Classification of Water Potability

LIYA LAPIERRE, PRATEEK KAKKAR, NITHARSAN SIVAKANTHAN

Background

Drinkable water is essential to human health

Interested in finding a series of tests to use to predict Potability (safe drinking water)

If possible, find the least number of tests needed to improve accessibility

The Data

Water Quality Dataset (Posted on Kaggle)

Metrics for ~ 3300 bodies of water

Predictors: pH level, Hardness, Solids, Chloramines, Organic Carbon, Trihalomethanes, Turbidity

Response: Potability

- ~ 2000 bodies of water are not Potable, ~1300 are Potable

Methodology

Data Cleaning

- SVD Imputation

Data Exploration

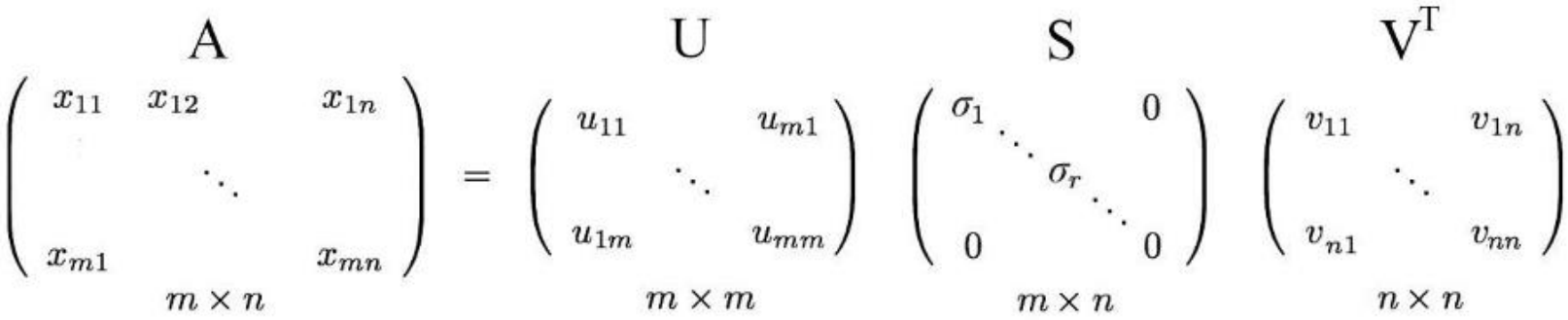
- Principal Component Analysis

Modeling

- Logistic Regression
- Random Forest Classifier
- K-Nearest Neighbor Classifier

Data Cleaning

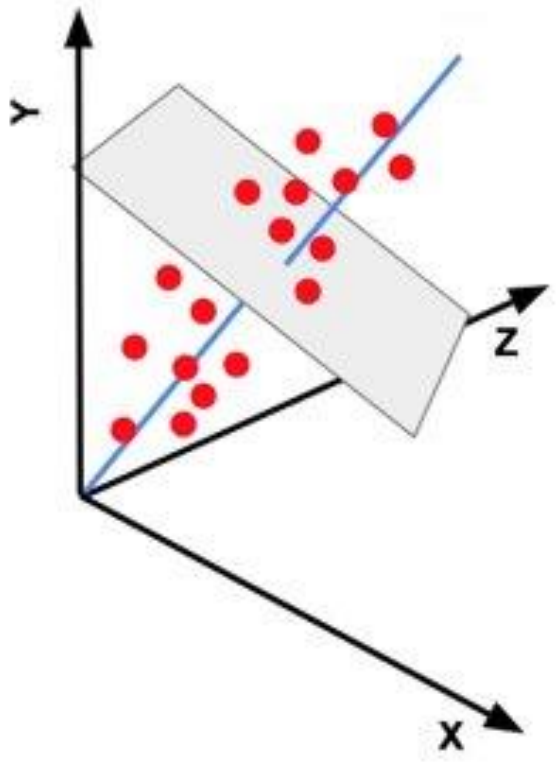
- Singular Value Decomposition Imputation



