Water Potability

Group 3 - Abhi Yammanuru, Ajay Bhagavatula, Ryan Offstein, William Abraham, Michael Herchenroder

Motivation

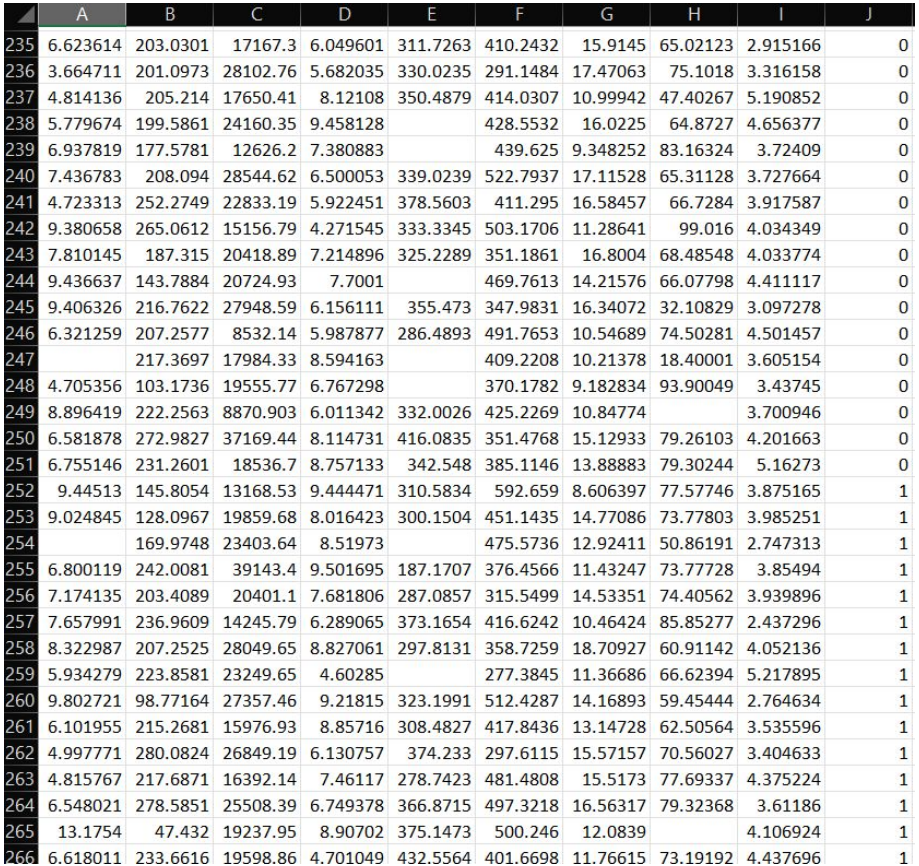
Our group decided to determine what factor would have the biggest impact on water potability (drinkability) for humans. The motivation behind this was to see what features people would have to look into when determining the drinkability of water. We looked at this from a humanitarian perspective, as drinkable water is an essential part of survival.

Conjectures

1. Turbidity is stronger than Solids for predicting Potability
2. pH is stronger than Conductivity for predicting Potability
3. Hardness is stronger than Organic carbon for predicting Potability

