Exploring Statistical and Machine Learning Methods for Predicting Water Safety

Ben Sunshine

Statistical & Machine Learning

Dr. Schuckers

5/7/2023

Introduction

Access to clean drinking water is critical for maintaining human health and wellbeing, as it helps prevent the spread of waterborne diseases and supports bodily functions. However, in many parts of the world, access to clean water remains a significant challenge, and waterborne diseases are a leading cause of illness and death. Exposure to water contaminants such as heavy metals and chemicals can also have serious health effects, exacerbating the problem further. Communities without access to clean water face a multitude of challenges, including higher rates of illness and reduced productivity. As such, organizations are investing in technology companies that are developing machine learning models to detect water supply contamination and help communities respond more quickly. For example, Microsoft has implemented convolutional neural networks to analyze water samples for contamination in their "Clean Water AI" IoT devices. These devices will be used by cities to continuously monitor water sources, identifying, and responding to contamination quickly. By improving access to clean water and developing accurate methods for assessing water safety, we can promote sustainable development and improve the health and wellbeing of communities worldwide. This project aims to support this effort by utilizing multiple statistical and machine learning methods to make accurate predictions about water safety. By developing reliable models to detect contamination in water samples, we can help ensure communities have access to clean drinking water, reduce the risk of illness, and promote economic and social development.

The dataset used in this project was sourced from Kaggle and consists of 8,000 water samples, each characterized by the concentration of 20 different contaminants including aluminium, ammonia, arsenic, barium, cadmium, chloramine, chromium, copper, fluoride, bacteria, viruses, lead, nitrates, nitrites, mercury, perchlorate, radium, selenium, silver, and uranium. The dataset also includes a binary target variable called 'is_safe', which indicates whether a water sample is considered safe (labeled as '1') or unsafe (labeled as '0'). During the initial data analysis, only three observations were found to be missing their 'ammonia' values. As the proportion of missing data was negligible, these instances were removed from the dataset instead of performing imputation on the missing values.

In this report, three supervised modeling methods were implemented for predicting the target variable, 'is_safe', in water samples. These methods include linear discriminant analysis (LDA), a Random Forests, and a Deep Neural Network (DNN). After applying 5-fold cross validation to each method, the Deep Neural Network produced the highest mean accuracy (0.9632) followed by the Random Forest (0.9504) and LDA (0.8968). Similarly, the standard deviation of the accuracies was lowest for the Random Forest method (0.0026) followed by LDA (0.0027) and the Deep Neural Network (0.0051). These findings suggest that the Random Forest and Deep Neural Network methods are more effective for predicting water safety levels than LDA.

Explanation of Methodologies

To ensure reliable estimates of the models' performance on new data, all modeling methods used in this project were subject to 5-fold cross validation. K-fold cross validation is a widely used technique in machine learning, which involves dividing the dataset into k subsets of roughly equal size. In each iteration of the validation process, the model is trained on k-1 of the subsets and tested on the remaining subset. This process is repeated k times, with each subset serving as a validation set exactly once. In this project, five folds were chosen for cross validation, as this allowed each model to be trained on 80% of the data, leaving the remaining 20% for validation. This approach helps to reduce overfitting and provides a more reliable estimate of the models' performance on unseen data. By performing 5-fold cross validation, we can be more confident in the accuracy of our models and their ability to generalize to new data.

As the first step in solving the classification problem, I utilized Linear Discriminant Analysis (LDA), a technique that helps in dimensionality reduction and identification of a vector that maximizes the separation between classes. Unlike methods such as logistic regression, LDA can handle high-dimensional models. However, it is sensitive to outliers and requires independent predictors that follow a normal distribution with equal variance across classes.

A Random Forest was the second machine learning algorithm I used to predict the safety level of water samples. Random Forests are an excellent method for regression and classification problems because they are known to produce highly accurate results by averaging the results of multiple decision trees. They also provide easy to tune hyperparameters and great insight regarding feature importance, allowing for simple feature selection to improve the model. On the other hand, Random Forests may overfit data if unimportant features are included and can be computationally expensive. In addition, while Random Forests provide insight into feature importance, they lack interpretability compared to simpler models.

