Data Presentation and Visualization

Data Science II (COSC 4337)

Dr. Ricardo Vilalta

Submitted by:

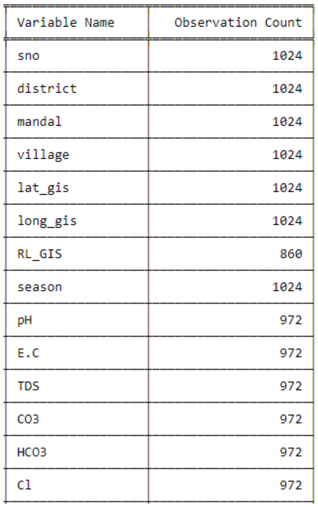
Brayan Gutierrez (1865588)

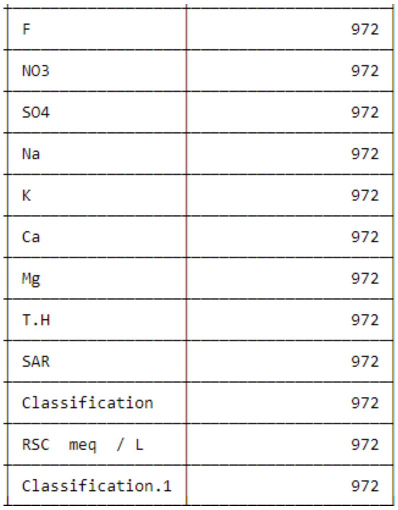
Jericka Ledezma (1968730)

Katie To (2026435)

In Telangana, India, our research team is examining the quality of groundwater across different administrative divisions, known as mandals, to ensure agricultural productivity and soil health are preserved. Utilizing data from the "Telangana Open Data Portal," our goal is to uncover potential connections between water quality metrics and their effects on irrigation. This endeavor aims to offer practical insights for farmers, agricultural experts, and policymakers to enhance water management strategies. Our main aim is to develop a thorough understanding of groundwater quality across Telangana's agricultural areas. By examining the relationships among various water quality indicators and their combined effects on soil and plant well-being, we seek to classify water sources into suitable categories for irrigation. This classification will aid in creating guidelines for agricultural groundwater usage, thus reducing risks associated with unsuitable groundwater sources around the world. All visualizations were generated in our project's working code entitled "Data Pre-Processing Assignment," "Data Modeling Assignment," and "Comparisons."

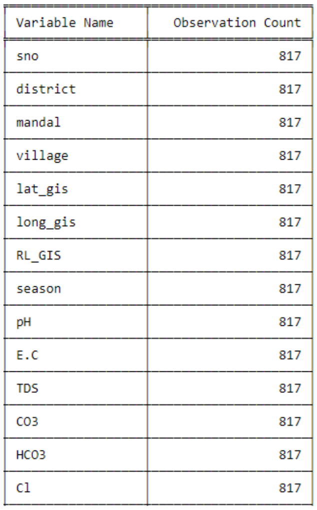
In the initial phase of our project, we began by scrubbing the data to address missing rows or observations, ensuring completeness. We also handled outlier observations that could potentially skew our classification models. Additionally, we carefully selected groundwater quality metrics to support our classification efforts. After importing the data into Python, we first decided to treat the missing observations. Below are tables containing the number of observations in each variable of our data.



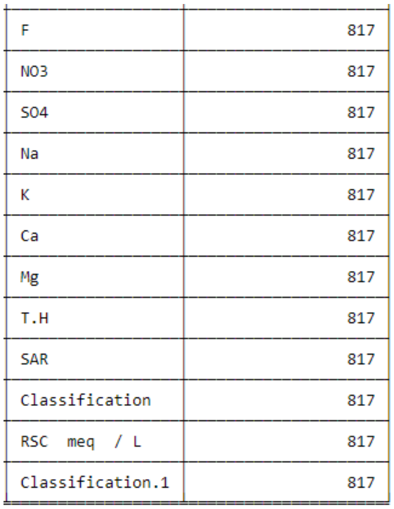


Variable Name	Observation Count
sno	1024
district	1024
mandal	1024
village	1024
lat_gis	1024
long_gis	1024
RL_GIS	860
season	1024
pH	972
E.C	972
TDS	972
CO3	972
HCO3	972
Cl	972
F	972
NO3	972
SO4	972
Na	972
K	972
Ca	972
Mg	972
T.H	972
SAR	972
Classification	972
RSC meq / L	972
Classification.1	972

Notice that the number of observations is not the same for every variable, meaning that there are observations that are missing information for some variables. To address this issue, we omitted the observations that did not have all of the needed information from the data. Below are tables containing the number of observations in each variable after omitting observations. Now notice that all of the variables have the same number of observations, so we are able to proceed to our next step, outlier treatment.

