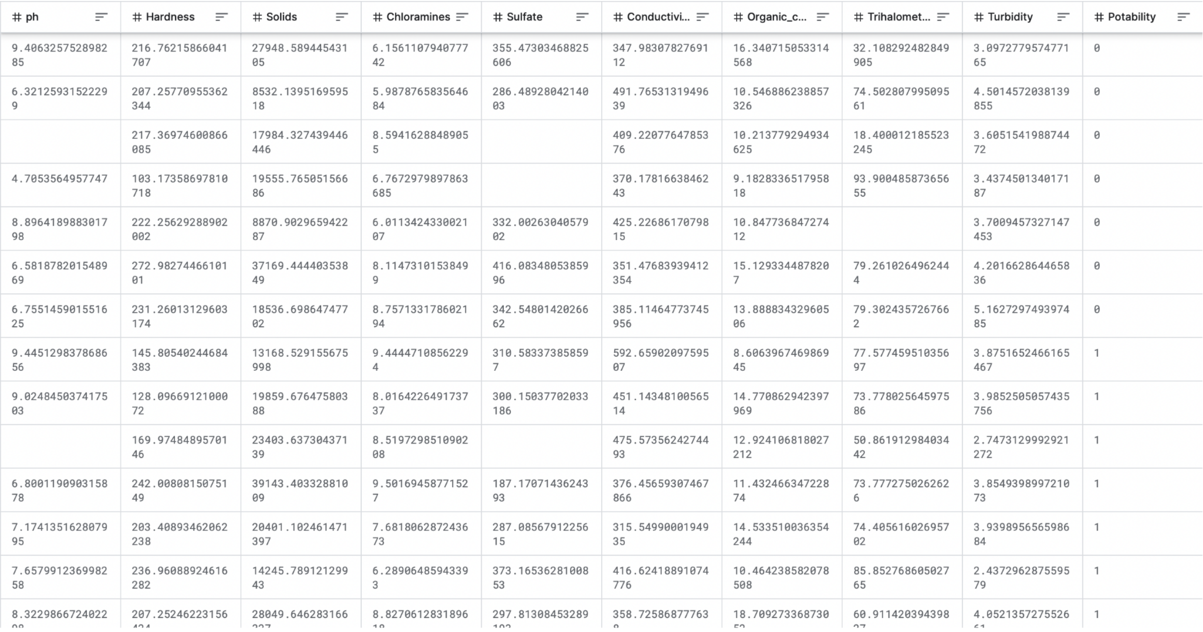
# DS 2010 Final Report

Abhiram Yammanuru, Michael Herchenroder, Ryan Offstein, William Abraham, Ajay Bhagavatula

## About Dataset & Motivation:

Our group decided to determine what factor would have the biggest impact on water potability (drinkability) for humans. Our main motivation behind this was to see what features people would have to look into when determining the drinkability of water. We also looked at this from a humanitarian perspective, as drinkable water is an essential human need, and access to drinking water is a serious issue faced in many areas of the world.

We found a dataset with different variables that each could have a potential impact on the drinkability. The specific variables found in the dataset that impact the potability are Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, and Turbidity. With all these variables that could potentially impact the potability, we decided to narrow down on a few for our conjectures. We specifically wanted to compare these variables to each other to see the best predictor of potability. The 3 conjectures/hypotheses that we came up with were that Turbidity is a better predictor of potability than solids, Ph is a better predictor of potability than conductivity, and Hardness is a better predictor of potability than Organic carbon. They can be seen listed below the dataset.