The last method I used for this classification problem was implementing a Deep Neural Network (DNN). This type of Artificial Neural Network has multiple hidden layers between the input and output layers. These networks are great for classification and regression problems. Some of their advantages include their ability to fit not linear data, work in diverse applications (i.e., image processing and language processing), and their ability to use to be customized and fine-tuned by the user. However, like Random Forests they can be computationally expensive and difficult to interpret. Also, to avoid overfitting DNN require careful regularization.

Linear Discriminant Analysis Results

When comparing the variance of the predictors between not safe (0) and safe (1) groups, using an F-test we obtain an f-statistic of 2.3215 on 19 degrees of freedom. This produces a 95% confidence interval between 0.9188838 and 5.8651936, indicating there is no significant difference in variance of the two groups, meaning LDA can be appropriately used. Additionally, a correlation plot was created to ensure independence of features as seen in figure 1. The LDA model achieved an average accuracy of 0.8961 with a low standard deviation of 0.0027 across the 5-fold cross-validation. This accuracy was achieved by using all 20 of the features in the dataset. The accuracy ranged from 0.8881 to 0.9019, indicating consistent performance of the model. Visual analysis of the model plots reveals that LDA effectively separates the classes into normal distributions. Figure 2 shows an example of the probability distributions of the groups from the first fold of the cross-validation. Upon inspecting the coefficients of the models, mercury had the largest weight and the highest variability among all the features (figure 4).