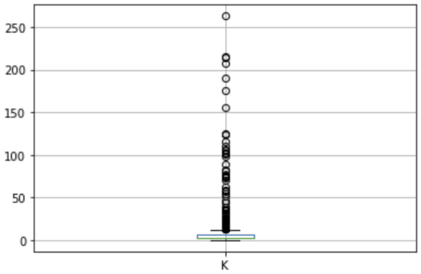


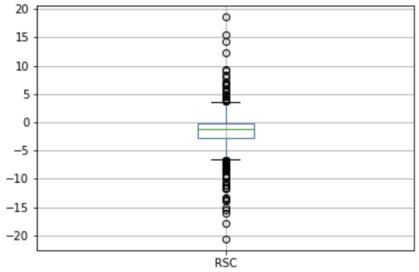
Variable Name	Observation Count
sno	817
district	817
mandal	817
village	817
lat_gis	817
long_gis	817
RL_GIS	817
season	817
pH	817
E.C	817
TDS	817
CO3	817
HCO3	817
Cl	817



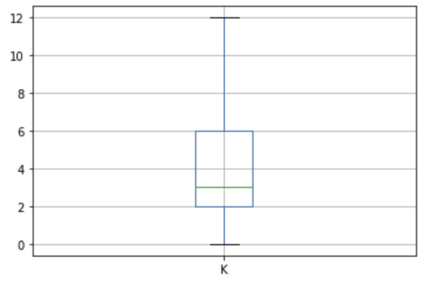
F	817
NO3	817
SO4	817
Na	817
K	817
Ca	817
Mg	817
T.H	817
SAR	817
Classification	817
RSC meq / L	817
Classification.1	817

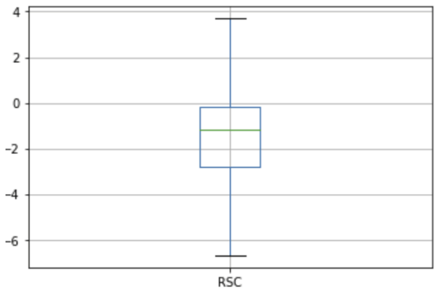
To check for any outliers present in the data, we visualized the numerical variables using a boxplot. A boxplot is a visual representation that demonstrates the spread, skew, and outliers of the data. As can be seen in a few of the numerical boxplots, there are outliers present in the data, which are represented by the individual points in the upper and lower sections of the plots. No outliers were present in the categorical variables.



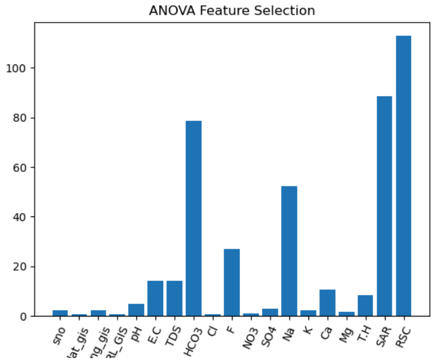


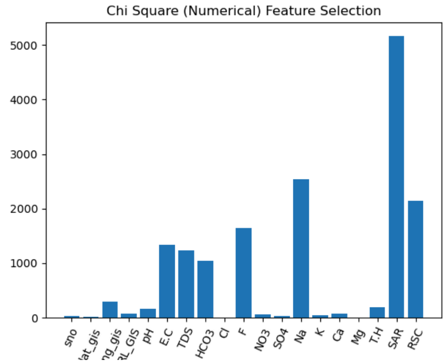
We treated these outliers in the numerical variables by flooring and capping all of the data, including the outliers, to the 25th and 75th percentiles of our data. This means that the lowest value of our data will be the 25th percentile in our data and the highest value will be the 75th percentile in our data, thus eliminating the outliers. The treatment results can be seen below.

