T5 DATASCIENCE BOOTCAMP 6 OCT 2021

PREDICT THE WATER POTABILITY USING LOGISTIC REGRESSION

Raghad Aloraini - Raghad Alawad

Introduction:

The dataset we used are taken from the kaggle website (Water Quality),
We wanted to predict if the water is drinkable or not.

Dataset features:

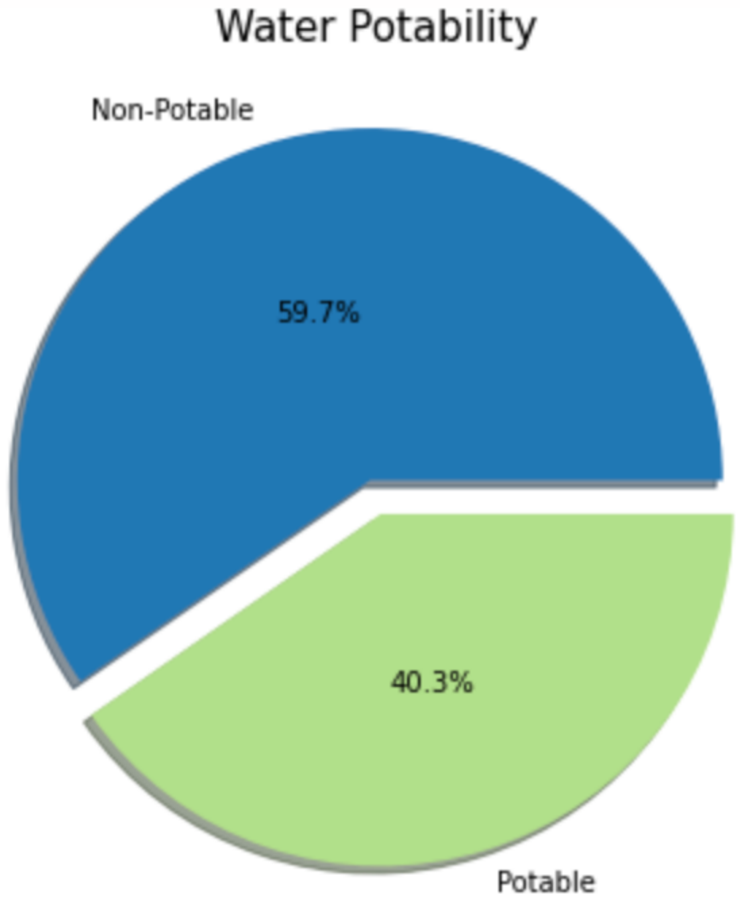
- **ph: pH of 1. water (0 to 14).**
- **Hardness: Capacity of water to precipitate soap in mg/L.**
- **Solids: Total dissolved solids in ppm.**
- **Chloramines: Amount of Chloramines in ppm.**
- **Sulfate: Amount of Sulfates dissolved in mg/L.**
- **Conductivity: Electrical conductivity of water in µS/cm.**
- **Organic_carbon: Amount of organic carbon in ppm.**
- **Trihalomethanes: Amount of Trihalomethanes in µg/L.**
- **Turbidity: Measure of light emiting property of water in NTU.**
- **Potability: Indicates if water is safe for human consumption. Potable: 1, not potable: 0**

Project steps:

1. EDA
2. Feature engineering
3. Linebase model
4. Parameter tuning
5. Results comparison

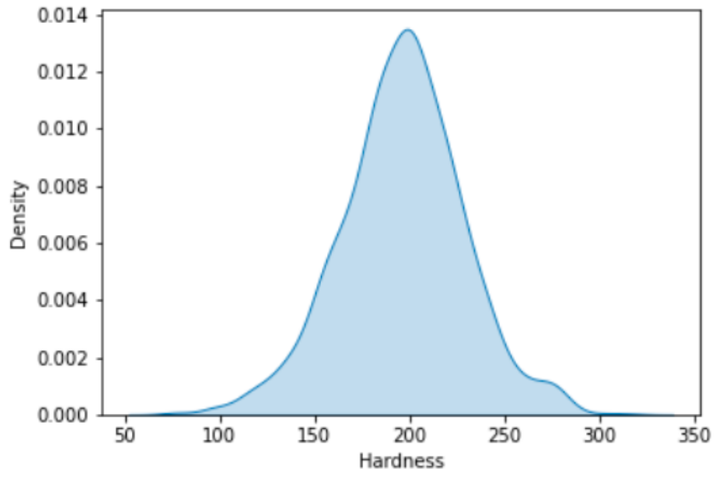
1 Exploratory Data Analysis

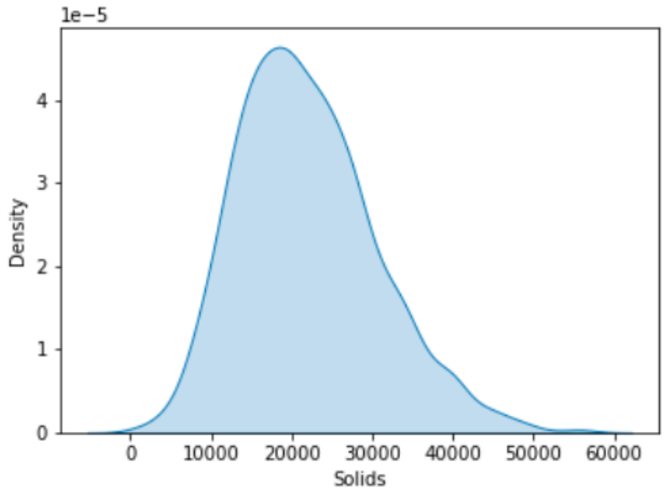
Visualization for number of potable and non-potable data