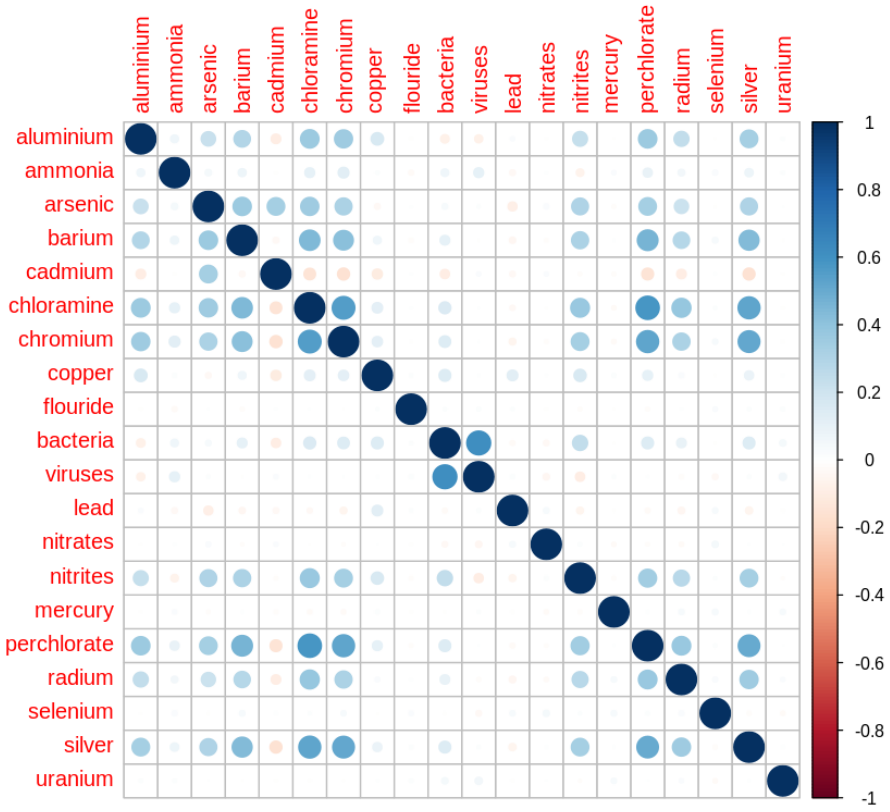

Figure 1

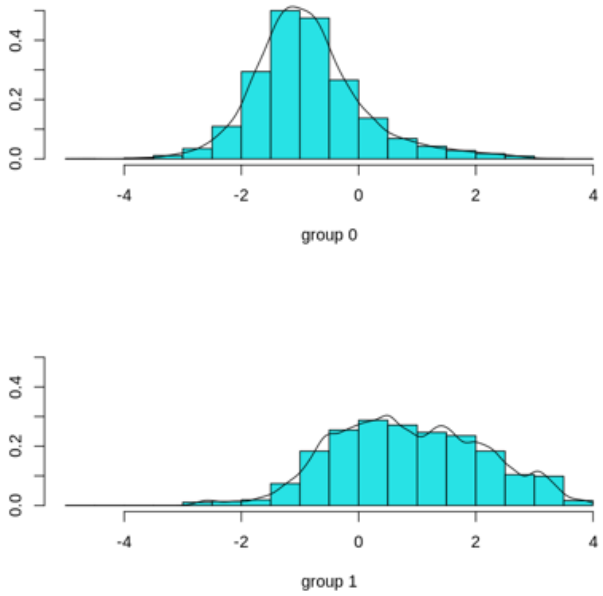

Figure 2

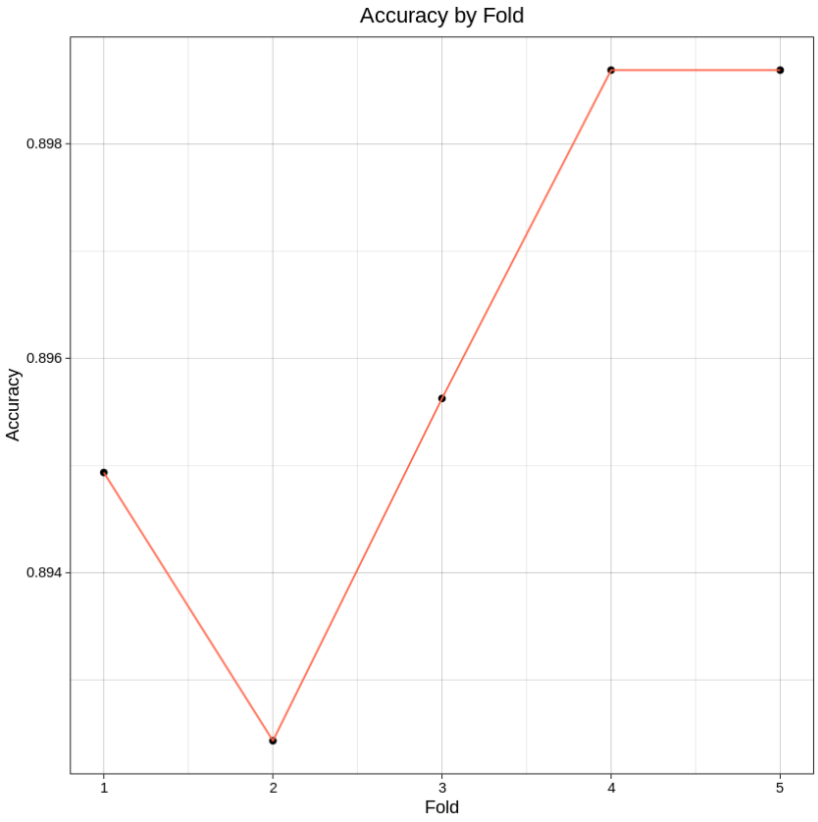

Figure 3

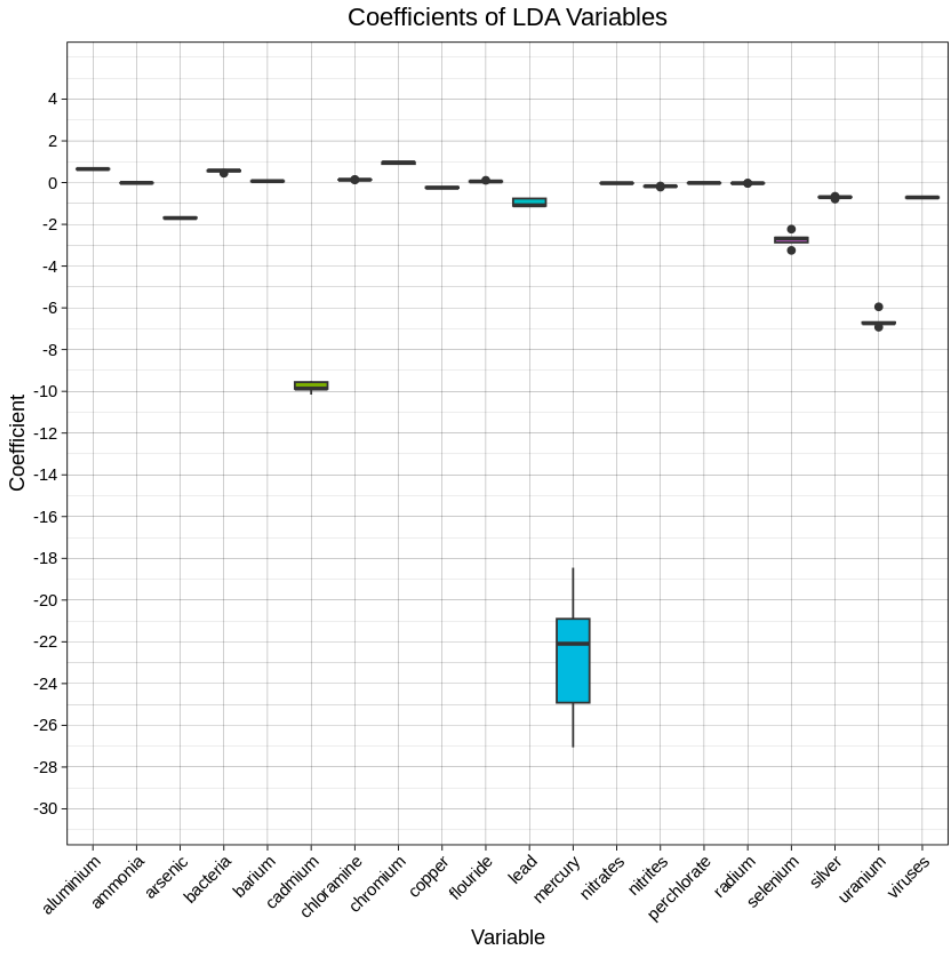
Coefficients of LDA Variables
Coefficient
4
2
0
-2
-4
-6
-8
-10
-12
-14
-16
-18
-20
-22
-24
-26
-28
-30
aluminium
ammonia
arsenic
bacteria
barium
cadmium
chloramine
chromium
copper
flouride
lead
mercury
nitrates
nitrites
perchlorate
radium
selenium
silver
uranium
viruses
Variable

Figure 4

Random Forest Results

Feature selection is a critical aspect of fitting Random Forest models to ensure optimal performance. After an initial Random Forest model was fitted, 10 features were selected based on their higher importance scores. These features include aluminium, ammonia, cadmium, bacteria, viruses, nitrates, perchlorate, radium, silver, and uranium. The final Random Forest model utilized 2,500 trees, randomly sampled 10 features at each split, and had a shrinkage of 0.1. The model achieved an average accuracy of 0.9504 across 5-folds of cross-validation, with accuracy scores ranging from 0.9461 to 0.9531. These accuracies were relatively consistent with a standard error of 0.0026, outperforming LDA. The false positive rate, where a water sample was misclassified as safe, had an average of 0.0127, while the false negative rate, where a water sample was misclassified as unsafe, had an average of 0.3312. Also, cadmium, aluminium, and perchlorate were identified as the three most important features with the highest mean decrease in accuracy and mean decrease in Gini-Impurity, as shown in figure 6 and 7.