The Data

- 9 Features
- 3000+ Samples
- Binary Output Variable
- Needed cleaning

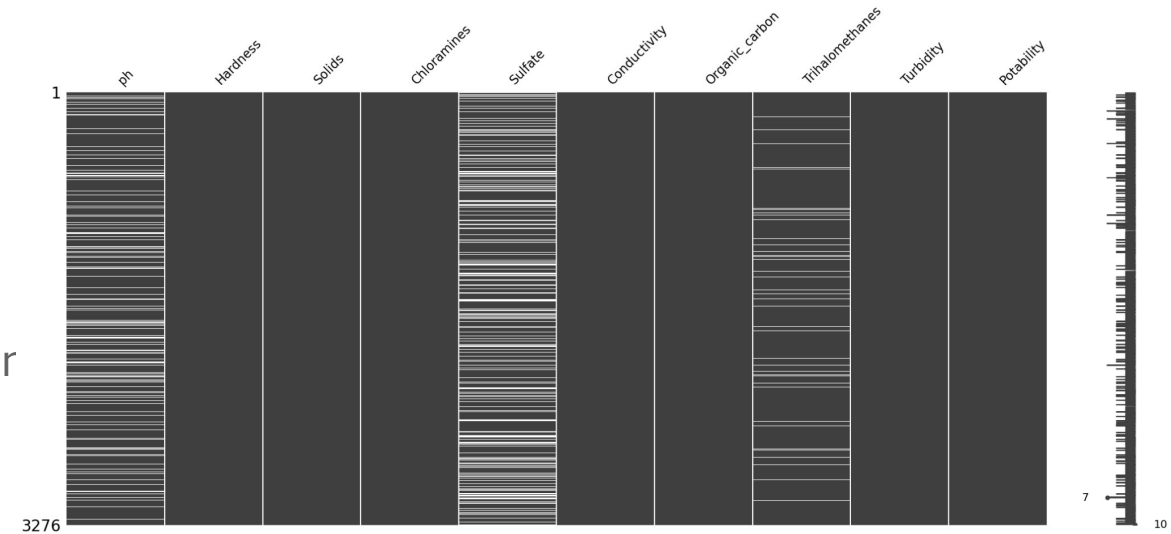


	A	B	C	D	E	F	G	H	I	J
235	6.623614	203.0301	17167.3	6.049601	311.7263	410.2432	15.9145	65.02123	2.915166	0
236	3.664711	201.0973	28102.76	5.682035	330.0235	291.1484	17.47063	75.1018	3.316158	0
237	4.814136	205.214	17650.41	8.12108	350.4879	414.0307	10.99942	47.40267	5.190852	0
238	5.779674	199.5861	24160.35	9.458128		428.5532	16.0225	64.8727	4.656377	0
239	6.937819	177.5781	12626.2	7.380883		439.625	9.348252	83.16324	3.72409	0
240	7.436783	208.094	28544.62	6.500053	339.0239	522.7937	17.11528	65.31128	3.727664	0
241	4.723313	252.2749	22833.19	5.922451	378.5603	411.295	16.58457	66.7284	3.917587	0
242	9.380658	265.0612	15156.79	4.271545	333.3345	503.1706	11.28641	99.016	4.034349	0
243	7.810145	187.315	20418.89	7.214896	325.2289	351.1861	16.8004	68.48548	4.033774	0
244	9.436637	143.7884	20724.93	7.7001		469.7613	14.21576	66.07798	4.411117	0
245	9.406326	216.7622	27948.59	6.156111	355.473	347.9831	16.34072	32.10829	3.097278	0
246	6.321259	207.2577	8532.14	5.987877	286.4893	491.7653	10.54689	74.50281	4.501457	0
247		217.3697	17984.33	8.594163		409.2208	10.21378	18.40001	3.605154	0
248	4.705356	103.1736	19555.77	6.767298		370.1782	9.182834	93.90049	3.43745	0
249	8.896419	222.2563	8870.903	6.011342	332.0026	425.2269	10.84774		3.700946	0
250	6.581878	272.9827	37169.44	8.114731	416.0835	351.4768	15.12933	79.26103	4.201663	0
251	6.755146	231.2601	18536.7	8.757133	342.548	385.1146	13.88883	79.30244	5.16273	0
252	9.44513	145.8054	13168.53	9.444471	310.5834	592.659	8.606397	77.57746	3.875165	1
253	9.024845	128.0967	19859.68	8.016423	300.1504	451.1435	14.77086	73.77803	3.985251	1
254		169.9748	23403.64	8.51973		475.5736	12.92411	50.86191	2.747313	1
255	6.800119	242.0081	39143.4	9.501695	187.1707	376.4566	11.43247	73.77728	3.85494	1
256	7.174135	203.4089	20401.1	7.681806	287.0857	315.5499	14.53351	74.40562	3.939896	1
257	7.657991	236.9609	14245.79	6.289065	373.1654	416.6242	10.46424	85.85277	2.437296	1
258	8.322987	207.2525	28049.65	8.827061	297.8131	358.7259	18.70927	60.91142	4.052136	1
259	5.934279	223.8581	23249.65	4.60285		277.3845	11.36686	66.62394	5.217895	1
260	9.802721	98.77164	27357.46	9.21815	323.1991	512.4287	14.16893	59.45444	2.764634	1
261	6.101955	215.2681	15976.93	8.85716	308.4827	417.8436	13.14728	62.50564	3.535596	1
262	4.997771	280.0824	26849.19	6.130757	374.233	297.6115	15.57157	70.56027	3.404633	1
263	4.815767	217.6871	16392.14	7.46117	278.7423	481.4808	15.5173	77.69337	4.375224	1
264	6.548021	278.5851	25508.39	6.749378	366.8715	497.3218	16.56317	79.32368	3.61186	1
265	13.1754	47.432	19237.95	8.90702	375.1473	500.246	12.0839		4.106924	1
266	6.618011	233.6616	19598.86	4.701049	432.5564	401.6698	11.76615	73.19192	4.437696	1