Source: (Hui, 2019)

Data Exploration

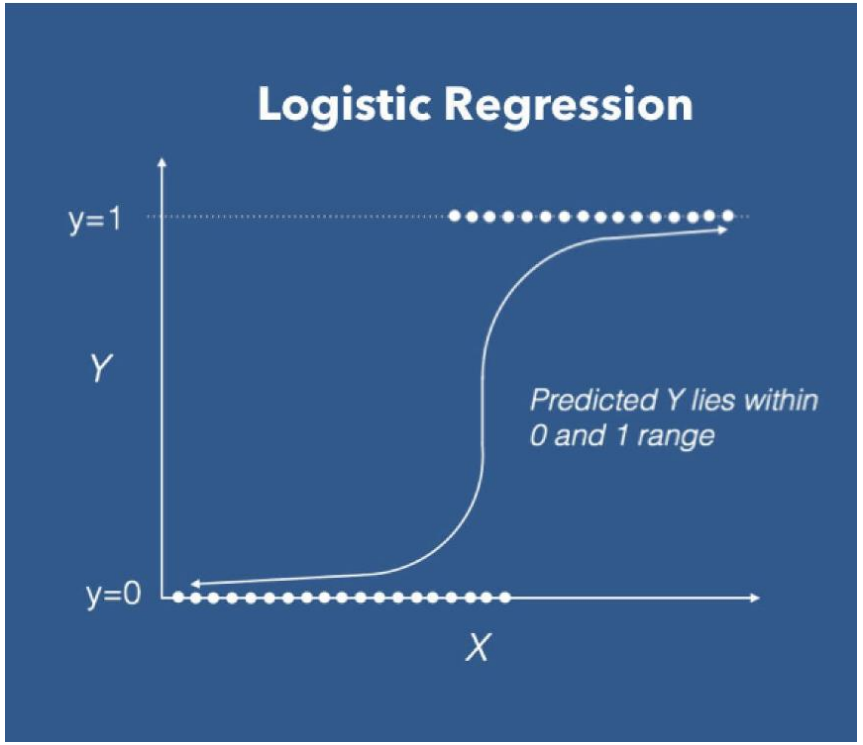
- Principal Component Analysis



Source: (LearnOpenCV)