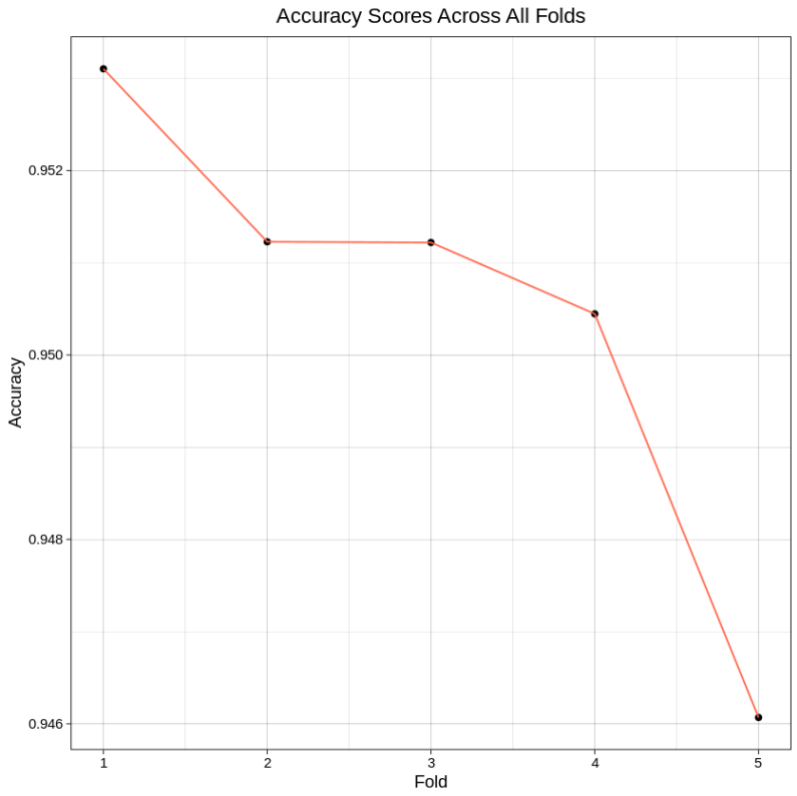

Figure 5

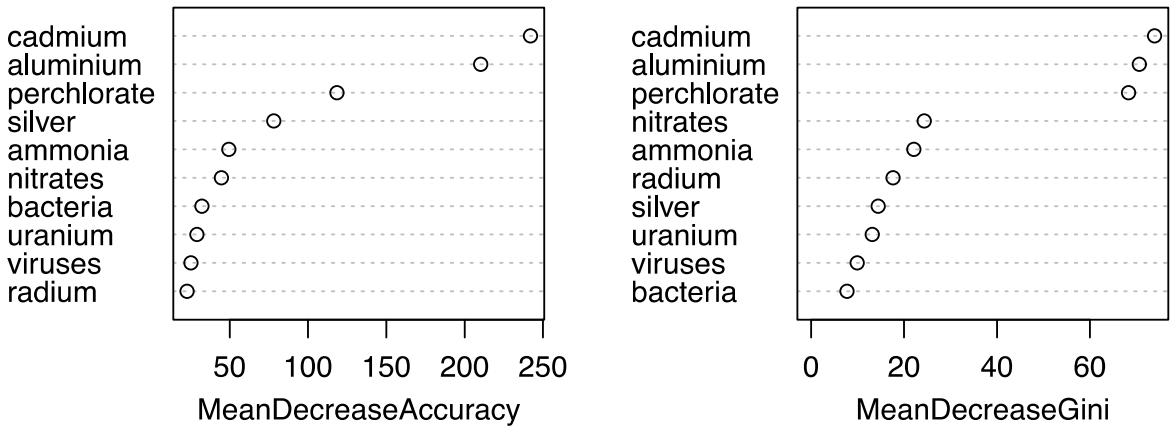

Figure 6

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
aluminium	120.61331	270.758501	210.21016	70.628167
ammonia	47.15620	18.273604	49.40434	22.109464
cadmium	76.12566	323.286655	242.08286	73.943091
bacteria	26.98547	19.152148	32.21502	7.730313
viruses	21.34564	19.141535	25.21572	9.922263
nitrates	28.49397	46.363370	44.64109	24.337416
perchlorate	110.84951	40.080007	118.49040	68.337510
radium	22.69006	3.578065	22.78500	17.614873
silver	78.07073	16.179000	78.13803	14.408982
uranium	18.03439	30.599348	29.11301	13.136191