Cleaning Method

- Removal
- Min-Max Normalization
- Simple Mean imputation
- Classified Mean imputatior

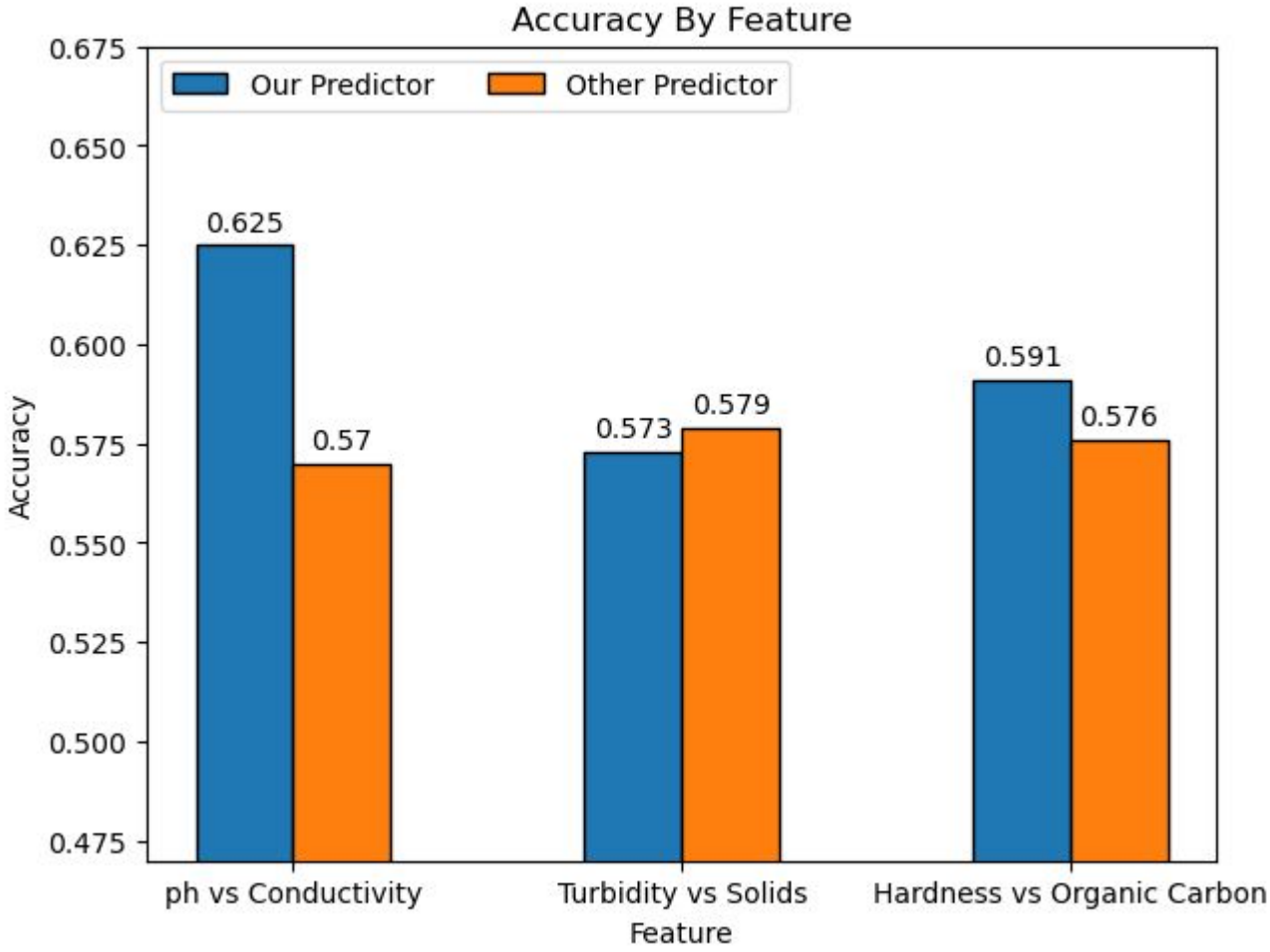
Data with missing values visualized



Classified Mean Value Imputation Code

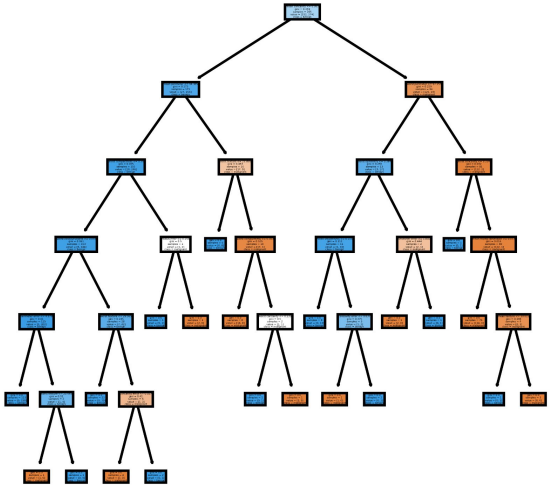
```
#Nan to Mean
df['ph']=df['ph'].fillna(df.groupby(['Potability'])['ph'].transform('mean'))
df['Sulfate']=df['Sulfate'].fillna(df.groupby(['Potability'])['Sulfate'].transform('mean'))
df['Trihalomethanes']=df['Trihalomethanes'].fillna(df.groupby(['Potability'])['Trihalomethanes'].transform('mean'))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