Classification Methods

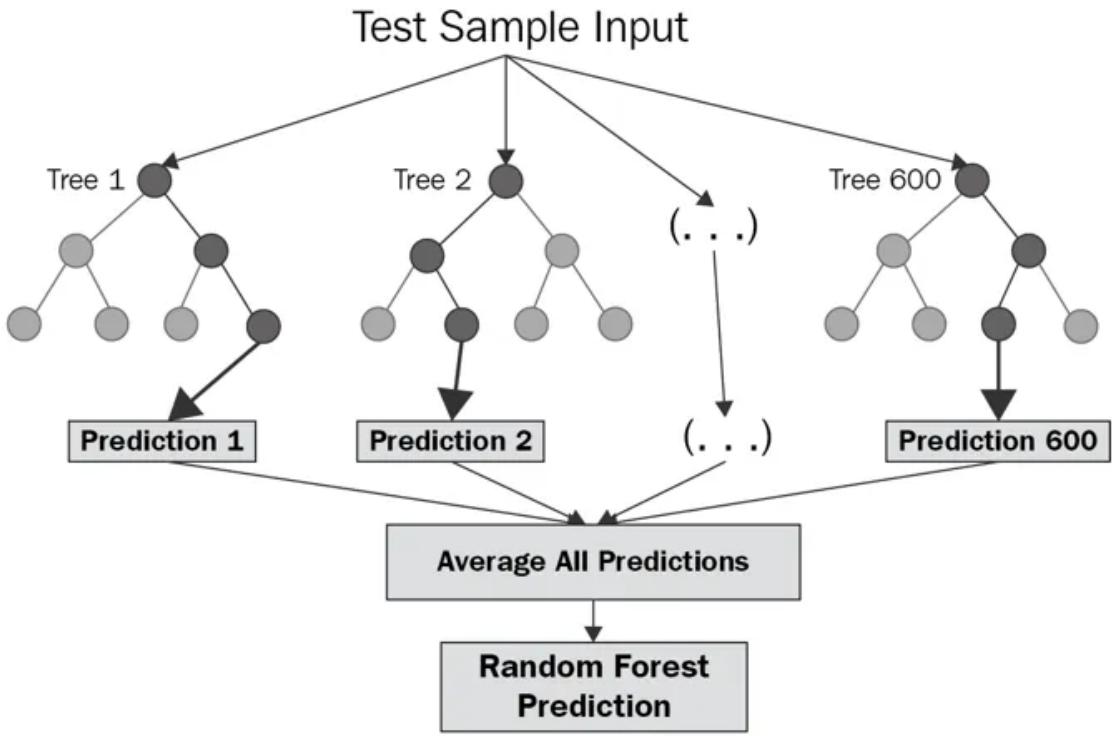
- Logistic Regression



Source: (Rajput, 2018)

Classification Methods

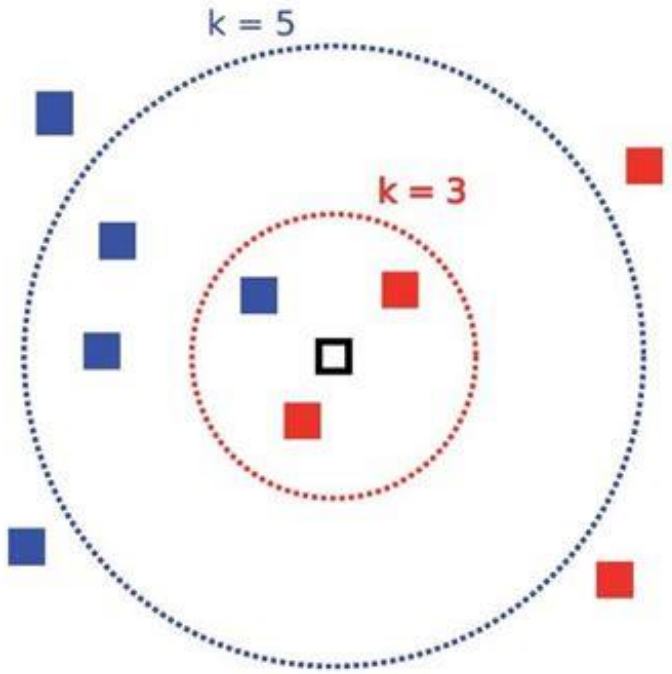
- Random Forest Classifier



Source: (corporatefinanceinstitute)

Classification Methods

- K Nearest Neighbors



Source: (ResearchGate)

Results

Logistic Regression

- Accuracy: 61%
- Recall: 0%
- Precision: 0%
- Specificity: 99%

Potable?	Yes	No
P- Yes	0	1
P- No	255	400

Random Forest

- Accuracy: 68%
- Recall: 35%
- Precision: 66%
- Specificity: 89%

Potable?	Yes	No
P- Yes	89	46
P- No	166	355

KNN

- Accuracy: 65%
- Recall: 25%
- Precision: 62%
- Specificity: 91%

Potable?	Yes	No
P- Yes	63	38
P- No	192	363

Results Plotted

Conclusions

- We can predict, with moderate accuracy, the potability of water using several machine learning algorithms
- KNN models were very accurate at identifying the samples that were not potable
- The models with only three predictor variables (pH, sulfate, chloramines) were still moderately successful

Future Considerations

- Bacterial levels are an important indicator of potability
- Consider using undersampling or oversampling due to unbalanced data
- Tune model based on other metrics