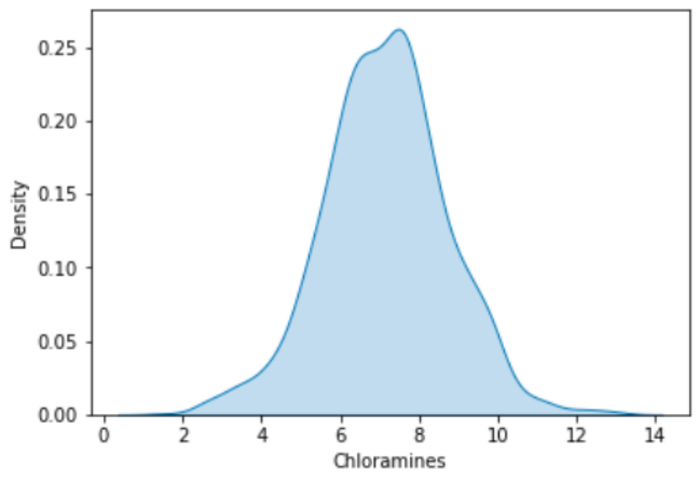


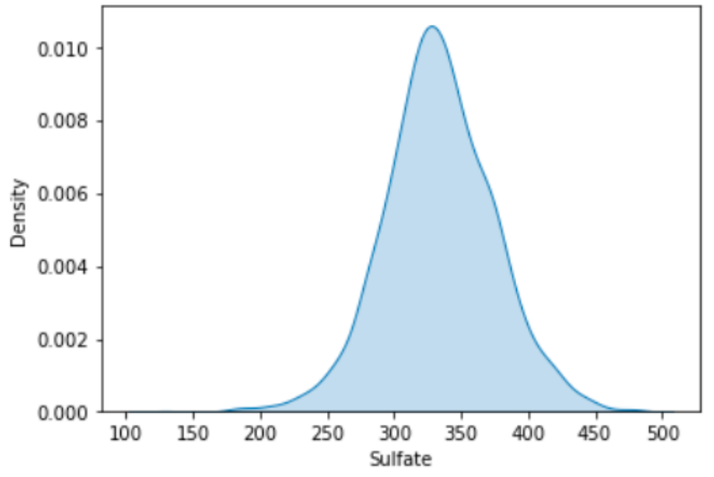
1 Exploratory Data Analysis

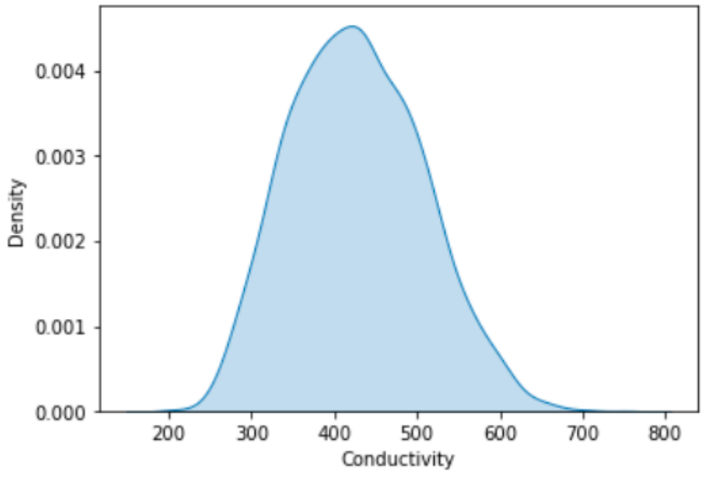
Distribution Plots

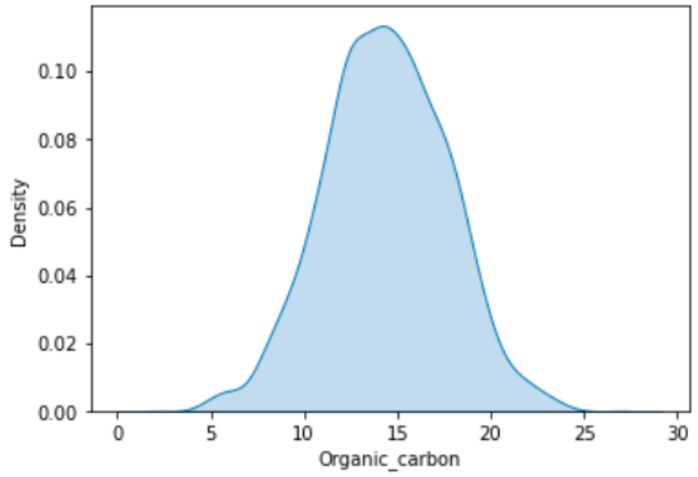








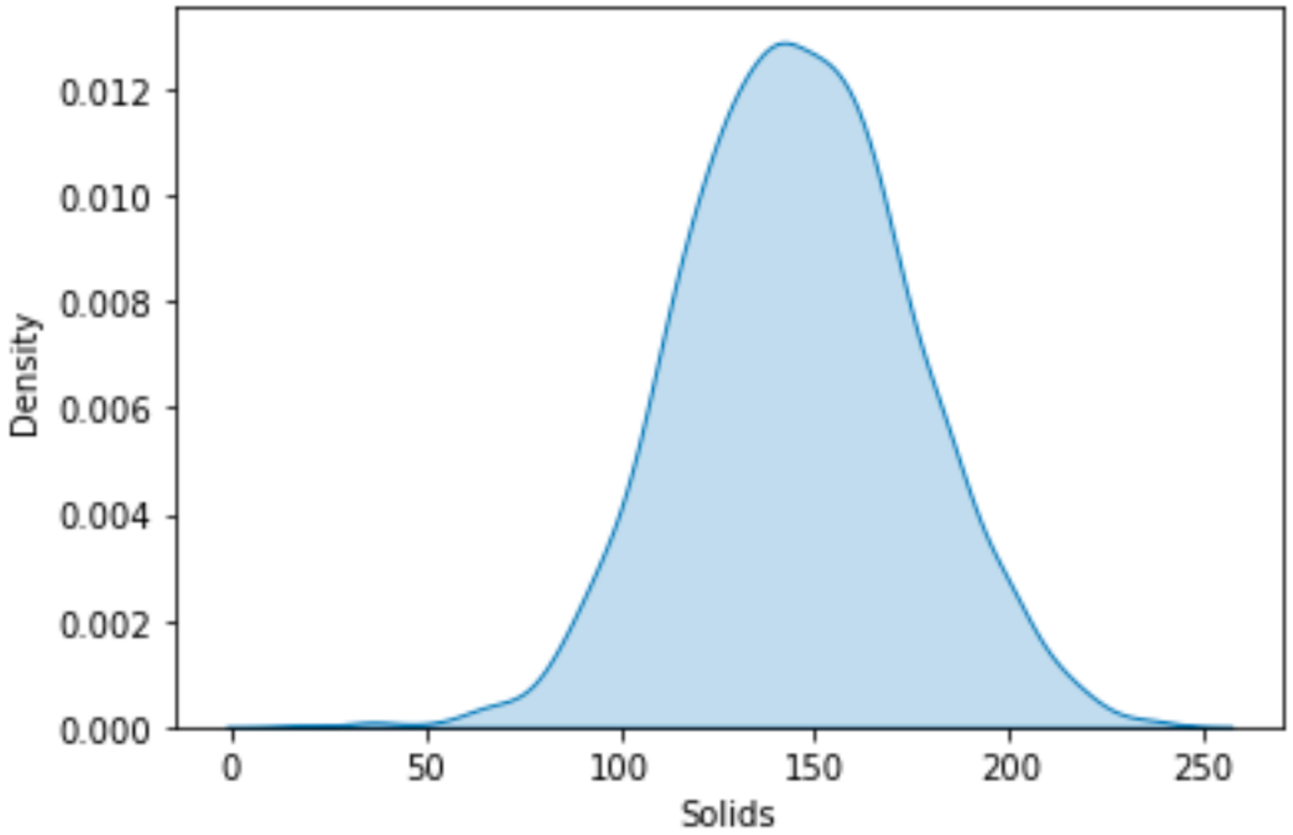




2

Feature engineering

Solids graph after applying a transformation to fix the skewness



3

Linebase model

Logistic Regression Model

```
0.5235732009925558

[[141  99]
 [ 93  70]]
```

	precision	recall	f1-score	support
0	0.60	0.59	0.59	240
1	0.41	0.43	0.42	163
accuracy			0.52	403
macro avg	0.51	0.51	0.51	403
weighted avg	0.53	0.52	0.52	403