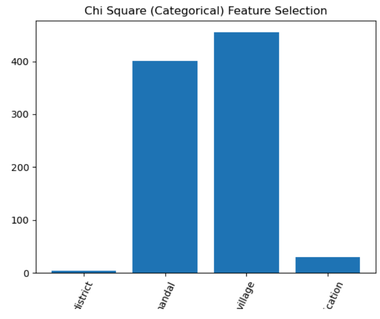




With the outliers now treated, we are able to select our quality metrics used for our classification models. To find these metrics, we used two different selection methods, Analysis of Variance (ANOVA) and Chi-Square. ANOVA selection is like picking the most important ingredients for a recipe. It looks at how much each ingredient affects the final taste. By focusing on the ingredients that make the biggest difference, it helps simplify cooking and make better-tasting dishes. In machine learning, it helps us choose the most important factors for making accurate predictions while ignoring less important ones. Chi-square selection is like sorting through a bag of candies to find the most popular flavors. It compares how often each candy flavor appears in different bags. If a flavor shows up more often than expected, it's considered significant. Similarly, in machine learning, Chi-square feature selection identifies the most influential features for making predictions by comparing how often they occur with the outcome. Using these selection methods, we found that the following quality metrics should be used in our classification models: RSC, SAR, Na, E.C, TDS, HCO3, pH, Mandal, and Village. These quality metrics scored the highest in our selection methods, as seen in our figures below.

