|  |  |
| --- | --- |
| **DATA 430 Technical Report Assignment 2: Bayesian Classification** | **Monique Reed** |
| **Naïve Bayes: Water Quality Data** | |
| **URL to dataset: https://www.kaggle.com/datasets/mssmartypants/water-quality** | |

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 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 485,000 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 number 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. |
| The objective of the following analysis is to build a Gaussian Naïve Bayes Classifier model to predict whether a specific sample of water is safe to drink based upon the measure of contaminants or contents found in that water. |
| **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. |
| Some preprocessing steps had to be done before exploratory data visualization. Being that it is one of the most essential pieces of creating a successful machine learning algorithm. In code box 3 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 properly fit the model, one of the first things that needs to be done is the even splitting of the data that is made available to us. Utilizing the common test/train/split built in python function. With the training set consisting of 80% of the data and the testing set consisting of the remaining 20%. One thing that is new here is the usage of random state. Random\_ states are built in functions to randomly shuffle the data. Then, feature scaling is preformed to make sure that the data doesn’t seem improperly skewed. From there, a classifier model is initialized, and the model is 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 was fitted, it is vital to pull metrics from it to determine whether the analysis done was accurate. The accuracy score of about 83%. Thankfully, python has some built in functions that allows for the validation of model accuracy. The confusion matrix for the model is as follows:  A screenshot of a computer  Description automatically generated |
| **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 objective of this model was to classify whether water is safe to drink based upon the contaminants presented in the water. Based upon the results given, I am positive that evaluating such features gives a positive result. |
| **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 vast array of features in the model, the classification model seemed to do impressive, the number of true positives is comforting to see in such model. The image below is the confusion matrix for the following model: |
| **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. |
| The classification model has determined the number of contaminants found in the water can be corelated to 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. |
| Looking into once or two specific features that have a correlation to the contaminants or even the addition to the data of the location of the contaminants or bodies of water could be beneficial for future analysis. |

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

**References**

Debroy, S. (2023, February 9). *Naïve Bayes classification in Python*. Medium. https://medium.com/@shuv.sdr/na%C3%AFve-bayes-classification-in-python-f869c2e0dbf1

prashant111. (2020, August 28). *Naive Bayes classifier in Python*. Kaggle. https://www.kaggle.com/code/prashant111/naive-bayes-classifier-in-python#9.-Split-data-into-separate-training-and-test-set-