4

Parameter tuning

Logistic Regression Model

Cross Vallidation

Search for an optimal value of K for KNN

Grid Search

Find the best parameters

5

By the numbers

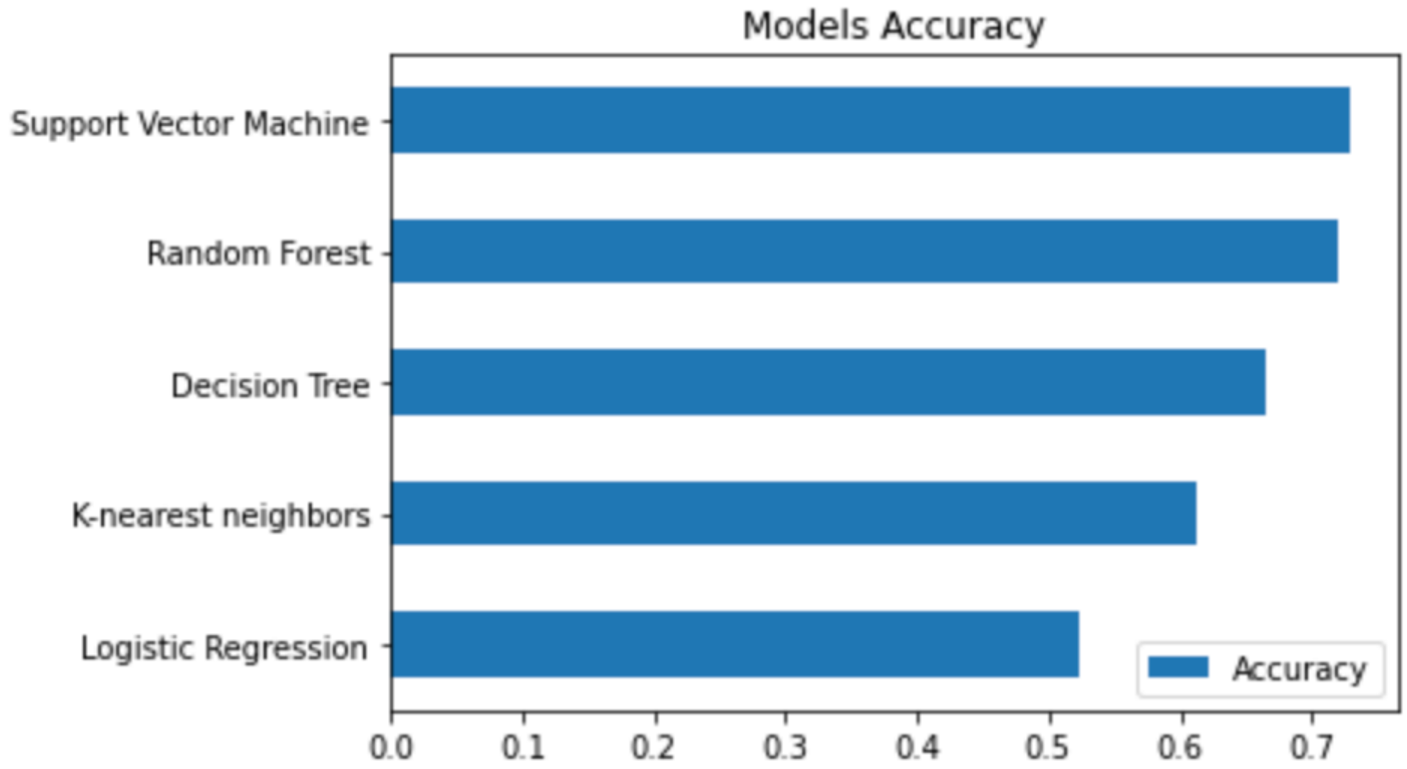
Accuracy and Recall scores before and after parameter tuning

Model	Accuracy	Recall
K-nearest nieghbours Before	0.63	0.79
K-nearest nieghbours After	0.61	0.70
Decision Tree Before	0.56	0.63
Decision Tree After	0.66	0.80
Random Forest Before	0.714	0.89
Random Forest After	0.719	0.91
Support Vector Before	0.729	0.93

5

Accuracy Visualization

Models Accuracy scores



Future Work:

- Increase data to have better results
- Train more models

Thank you!