Project Number 2
Sameer Nepal
DSC 680

Report on Project 2

Topic: This project classifies whether the water is safe to drink or not depending upon the impurities present in the water. There can be many impurities like physical, chemical, biological, and radiological and we will mainly focus on the chemical impurities in the water.

Business Problem: Drinking water quality is very important for healthy body and tracking the contamination helps us in preventing the use of those contaminated water. In the world, there has been issue on access to clean drinking water. The World Health Organization has reported that approximately 10% of the global population lacks access to improved drinking water sources. One of the United Nations Sustainable Development Goals aims to achieve universal access to water and sanitation by 2030 [Appendix A]. The model that will be generated using this project will classify whether the water is safe to drink or not. This project will help in making the community healthy and provide safe drinking to the public.

Background/History: Almost 10% of the population in the world do not have access to clean drinking water. The major contamination that happens with water is biological which causes waterborne diseases illness and that nearly kills one million people every year. Aside from that chemical contamination is another major reason which mostly happens in low and medium income countries because the law and order is not that strict in those countries and people are doing whatever they want for money. Being exposed to chemicals present in drinking water can result in various chronic illnesses, such as cancer and cardiovascular disease, as well as negative reproductive outcomes and health impacts on children, including neurodevelopmental issues. This model will look into the data for the concentration of metals present in water and categorize whether the water is safe to drink or not.

Data Explanation: The data that I am using for this project is from Kaggle and it has information about the concentration of chemicals present in water sample. The data has 7999 rows with 21 columns. This dataset has been generated using fictional water quality data in an urban setting. The owner of this dataset suggests utilizing this dataset for educational purposes, practicing data analysis, and obtaining the requisite knowledge. This project gives the techniques on how accurate the model is. We can ingest the actual data for the concentration of chemicals in water to identify if the water is potable or not

Data link: https://www.kaggle.com/datasets/mssmartypants/water-quality

Exploratory data analysis was performed to the data to study the data. The dependent variable for our analysis is the portability of water and the independent variables are the ingredients in the water which are mostly the chemical elements. The data was explored for any possible nulls and were removed from the dataset. The dataset was imbalanced as there were 7084 records for the impurity of water and 912 for purity of water. While modeling classification we try to have equal number of positive and negative case and therefore we will be using oversampling technique to get same number of rows for both cases.

Methods: The main techniques I will be using is the classification analysis where we will predict whether the water is safe to drink or not. For the classification I have used many models and calculated the accuracy and proceed further with the one that have highest accuracy. The classification techniques that I have used are Logistic Regression, Decision tree, Random Forest, Gradient Booster and K-Neighbors.

Analysis: For the analysis of the data, I started looking at any missing values and dropped three records. Then I looked into the count of the classification of the target variable which came off to be imbalanced and hence I used the oversampling techniques to get the same number of records for two cases of classification. I also looked at the correlation between the variables to see how the data are correlated. The correlation is not very strong, but I think this is okay because having perfect correlation is biased and training the model with highly correlated data might not be the good model.

The dataset is imbalanced, and I am using RandomOverSample class to balance the target variable. This will create equal number of 1s and 0s for the target variable. If we do not do this then the date will be biased towards 0 because we had 7084 0s and 912 1s. I have compared five different models Logistic Regression, Decision tree, Random Forest, Gradient Booster and K-Neighbors.

To compare these and get good result I have use cross validated search. During this process the parameters are randomly selected and the accuracy score of these models were analyzed. The score was best for gradient boosting classifier. Then I divided the data into training and test sets in the ratio of 0.2 and fitted the data to the GradientBoostingClassifier model and calculated the accuracy score for the model which came out to be 96%