Results


Accuracy By Feature
Our Predictor
Other Predictor
0.675
0.650
0.625
0.600
0.575
0.550
0.525
0.500
0.475
Accuracy
0.625
0.57
0.573
0.579
0.591
0.576
ph vs Conductivity
Turbidity vs Solids
Hardness vs Organic Carbon
Feature

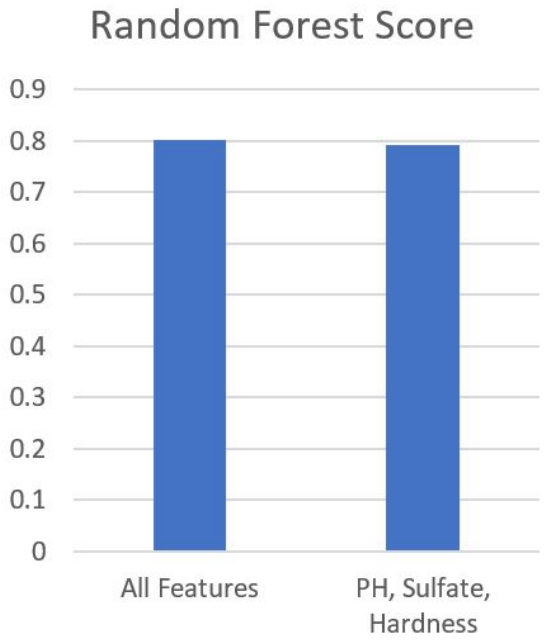
Analysis

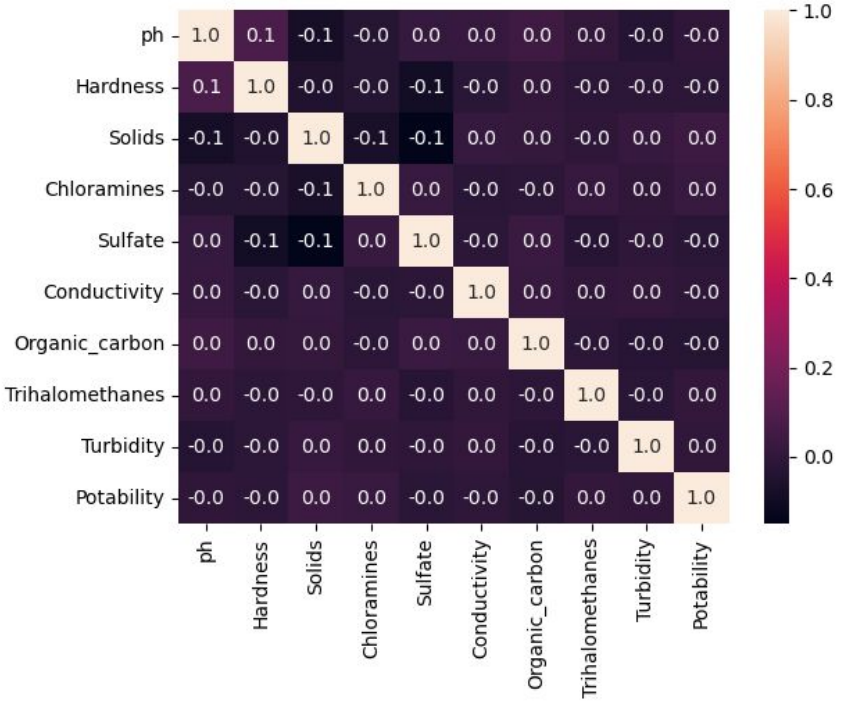
- Models used
- Training-Test split
- Features used
- Accuracy



Using More Features

- Because one feature didn’t score too well, we tried using more.
- We first tried all the features