| # ph | # Hardness | # Solids | # Chloramines | # Sulfate | # Conductivi... | # Organic_c... | # Trihalomet... | # Turbidity | # Potability |
|---|---|---|---|---|---|---|---|---|---|
| 9.406325752898285 | 216.76215866041707 | 27948.58944543105 | 6.156110794077742 | 355.47303468825606 | 347.9830782769112 | 16.340715053314568 | 32.108292482849905 | 3.097277957477165 | 0 |
| 6.32125931522299 | 207.25770955362344 | 8532.139516959518 | 5.987876583564684 | 286.4892804214003 | 491.7653131949639 | 10.546886238857326 | 74.50280799509561 | 4.5014572038139855 | 0 |
|  | 217.36974600866085 | 17984.327439446446 | 8.59416288489055 |  | 409.2207764785376 | 10.213779294934625 | 18.400012185523245 | 3.605154198874472 | 0 |
| 4.7053564957747 | 103.17358697810718 | 19555.76505156686 | 6.7672979897863685 |  | 370.1781663846243 | 9.182833651795818 | 93.90048587365655 | 3.437450134017187 | 0 |
| 8.896418988301798 | 222.25629288902002 | 8870.902965942287 | 6.011342433002107 | 332.0026304057902 | 425.2268617079815 | 10.84773684727412 |  | 3.7009457327147453 | 0 |
| 6.581878201548969 | 272.9827446610101 | 37169.44440353849 | 8.11473101538499 | 416.0834805385996 | 351.47683939412354 | 15.1293344878207 | 79.2610264962444 | 4.201662864465836 | 0 |
| 6.755145901551625 | 231.26013129603174 | 18536.69864747702 | 8.757133178602194 | 342.5480142026662 | 385.11464773745956 | 13.88883432960506 | 79.3024357267662 | 5.162729749397485 | 0 |
| 9.445129837868656 | 145.80540244684383 | 13168.529155675998 | 9.44447108562294 | 310.583373858597 | 592.6590209759507 | 8.606396746986945 | 77.57745951035697 | 3.8751652466165467 | 1 |
| 9.024845037417503 | 128.0966912100072 | 19859.67647580388 | 8.016422649173737 | 300.15037702033186 | 451.1434810056514 | 14.770862942397969 | 73.77802564597586 | 3.9852505057435756 | 1 |
|  | 169.9748489570146 | 23403.63730437139 | 8.519729851090208 |  | 475.5735624274493 | 12.924106818027212 | 50.86191298403442 | 2.7473129992921272 | 1 |
| 6.800119090315878 | 242.0080815075149 | 39143.40332881009 | 9.50169458771527 | 187.1707143624393 | 376.45659307467866 | 11.43246634722874 | 73.7772750262626 | 3.854939899721073 | 1 |
| 7.174135162807995 | 203.40893462062238 | 20401.102461471397 | 7.681806287243673 | 287.0856791225615 | 315.5499000194935 | 14.533510036354244 | 74.40561602695702 | 3.939895656598684 | 1 |
| 7.657991236998258 | 236.96088924616282 | 14245.78912129943 | 6.28906485943393 | 373.1653628100853 | 416.62418891074776 | 10.464238582078508 | 85.85276860502765 | 2.437296287559579 | 1 |
| 8.3229866724022 | 207.25246223156 | 28049.646283166 | 8.8270612831896 | 297.81308453289 | 358.72586877763 | 18.709273368730 | 60.911420394398 | 4.0521357275526 | 1 |

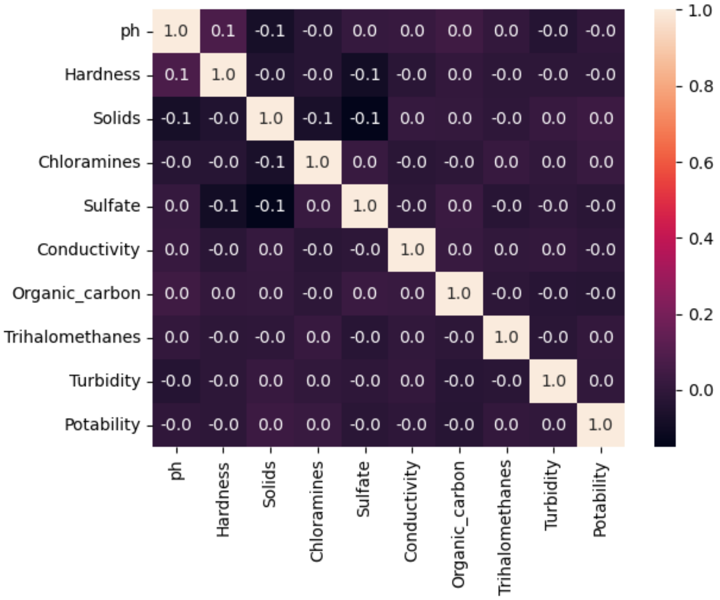
## Conjectures:

1. **Ph is a better predictor of potability than conductivity**
2. **Turbidity is a better predictor of potability than solids**
3. **Hardness is a better predictor of potability than Organic carbon**

We hypothesized that Turbidity, Ph, and Hardness were going to be the best predictors of potability from our data set so our conjectures are comparing those features, with other features in the dataset to back up that hypothesis.