Figure 7

Deep Neural Network Results

In this study, a deep neural network with a sequential model consisting of three hidden layers was implemented. 20 neurons made up the input layer, consisting of each of the features from the dataset. The first hidden layer had 64 neurons, the second had 32, and the third had 8 neurons, all with ReLu activation functions. This structure has been visualized in figure 8. To improve the model's generalization, batch normalization and dropout (with a 20% probability) were applied after each hidden layer, as seen in the model's architecture in figure 9. Because this is a binary classification problem, the output layer contained a single neuron with a sigmoid activation function, outputting the probability of an instance belonging to a certain class. The model was compiled with the Adam optimizer, starting with a learning rate of 0.01, but later relying on a learning rate schedule function. This function set the learning rate to 0.01 if the current epoch was below the 15th epoch, reduced it to 0.001 after the 15th epoch, and then set it to 0.0001 after the 35th epoch. Additionally, early stopping was used as an additional callback to prevent overfitting of the training data in the event the validation loss didn't improve over 25 epochs. The 5-fold cross-validation results of the model showed an average validation accuracy of 0.9632 with a standard deviation of 0.004 between accuracies, and an average validation loss of 0.0963 with a standard deviation of 0.009 between losses. The training/validation accuracy and loss histories are shown in figures 10 and 11, respectively.

