Towards Data Science.* [Online]
    Available at: https://towardsdatascience.com/hyperparameter-tuning-for-support-vector-machines-c-and-gamma-parameters-6a5097416167
    [Accessed 1 June 2020].

Appendix I – Glossary

Table 1: Glossary

S.no	Term	Definition
1.	SMOTE	A statistical technique for increasing the number of data points in a balanced way
2.	z-score	Outlier detection using standard deviation from the mean
3.	SVM	Support Vector Machine is a supervised machine learning model with associated learning methods
4.	Fit	How well the model generalise on similar data.
5.	Predict	Output of the model after training
6.	Classification Report	Measures the quality of predictions
7.	Confusion Matrix	Performance of Classifier on test data
8.	ROC	Performance of model at all thresholds
9.	AUC	Area under the ROC Curve
10.	Cross Validate	Statistical analysis generalizes to dataset
11.	RandomizedSearchCV	Checks the best parameters Randomly
12.	GridSearchCV	Checks the best parameters with all combinations
13.	Kernel	It allows to apply linear classifier on complex problems
14.	Gamma	Number of points that needs to be grouped
15.	C	Chance of avoiding misclassification
16.	Degree	Degree of polynomial function
17.	Dropout	It avoids overfitting
18.	Activation	Transform the output signal
19.	SoftMax	Convert vector of numbers to probabilities
20.	Early Stopping	Avoid overfitting
21.	Cross Entropy Loss	Measures the performance of the model
22.	Epoch	Times the algorithm works

Appendix 2 – Implementation details

Table 2: Model Measurements

			Training Value	Testing Value	Time
SVM	Base SVM		73.09	66.54	
	Random CV	RBF	75.87	66.16	4.44s
		POLY	74.64	67.48	3.44s
	Grid Search CV	RBF	75.25	65.78	32.2s
		POLY	74.64	67.48	36.1s
MLP	Base MLP		62.21	61.09	
	Base MLP with Early Stopping		66.67	65.22	
	Base MLP with Cross Validation		65.52		
	Random CV		66.51	64.12	1 min 30 s
	Grid Search CV		67.16	64.28	12 min 15 s

The following Table 2 shows the performance of models throughout the analysis. First the data is tested on both of the base models. Then as per the model different techniques has been used to check if the performance can be increased without touching the parameters. Then RandomizedSearchCV and GridSearchCV has been performed on this model. For SVM, it was difficult for all the kernels at a time to find the optimal boundaries and is time consuming. Hence SVM is further divided into two sectors. The time for each of the process has been noted and it can be used to check how long does it take for the model to determine the optimal values. From an overview perspective, we can see that the SVM measurements haven't varied much but for MLP it has increased quite decently.