The features we are comparing are relatively similar too so we can see which is the better predictor. For example, Turbidity is the opaqueness and thickness of the water, whereas solids is the mass of the suspended solids in the water. Opaqueness and thickness of the water are strong indicators of solids present, so they are similar in that way. Ph and conductivity are related in the sense that they take into consideration the $H^+$ ions in the water. Lower Ph's correlate with lower conductivity and vice versa. The hardness of water is related to Organic Carbon because $CO_2$ induces the dissolution of $CaCO_3$ in the reaction with water, it can contribute to an increase in the hardness of water.

Even though these features seemed to have an impact on each other due to similarities, we found that none of these features have any significant correlation with each other. As seen below, the correlation matrix we ran on each of the datasets shows that none of these features significantly impact each other. This worked to our benefit as we didn't want to have confounding variables, which would have made it a lot more challenging to see which features had the largest impact on potability.

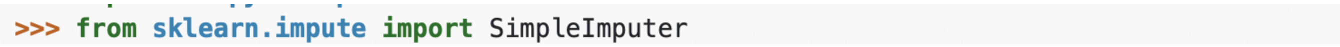


## Data Cleaning methods:

When first taking a look at our dataset, there were numerous missing values. The specific features that were missing values were Sulfate, Ph and Trihalomethanes. It was essential that we dealt with these in order to run different machine learning models on them. To remedy this, we attempted 4 different ways of cleaning our data. These were the deletion of missing values, Min-Max normalization, Simple Imputer, and Classified imputation.

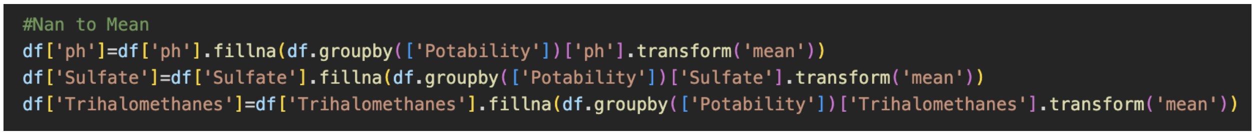
The first method that we used was the deletion of missing values, where if a row in the dataset was missing a value, we would remove it from the training set This was the easiest way of going about cleaning the data. Unfortunately it wasn't effective in providing results, as it removed a lot of the training data that we needed for the model. We did this in pandas by using the df.dropna[] function with Sulfates, Ph and Trihalomethanes.

We also tried simple imputation. Simple imputation essentially replaces the missing values with whatever values desired. The 3 main methods of simple imputation that we used were simple mean imputation, simple median imputation, and simple mode imputation. As stated in the name, simple mean imputation replaces the missing values of the feature with the mean value of that specific feature. Same goes with median and mode except it would make the missing value median and mode respectively. While this did allow us to run the models, and made the accuracy of these models higher, we believed that we could do better. We ran simple imputation using scikitlearns simple imputer package.



```
>>> from sklearn.impute import SimpleImputer
```

The final and most effective method that we did was classified imputation. It is similar to simple imputation as it also replaces the missing values with the mean value of the feature; however, it does that based on the class the datapoint is assigned to. Potability is a binary value, with 0 meaning that the water is potable and 1 meaning the water isn't potable. What this means is that the values will be imputed based on the classand would replace the empty data based on the 1 or 0 for the potability. This way it is a lot more unified in the way that the values are replaced which allowed us to get the highest accuracy for our model this way. The code that we used can be seen below.