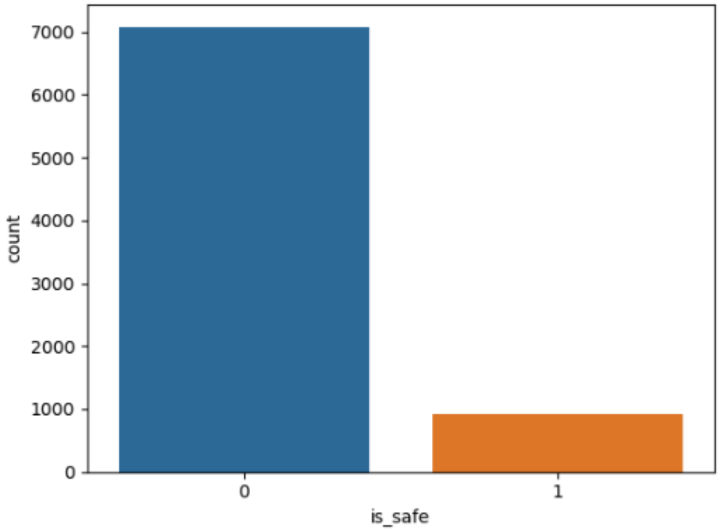

Fig 1: Imbalanced target variable

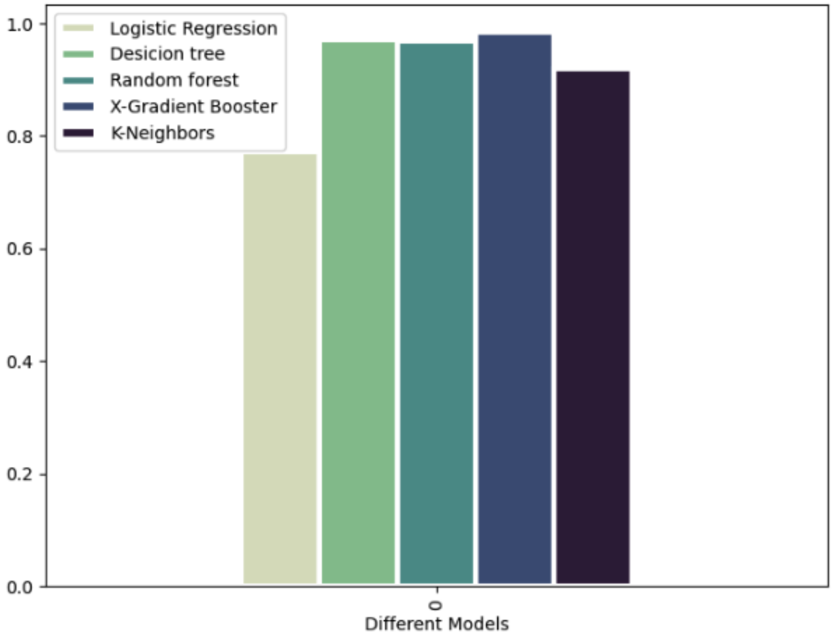

Fig 2: Score for different Model

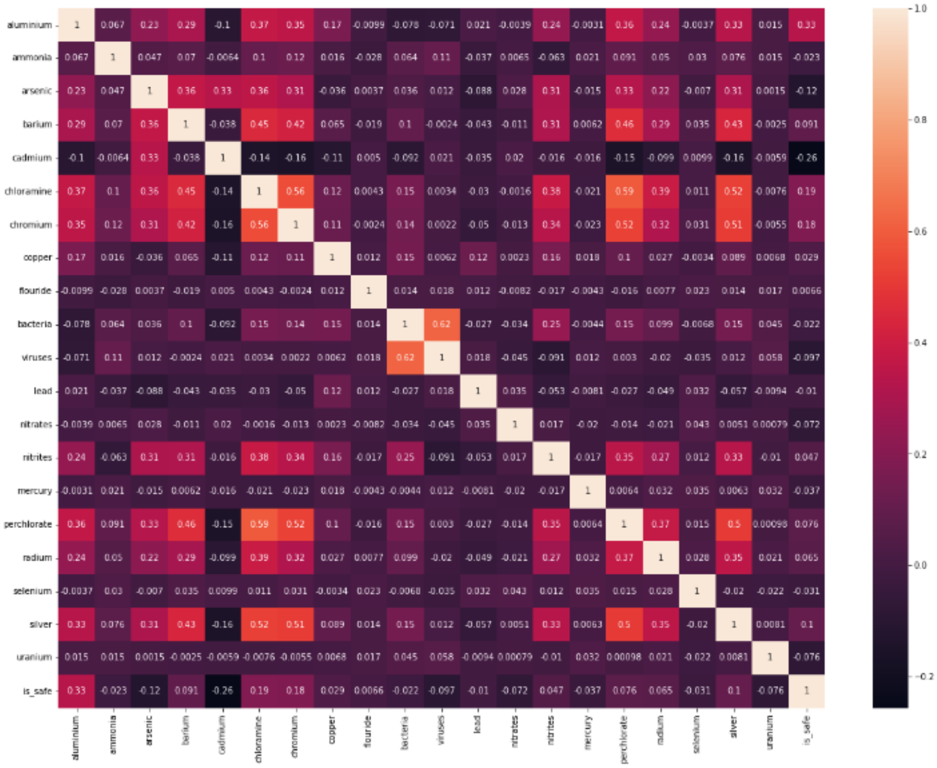

Fig 3: Heat Map for the features

Conclusion: Using the model that we developed in this project; we can predict whether the water is safe to drink or not. This will help in making our community healthy and saving life of many people. The accuracy of the model is very high which is necessary because we are dealing with the health of an individual and any negligence might severely affect an individual. This model can be used by health authorities of different countries and determine whether the water in their area is safe to drink and take necessary actions.

Assumptions: There are many assumptions that we have made while developing this model. The data that we used in this project is imaginary data, so it is assumed that the target variable i.e the safety classification of water is 100% accurate. Before using it practically we need to run the model with the actual data.

Limitations: The model is using the imaginary data and it is limited to educational purpose only. If we want to use the model practically then we need to run the model using real world data and analyze the accuracy of the model.

Challenges: The challenge for this project was to find the way to balance the target variable. When I checked for the count of the 1s and 0s of the target variable it looked very imbalanced (7080 records of 0s and 912 records of 1s) and running the model on that data would be biased towards 0s so I used RandomOverSample class to balance the target variable.

Future Use and Recommendation: The accuracy of the model is about 96%. Since the model is generated using the imaginary data, I recommend running the model with actual real world data and analyze the accuracy. The outcome of the analysis will determine whether the model should be used by different health authorities to check the portability of the water.

Implementation Plan: This model should only be implemented after we run the model using the real world data and analyze the result.