- Then a selection of 3 top features.

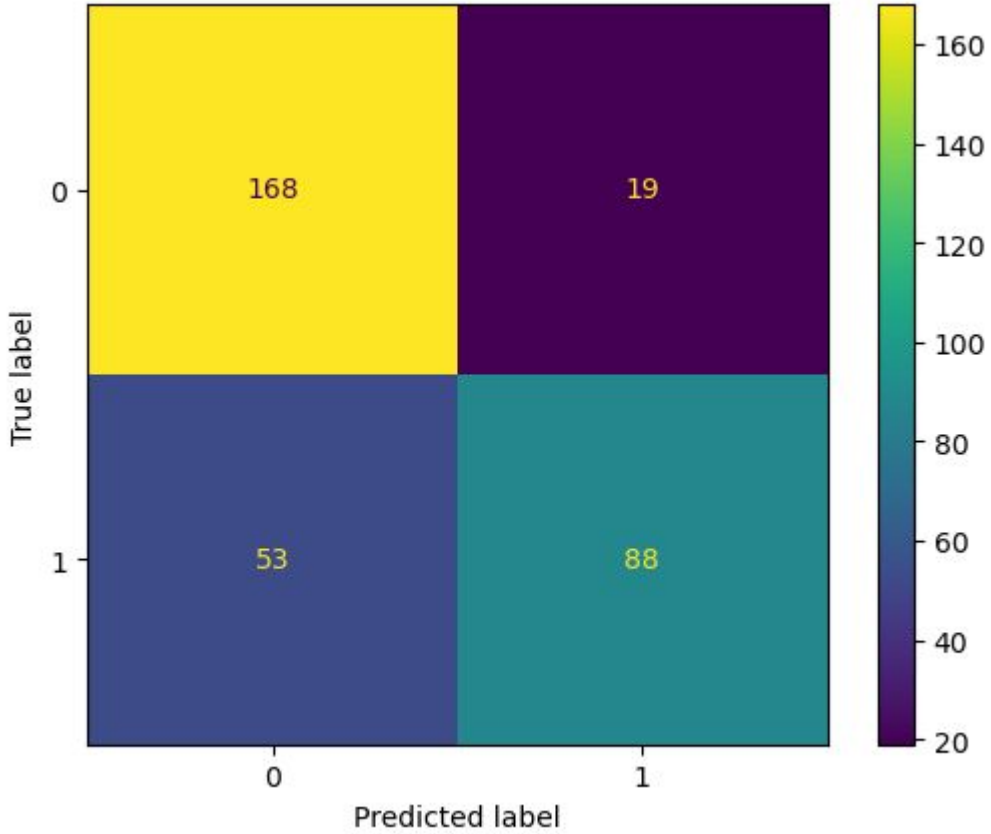




Correlation Matrix

No significant correlations were found between any of the features given

Confusion Matrix



Recommendations

- If given a limited budget, purely focus on measuring the Ph, Sulfate, and Hardness measurements of the water.

- If the budget allows, spend more money to get more samples for Ph, Sulfate, and Hardness instead of measuring all nine features.

- Invest in higher quality measurement materials to secure a greater accuracy of measurement for the samples.