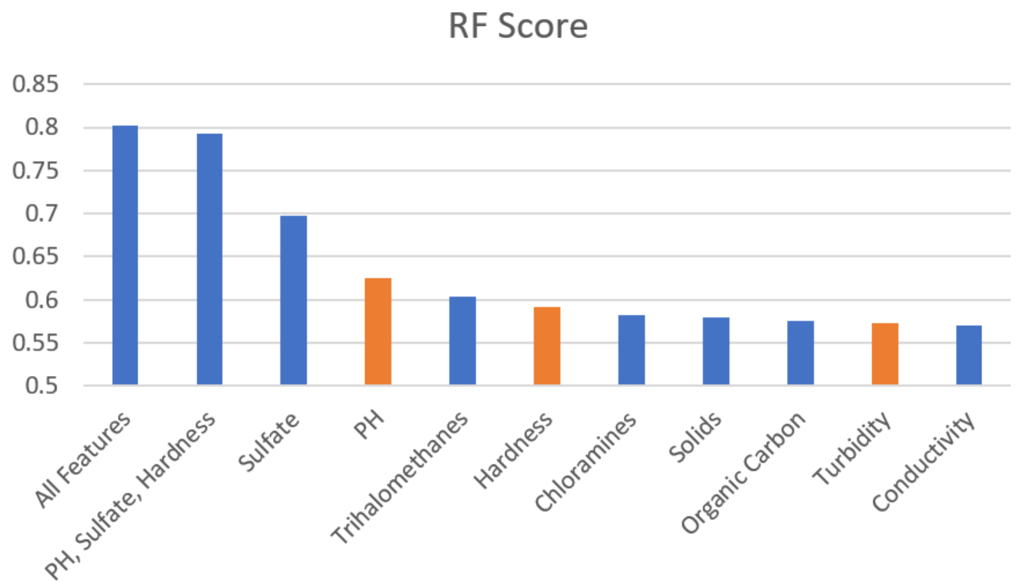


```
#Nan to Mean
df['ph']=df['ph'].fillna(df.groupby(['Potability'])['ph'].transform('mean'))
df['Sulfate']=df['Sulfate'].fillna(df.groupby(['Potability'])['Sulfate'].transform('mean'))
df['Trihalomethanes']=df['Trihalomethanes'].fillna(df.groupby(['Potability'])['Trihalomethanes'].transform('mean'))
```

Once our data wasn't missing values, the last cleansing procedure we applied to our data was min-max normalization. This is a technique in which the lowest value in a certain feature becomes a 0 and the highest value becomes a 1 and all the other values become scaled relatively. While at first this seemed like a good idea as the data would get normalized, it became challenging as a lot of the missing values were assigned 0, which inhibited and skewed the models quite a bit.

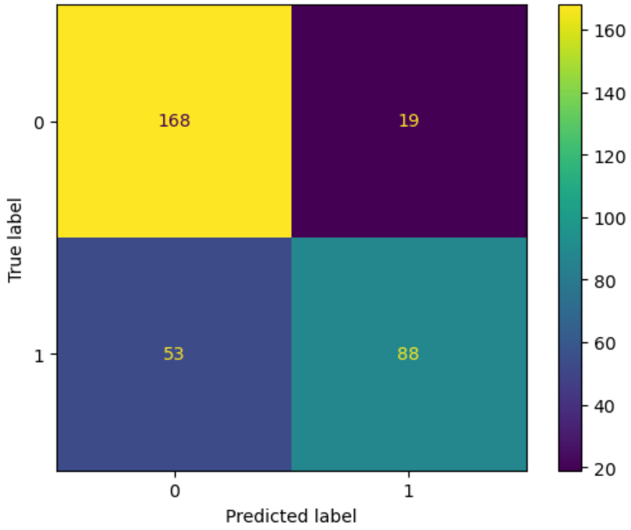
**Analysis:**

We trained several different classification models on our data test, including K-Nearest Neighbors, Random Forest, Logistic Regression, Gradient Descent, Support Vector Machine, and a Neural Network. Factors such as the method of data cleaning and imputation significantly impacted the accuracy of the models. We used 90% of the dataset for training and 10% for testing, as this struck a balance between giving the models enough data to make accurate predictions while preventing overfitting of the data. The high accuracy we achieved was 80.2% with the Random Forest Model that used all the features as predictors. The three features that best predicted potability were pH, Sulphates, and Hardness. Training a Random Forest Model using these three features as predictors resulted in an accuracy of 79.3%. Below is the accuracy of the Random Forest Model on the different individual and combinations of features in the dataset.