Ethical Considerations: The ethical concern that I need to be aware about is the private information of the people and the harm it can cause to the humanity. Since the data is imaginary data of water quality in an urban environment this should only be used after running the model using the real world data and analyzing the result.

References:

1. MsSmartyPants. (2021, June 30). *Water quality*. Kaggle. Retrieved April 30, 2023, from https://www.kaggle.com/datasets/mssmartypants/water-quality

2. Environmental Protection Agency. (n.d.). EPA. Retrieved April 30, 2023, from https://www.epa.gov/ccl/types-drinking-water-contaminants

3. Levallois, P., & Villanueva, C. M. (2019, February 21). *Drinking water quality and human health: An Editorial*. International journal of environmental research and public health. Retrieved April 30, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6406761/#B1-ijerph-16-00631

4. Levallois, P., & Villanueva, C. M. (2019, February 21). *Drinking water quality and human health: An Editorial*. International journal of environmental research and public health. Retrieved April 30, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6406761/

5. Püttmann, L. (2022, January 6). *Water quality dataset*. Deepnote. Retrieved April 30, 2023, from https://deepnote.com/@leonard-puttmann-a8ef/Water-quality-dataset-14a5f99f-9b81-49c9-9327-973b1c9f0b1f

6. Team, D. F. (2021, June 28). *Machine learning classification - 8 algorithms for data science aspirants*. DataFlair. Retrieved April 30, 2023, from https://data-flair.training/blogs/machine-learning-classification-algorithms/

Appendix A

Drinking Water Quality and Human Health: An Editorial

Drinking water quality is paramount for public health. Despite improvements in recent decades, access to good quality drinking water remains a critical issue. The World Health Organization estimates that almost 10% of the population in the world do not have access to improved drinking water sources [1], and one of the United Nations Sustainable Development Goals is to ensure universal access to water and sanitation by 2030 [2].