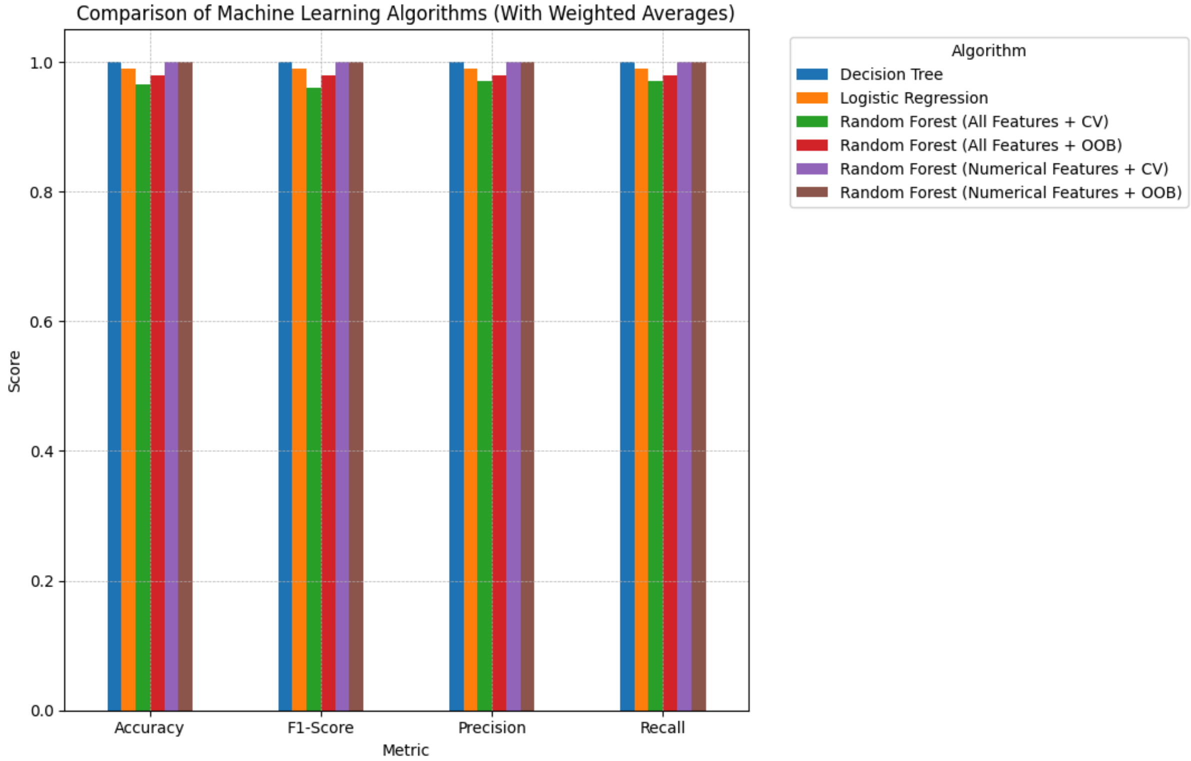






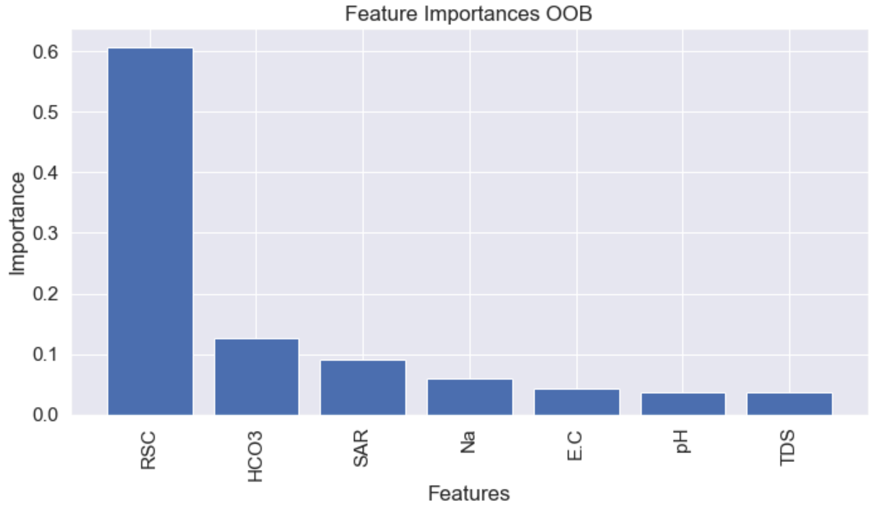
In the context of water safety for agriculture and health, similar to healthcare data, careful data cleaning is important. This ensures accuracy by addressing missing values and outliers. Also, using methods like ANOVA and Chi-Square help in choosing key quality metrics for precise classification models. These metrics are important for predicting water safety levels, maintaining agricultural safety, and upholding good public health.

With our cleaned data and our selected features, the team created a selection of models with three different algorithms to help classify water samples. The purpose of these models is to predict a water's safeness from the levels of the features we deemed as significant, classifying the water into safe or unsafe categories. With this, we could determine what level of contaminants would pass for safe water and which would be considered too much for an everyday use. As such, a model that could predict reliably and with a high accuracy is what we sought to find. We ended up with a total of 6 potential models: 1 logistic regression model, 1 decision tree model, and four random forest models, whose weighted average accuracy, f1-score, precision, and recall values we took to compare against one another in the figure below.



The accuracy checks how many predictions each model got right out of all the samples available. The precision is the ratio of correct identifications out of all everything that was classified as that class, we took the weighted average of the safe water precision and unsafe water precision and displayed it below for each model. This value is important especially in our case as it tells us our true classification rate, a high precision means that we can confidently say that this sample is safe water if we classify it as so. Next, we have the recall, which is the ratio of the correct identifications out of everything that should have been classified as that class. A high recall means that we have identified most of the safe water samples and minimizes the chance that safe water has been misclassified as unsafe water. The f1-score is a combination of precision and recall, so you can understand how well you are doing on both fronts.

While each model produced favorable results, the numerical out-of-bag random forest model was our preferred model, as it showed 100% accuracy, precision, recall, and F1 scores. Furthermore, the model exhibited low bias and variance, which is an ideal scenario as low bias tells us that the model can accurately find important patterns in the data it learns from and low variance tells us that the model is not sensitive to changes in the data we give it and is not overly complex. This combination of low bias and variance assures us that the model will respond well to new, unseen samples. This is perfect for our needs, as the scores reassures us that it can confidently and accurately predict safe water from unsafe water, which is essential when the water being used is meant for an agricultural purpose. A slip up in classification could easily poison the soil, nearby waterways, and the crops itself. More so, using this model means we could test samples from previously untested bodies of water that the model has not encountered before and receive a confident answer on its condition. We could greatly decrease the chances of exposure to unsafe water by testing samples with this model and seeing which bodies of water return the most unsafe classifications, thus giving us a direction where targeted clean up efforts need to be focused. Lastly, using this model to keep monitoring trends in water samples can show us where environmental protection efforts have been failing over time if they keep producing unsafe samples despite the clean up efforts.