**Confusion Matrix:**

Overall, our confusion matrix displayed the successes of our model. Obviously the ideal confusion matrix correctly guesses that water is drinkable when it actually is and vice versa, but our confusion matrix showed that our model was fairly accurate. For water potability, the most important thing we kept in mind was minimizing the false positives, as water that isn't drinkable when ingested is critically harmful to the human body. Our confusion matrix only had approximately 5 percent of the values to be false positive, which is a good statistic. Although minimizing false negatives is important, not drinking drinkable water is a good example of being safe rather than sorry, so we weren't as concerned with this. The model with the least percentage of false positives was Random Forest, and this was the main reason why we stuck with that model. One thing to keep in mind however, is the fact that there were more data points of undrinkable water vs drinkable which would skew the model towards making more negative predictions.



## Results of Conjectures:

For each of the features in the conjectures, we decided to run the Random Forest Model as it was the most accurate. We then compared them to each other to determine if our conjectures were correct or not. Below the summary of each conjecture a bar plot can be seen with all the accuracies of the model for the specific conjectures.

### Conjecture 1

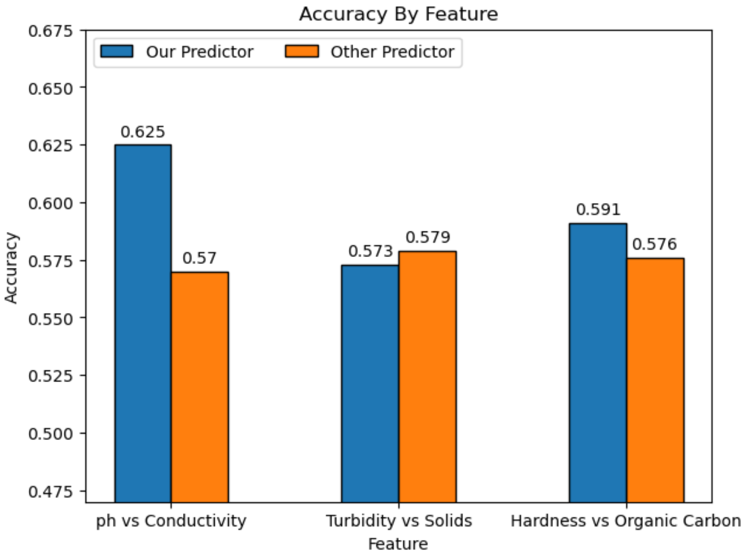
Our first conjecture was that **pH is a better predictor of potability than conductivity.** This conjecture was proven true due to our analysis, and was the only conjecture out of the 3 with a significant enough margin in between the accuracies. The accuracy of the pH model was 62.5 percent, when compared to the conductivity predictor at 57 percent. This conjecture was therefore proven correct

## Conjecture 2

Our second conjecture was that **Turbidity is a better predictor of potability than solids.** Although solids were a better predictor of potability than turbidity, the difference in accuracy was insignificant. Individually, both features were not good predictors of potability as they only correctly predicted potability around 60% of the time. We originally hypothesized that turbidity would be a better feature to base potability on, but solids are a marginally better predictor of drinkability. This conjecture was proven marginally wrong through our model.

## Conjecture 3

Our third conjecture was that **Hardness is a better predictor of potability than Organic carbon.** This conjecture was proven true when we analyzed the accuracy by feature for the dataset, as Hardness was given a higher accuracy rating than Organic Carbon. However, since both of these values were close to 50% accuracy, they were both ignored as single-column predictors because they were not accurate enough. We originally predicted that Hardness is a better predictor of potability when compared with It can be considered that this conjecture was marginally true.