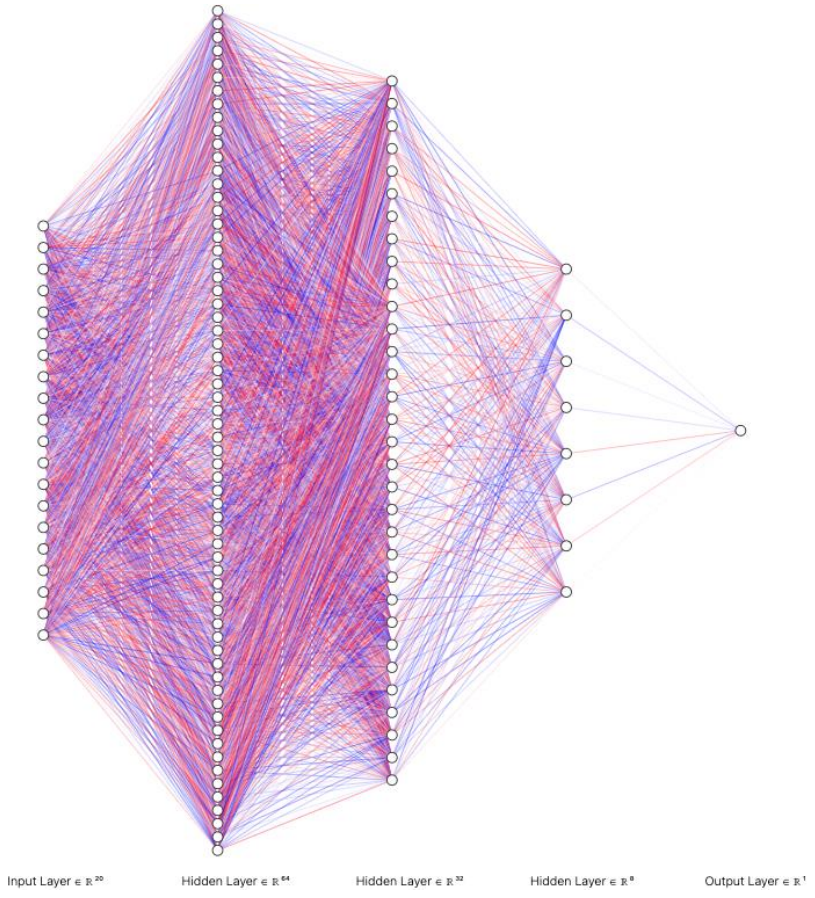

Figure 8

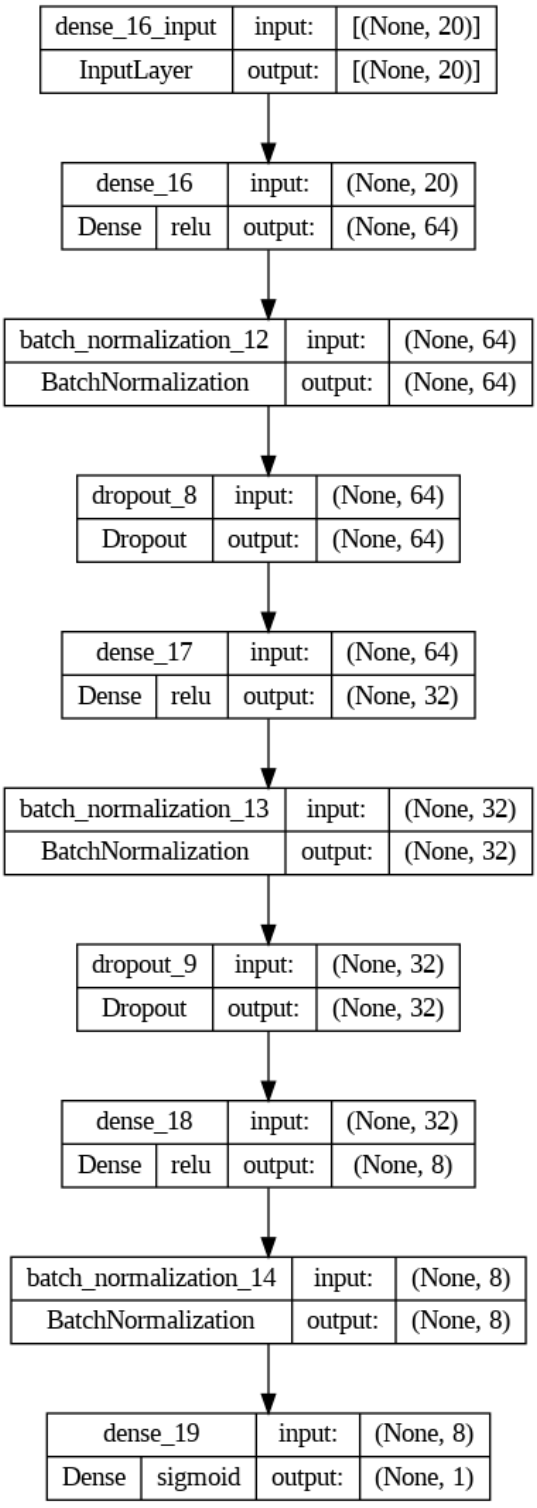
dense_16_input | input: | [(None, 20)]
InputLayer | output: | [(None, 20)]
dense_16 | input: | (None, 20)
Dense | relu | output: | (None, 64)
batch_normalization_12 | input: | (None, 64)
BatchNormalization | output: | (None, 64)
dropout_8 | input: | (None, 64)
Dropout | output: | (None, 64)
dense_17 | input: | (None, 64)
Dense | relu | output: | (None, 32)
batch_normalization_13 | input: | (None, 32)
BatchNormalization | output: | (None, 32)
dropout_9 | input: | (None, 32)
Dropout | output: | (None, 32)
dense_18 | input: | (None, 32)
Dense | relu | output: | (None, 8)
batch_normalization_14 | input: | (None, 8)
BatchNormalization | output: | (None, 8)
dense_19 | input: | (None, 8)
Dense | sigmoid | output: | (None, 1)