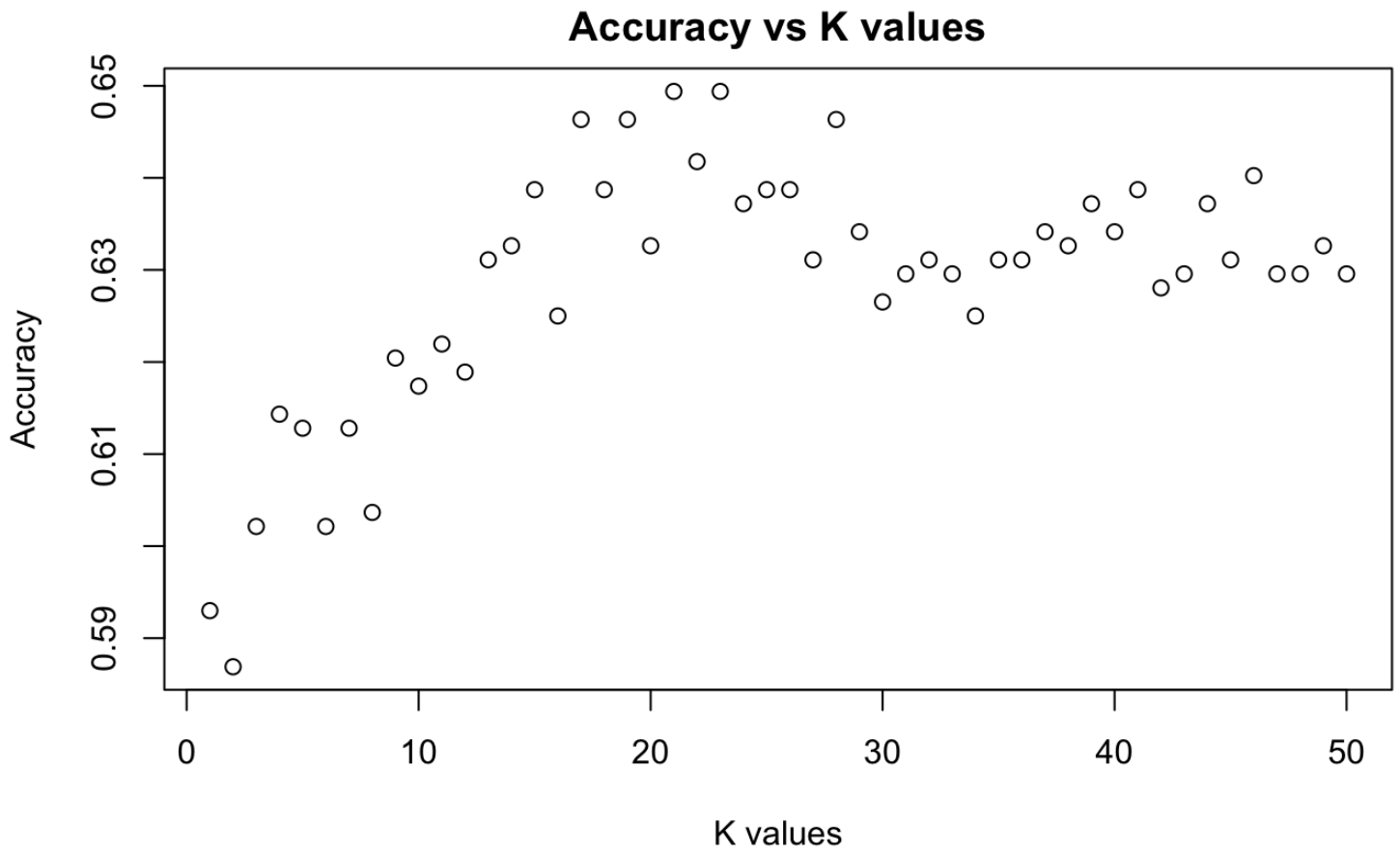
References

Environmental Protection Agency. (n.d.) National Primary Drinking Water Regulations. *EPA.gov.* https://www.epa.gov/sites/default/files/2016-06/documents/npwdr_complete_table.pdf