Above, we see the importance of the features decided by the out-of-bag random forest model. The chart shows how much each feature impacts the final prediction of the sample, and it shows that the residual sodium carbonate index (RSC) of the water is easily the most influential feature, followed closely by the amount of bicarbonate (HCO3) in the water. Knowing this, we can find ways to lower the RSC of the water and its bicarbonate amounts with water treatments, which could make previously unsafe water suitable for irrigation use.

While there are no glaring limitations of the model, it is important that there could be a possible bias in the model. The data itself is heavily imbalanced as the ratio between P.S. (safe water) and U.S. (unsafe water) appeared to be disproportionate, with 750 P.S. values and 67 U.S. values total in the dataset. As such, the model could exhibit bias towards the majority class. However, it is also important to note that most real-life data is imbalanced, and random forest models can combat this since they apply weights to the features and determine their importance.

According to The World Economic Forum, 70% of the water found in India is “unfit for consumption” and the country's population estimated to be about 1.417 billion in 2022 leaves 980 million of India’s residents at risk of water-borne illnesses. The purpose of this dataset is to attain a better understanding of water potability within the 30 cities in India found in the dataset. By using different prediction methods and eliminating certain features the findings can be presented to citizens, current businesses, and potential investors trying to improve the water potability in those cities. The citizens of the listed cities need to know about the water potability results because water is a part of the daily cooking, cleaning, and hygienic practices. If this information is readily available and dispersed by government officials it allows citizens to find alternative measures to acquire safe water like boiling the water first before consumption, buying packaged water, and if a prolonged lack of access is seen taking federal actions to ensure they get safe water. The potential businesses need to know about their city’s water potability since it impacts daily business functions. Based on the high result of contaminants in the local water sources, in 2021 the Bhadhadri city government created a district environmental plan to mitigate the water crisis. The district plan states in the water management sections how it plans for the “rejuvenation of water bodies i.e. Lakes and tanks for desilting and improving the holding capacity have been taken up in Mission Kakatiya program initiated by the Government of Telangana”. Another stakeholder of the water potability results is potential investors looking to move businesses into these cities and environmental startups looking for cheap but profitable solutions to bring safe water to the citizens. Uravu Labs is an ecological startup founded in 2017 based in Bengaluru, India that develops a renewable water technology that utilizes inexhaustible atmospheric moisture and renewable energy to produce drinking water.

In conclusion, knowing the potability of the water in the Indian cities listed in the dataset is critical for the well-being of the commonwealth, the functioning of businesses, and the attraction of investors and new ventures. Since a lot of water is considered dangerous, people must be aware of this to take preventative measures. Water safety is crucial since business operations depend on clean water. The district environmental plan in Bhadhadri is one example of an initiative demonstrating proactive measures to alleviate the water shortage. Startups and investors are essential in supplying cutting-edge fixes for raising water quality. In general, promoting a healthy environment and guaranteeing the well-being of communities in India requires an understanding of water potability.

Works Cited

“Water Pollution Is Killing Millions of Indians. Here’s How Technology and Reliable Data Can Change That.” *World Economic Forum*, www.weforum.org/agenda/2019/10/water-pollution-in-india-data-tech-solution/#:~:text=are%20getting%20toxic.-,It’s%20estimated%20that%20around%2070%25%20of%20surface%20water%20in%20India,a%20tiny%20fraction%20adequately%20treated. Accessed 30 Apr. 2024.

Introduction Bhadradri Kothagudem District, tspcb.cgg.gov.in/DEP/Bhadradri Kothagudem DEP Plan-converted.pdf. Accessed 30 Apr. 2024.

Uravu Labs - Products, Competitors, Financials, Employees, Headquarters Locations, www.cbinsights.com/company/uravu-labs. Accessed 30 Apr. 2024.