Figure 9

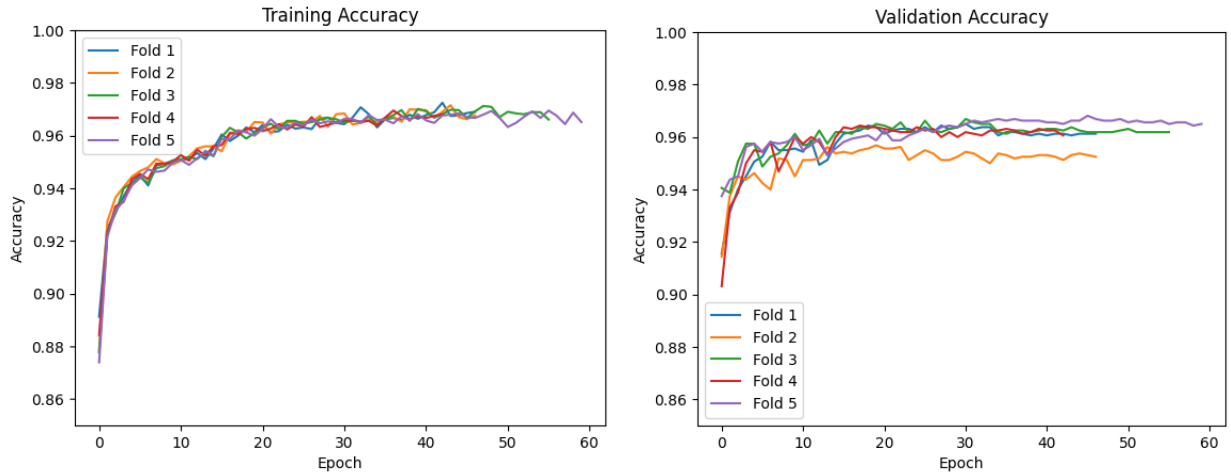

Figure 10

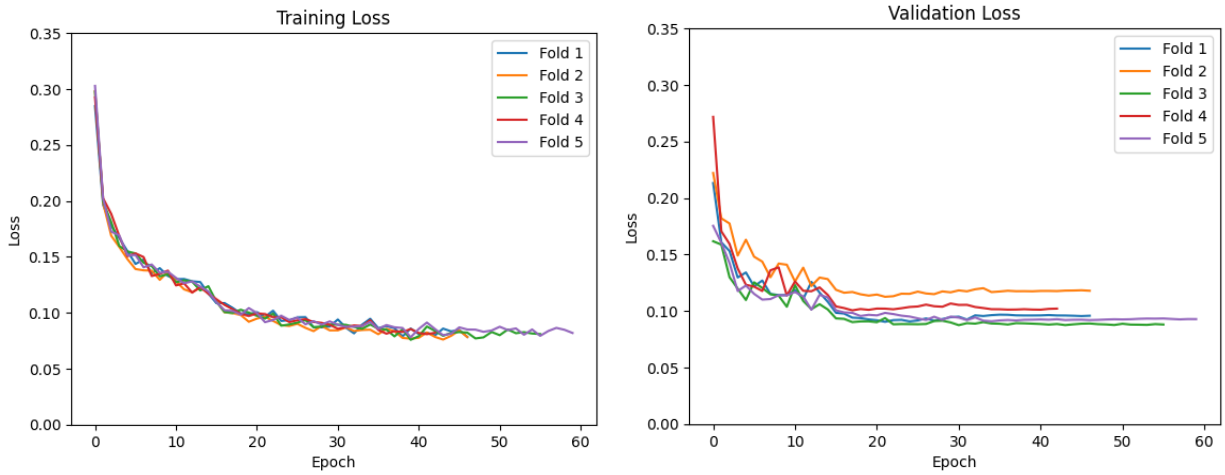

Figure 11

Conclusions

The use of statistical and machine learning methods in predicting water safety is a critical aspect in promoting health, sustainability, and the well-being of communities worldwide. In this study, I explored three major algorithms, namely Linear Discriminant Analysis, Random Forests, and Dense Neural Networks to predict the safety of water samples. To ensure that the models would generalize well, I used 5-fold cross-validation to evaluate their performance on unseen data. The results showed that the Dense Neural Network had the highest mean accuracy of 0.9632, followed by the Random Forest with a mean accuracy of 0.9504, while the Linear Discriminant Analysis had a mean accuracy of 0.8968. However, the Dense Neural Network had the highest standard deviation of accuracies (0.0051), followed by LDA (0.0027) and the Random Forest (0.0026), respectively. From the results, both the Dense Neural Network and the Random Forest performed exceptionally well in classifying water samples. However, it is not an easy decision to determine which model performed better since the DNN had a higher mean accuracy but more variation between accuracies. On the other hand, the Random Forest had a slightly lower mean accuracy but had lower variation between accuracies. The takeaway from this project is that the choice of the best model to use in predicting water safety should depend on the available data. Random Forest models may perform better when fewer variables are available in another data set, while the Dense Neural Network may achieve better accuracy with datasets containing more information. Therefore, it is important to consider all models and select the one that performs best for the given dataset and the goals of the study.