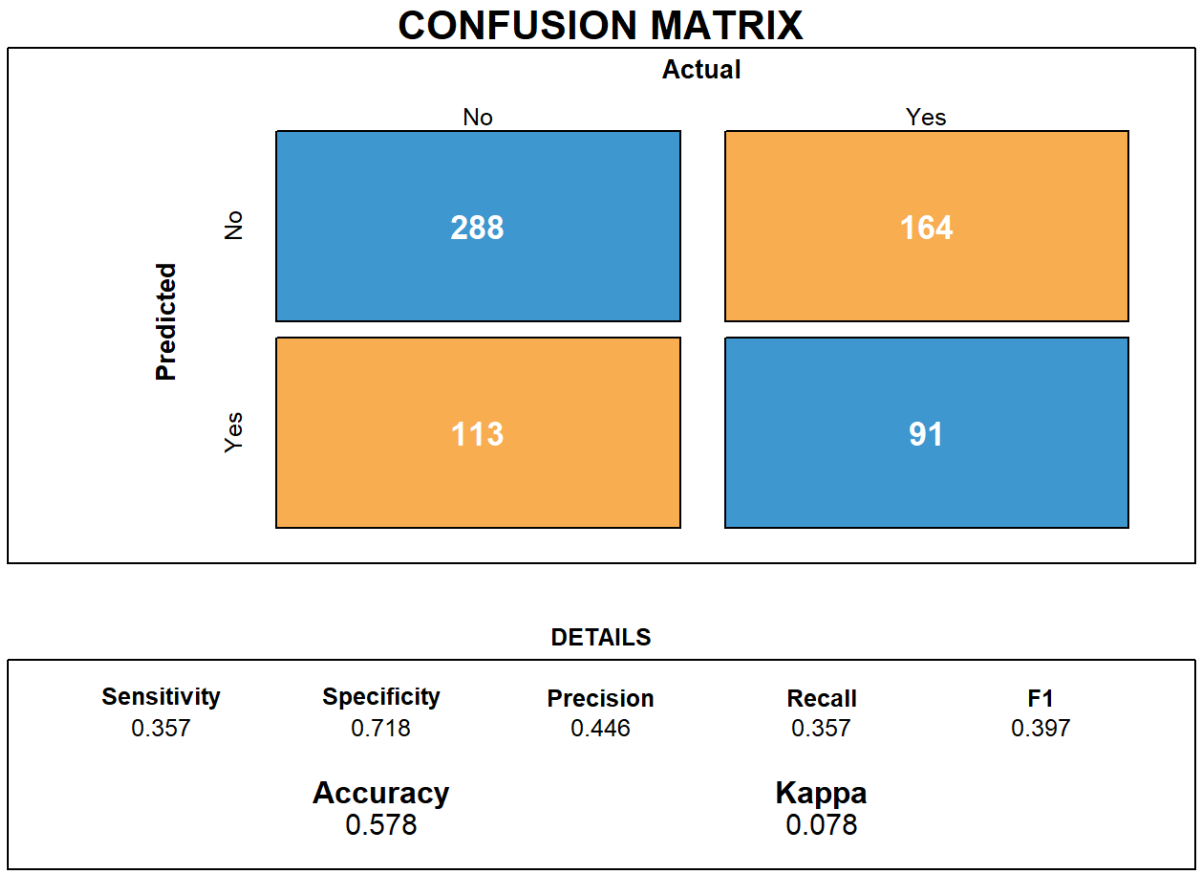
Kadiwal, Aditya. (n.d.) Water Quality. *Kaggle*. https://www.kaggle.com/datasets/adityakadiwal/water-potability?resource=download

Appendix

Tuning KNN


Accuracy vs K values
Accuracy
0.59
0.61
0.63
0.65
0
10
20
30
40
50
K values

Random Forest w/ Less predictors



KNN w/ Less predictors