|  |  |
| --- | --- |
| **4DATA 430 Technical Report Assignment 1 (a & b): Logistic Regression** | **Monique Reed** |
| **Water Quality Analysis** | |
| **URL to dataset:** https://www.kaggle.com/datasets/mssmartypants/water-quality | |

**Assignment 1a (due Week 2):** you should complete the following sections ONLY:

* Overview (Problem Domain)
* Overview (Objective)
* Analysis (Exploratory Analysis)

**Assignment 1b (due Week 3):** all sections of this template should be completed. Modifications of the three sections submitted in Assignment 1a should be made based on feedback from the instructor.

This template should be used in conjunction with the assignment instructions. The size of the text area below will expand to the length of your response; the area should not be interpreted as a required or suggested length of response. Responses within the text area should be single spaced with Times New Roman 12pt font. The body of the document will likely be 6-9 pages, not including the Appendix; length may vary depending on the specifics of the analysis and the dataset. As needed, APA format in-text citations should be included, along with a full references list at the end of the document.

|  |
| --- |
| **Overview** |
| **Problem Domain**: give some background and context about the problem domain (application area). For instance, if you are doing the analysis for predicting heart disease, provide some context about the disease and include some interesting statistics about it. Also, discuss how the method is relevant for the chosen problem. |
| Analyzing whether or not water is safe based upon the quality of the water is extremely important. Water pollution can occur quite easily from multiple sources such as industrial waste, oil pollution, plastic pollution, wastewater, pesticide and water runoff, and radioactive pollution, but the effects of such are disastrous in consumers. In fact, “Drinking water that contains hazardous micro-organisms can spread diarrhea and diseases such as cholera. Every year, polluted drinking water causes 485,ooo diarrhea-related deaths (Developmentaid, 2024).” By logistic regression and the proper use of the appropriate dataset, a conclusion could be made whether or not a certain volume of water is safe to drink by the amount of contaminants in the water.  References:  Developmentaid. (2024, February 27). *Water pollution in the world: Major causes and statistics*. DevelopmentAid. https://www.developmentaid.org/news-stream/post/152754/water-pollution-in-the-world |
| **Objective**: clearly state the objective of the analysis in relation to the kind of algorithm you are employing. Use specific language as to what question(s) you are trying to answer using the specific analysis/modeling type. |
| Using a logistic regression model, we are trying to predict whether water is safe to drink will be determined by the number of contaminants found in the water. Which contaminants and what number of contaminants will deem water as unsafe to drink or unusable for other purposes? How can we predict whether or not the water is safe to consume? |
| **Analysis** |
| **Exploratory Analysis**: describe the data including the source, the collection method, and variables. Perform exploratory analysis. Also, select few key variables (including the target variable for supervised learning) and study their distributions using plots such as histograms, box plot, bar chart, etc. |
| The data was downloaded from Kaggle.com and imported into VS Code for analyzation. The data is false imaginary data to mimic the water quality of an urban area. By a prompt of commands, it was determined that the data contained 7999 rows and 21 columns. A myriad of variables such as radium, silver, viruses, bacteria, and Is\_safe, the primary target variable, were analyzed. Below are images of analysis. |
| **Preprocessing**: armed with the exploratory analysis, perform the necessary preprocessing, both general and specific types appropriate for the modeling type being employed. |
| The preprocessing stage needed to be done before creating a proper heatmap. Data cleaning is one of the most essential pieces of the preprocessing stage. And the only strings that can be even remotely around a heatmap are the labels themselves. So, in box 6 of the Jupyter Notebook files, the string that characterizes null values, “#NUM!”, was turned into a standard null value and then dropped. After that, further examination was conducted to count the remaining null values and understand how many duplicated values were made. |
| **Model Fitting**: explain the key steps and activities you perform to fit the model. Experiment (as appropriate) with parameters tuning. This is key, what separates highly accurate model from a less accurate ones is the amount of performance tuning performed. |
| To fit the model, the first thing that was done was a test-train split, the separation of testing and training data first starting with the fundamental X and Y values that were to be used in the Logistic regression model; from there, we utilized the test\_train\_split function, using 80% of training data and 20% testing data. Moreover, strictly for sanity check purposes, we print the training and testing set data sizes. A partially unnecessary portion of this was feature scaling; even though we didn't need it, it's a widespread machine-learning technique that could have a good habit of utilizing. Feature scaling is mainly used when the data isn't visually proportional. Finally, the model fitting begins after the logistic regressing function is initialized, and the training data is used to fit. |
| **Results** |
| **Model Properties:** explain the components of the fitted model and their characteristics. Leverage functions to summarize the model properties. Also, leverage visualization as required. |
| After the model for fitted, we began to pull some metrics from it, an accuracy score of the model was given of 90% and a confusion matrix printed as follows:  [[1389 30]  [ 123 58]] |
| **Output Interpretation**: explain the result and interpret the final model output using terms that reflect the application area and in relation to the stated objective. This is where you check whether or not the stated objective is met. |
| The main objective of this project was to determine whether or not water was safe to drink based on the amount of substances found in the water. Based on the model created, I am optimistic that evaluating the contaminants in the water is essential. |
| **Evaluation**: employ appropriate metrics to quantitatively evaluate the performance of the fitted model. For supervised classification, this includes simple accuracy, precision & recall (or sensitivity & specificity), all of which can be generated from a confusion matrix, or ROC. |
| Given the presented array of data, the model seemed to perform exceptionally well. Even though the given amount of false negatives was low for the given situation, the number of true positives is statistically significant. |
| **Conclusion** |
| **Summary**: highlight the main findings in relation to the stated objective. You don’t need to discuss the details of the analysis and the model such as accuracy here, just focus on the key findings. |
| Based on the model, the amount of contaminant found in water can be used to determine whether the water is safe to drink. |
| **Limitations & Improvement areas**: discuss the limitations of the analysis and identify potential improvement areas for future work. This could be related to the data, algorithm, or a combination of the two. |
| For future work, there should be a refined number of features included in the model; there has to be at least one specific ant contaminant that has a closer correlation to the safety of the water than others. That feature could be helpful in the future. |

|  |
| --- |
| **Appendix** |
|  |

**References**