## Challenges:

We ran into a lot of challenges when working with this dataset. One of the main issues that we ran into was missing values. The columns with the missing values were Ph, Sulfate, and Trihalomethanes. With these missing values, it's impossible to run machine learning algorithms on the data to predict the potability.

While these allowed us to run the machine learning algorithms that we desired on the dataset, the accuracies of these models weren't ideal as they were around 50 percent. We soon realized that the closer to 50 percent that they got, the models were doing nothing better than guessing because our dataset condition is binary with 0 meaning the water is not potable and 1 means that the water is potable.

To combat this, we decided to do a more complex version of the imputation that we did earlier. This type of imputation is called classified imputation. Classified Imputation input values that provided the water was potable separate from those values that stated that the water wasn't potable. The values that provided 1 as the result would have a different imputation than the values that provided 0. This would fill in the empty values more purposefully and provided us with a Random Forest model closer to 80 percent accuracy.

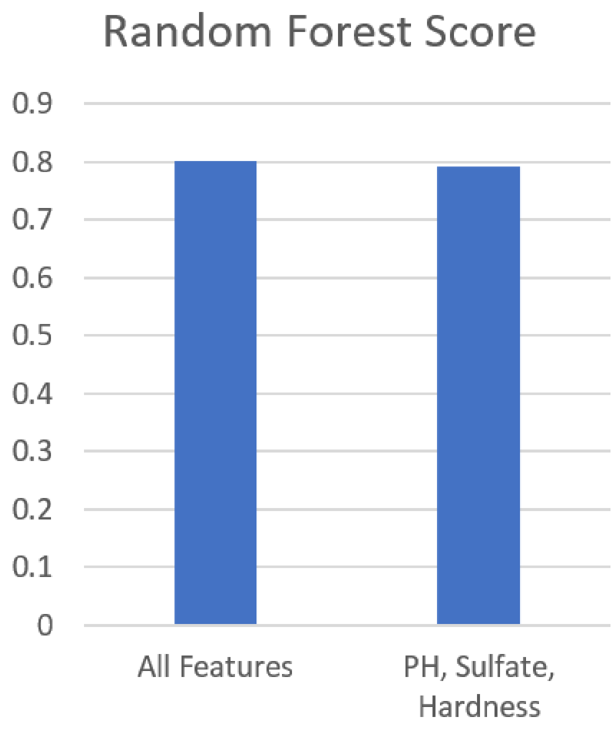
## Conclusion:

To summarize, our main motivation behind this was to see what features people would have to look into when determining the drinkability of water. We looked at this from a humanitarian perspective, as drinkable water is an essential part of survival.

We cleaned the data, and ran multiple machine learning models on the dataset to determine the feature(s) with the biggest impact of potability. To ensure that our findings were correct, we made a confusion matrix to determine how correctly the model is outputting potability.

After thorough investigation with numerous models with all the different features, we have come to the recommendation to look out for specifically pH, Sulfate and Hardness as the biggest contributors to if the water is potable or not. In our bar chart below, when comparing the Random Forest Model accuracy with potability, it is nearly identical. The accuracy for all of the features is 80%, when compared to the 79.3% that is provided with the 3 features listed above.

From a business standpoint with minimal resources, our recommendation for any humanitarian effort to improve potability would be testing of Hardness, pH, and Sulfate in the water. Therefore they could use the resources gained from making less measurements to collect more samples for those three features.



**Member Contributions:**

Abhi Yammanuru - Organized meeting times, performed simple imputation on the dataset, worked on slides, worked on introduction, conjectures, data cleaning methods and challenges sections of the essay

Michael Herchenroder- Assembled and organized the slides of the presentation, and helped figure out methods of analysis and data cleaning.

Ryan Offstein- Wrote code for imputing the data, trained several of the machine learning models, specifically the support vector machine model, created the confusion matrix, worked on several of the slides.

William Abraham- Wrote speaker notes for slides, added graphs, cleaned up slide visibility, thought of different ideas for report structure, wrote, connected and compiled code in jupyter notebook

Ajay Bhagavatula- Wrote code to clean and impute data, trained several machine learning models, created visualizations, worked on the analysis section of the essay, worked on the presentation