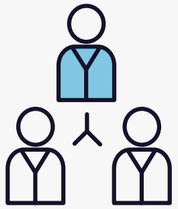
Water Potability in India

Statistical Learning (MOD. B) Exam

Gioele Ceccon 2079425

Pietro Renna 2089068

Roadmap



Reasoning

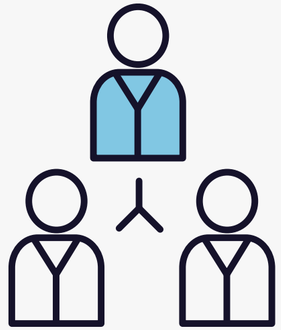
Data Presentation

Exploratory Analysis

Models

Conclusions

Reasoning



Water Potability in India

Several issues

Pollution

Lack of infrastructures

Contamined rivers

The task

Binary classification problem about the potability of water samples

Potable

Non-Potable

Goals

1. Obtain an effective method to predict wheter new samples are safe for human consumption
    - Different models evaluation
2. Identify the predictors that significantly affect water quality and assess their influence
    - Evaluate the relative importance of each factor in determining potability

Data Presentation

1361 observations: water samples

Predictors: 14 variables

Original Dataset

Geographical

- Stationcode
- Location
- Latitude
- Longitude
- CapitalCity
- State

Chemicals

- Nitrate
- DO
- BOD
- pH
- Conductivity

Generals

- Fecal coliform
- Total coliform
- Temperature

Response

Potability

Working Dataset

1361 observations: water samples

Predictors: 8 variables

Chemicals

- Nitrate
- DO
- BOD
- pH
- Conductivity

Generals

- Fecal coliform
- Total coliform
- Temperature

Response

Potability

A first look at the dataset

- Presence of NA values
- Presence of outliers
- Imbalanced Potability class

Temperature	DO	pH	Conductivity
Min. : 5.00	Min. : 0.000	Min. :5.000	Min. : 2.0
1st Qu.:23.50	1st Qu.: 6.329	1st Qu.:7.375	1st Qu.: 216.7
Median :26.21	Median : 7.200	Median :7.700	Median : 417.2
Mean :25.16	Mean : 7.054	Mean :7.605	Mean : 1232.1
3rd Qu.:27.76	3rd Qu.: 7.800	3rd Qu.:7.984	3rd Qu.: 781.6
Max. :65.00	Max. :30.367	Max. :9.575	Max. :65700.0
NA's :34	NA's :9	NA's :1	NA's :38

Table 3: Table continues below

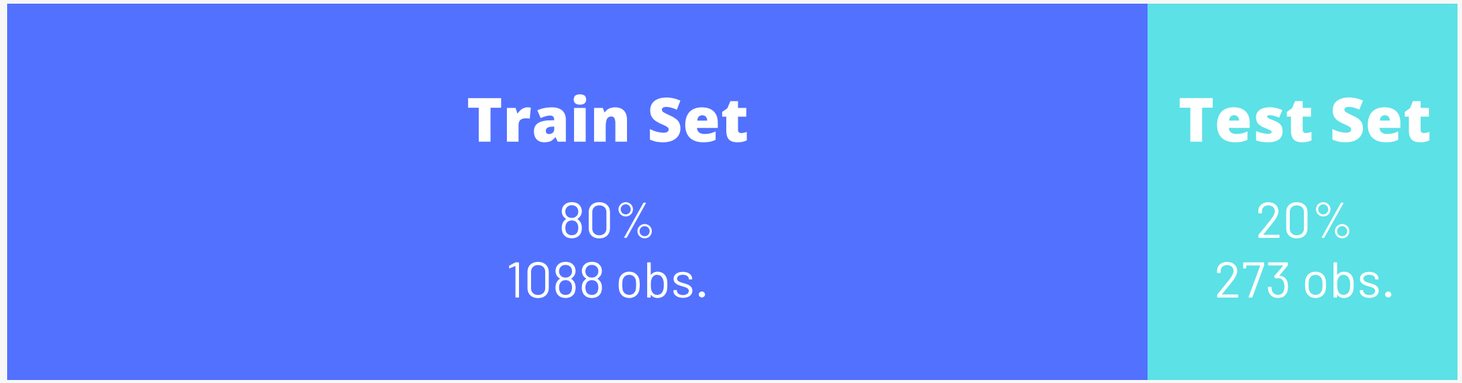
BOD	Nitrate	Fecalcoliform	Totalcoliform
Min. : 0.000	Min. : 0.0000	Min. : 1	Min. : 0
1st Qu.: 1.000	1st Qu.: 0.3336	1st Qu.: 14	1st Qu.: 2
Median : 2.060	Median : 0.9825	Median : 120	Median : 251
Mean : 4.191	Mean : 9.1467	Mean : 429155	Mean : 1431256
3rd Qu.: 3.933	3rd Qu.: 2.5428	3rd Qu.: 774	3rd Qu.: 1600
Max. :222.000	Max. :920.0000	Max. :230000000	Max. :670000000
NA's :68	NA's :210	NA's :189	NA's :135

Potability
0: 298
1:1063

Train Test Split

Full Dataset
1361 observations

Train Test Split



Checking Potability proportions after Train-Test split

Train

Potable	Non- Potable
0.784	0.216

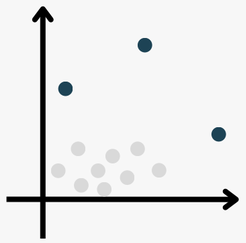
Test

Potable	Non- Potable
0.769	0.231

Pre-Processing

Set of techniques and data manipulation to ensure a proper format for the analysis

Outliers



- Handled with the *Modified Z-Score*
- **Only on Train Set**

NA Values

x1	x2	x3
1.67	*NA*	3.45

- Handled with the *Median imputation*
- Both Train and Test set

Exploratory Analysis

Univariate analysis

<u>Goal</u>: analyze and understand the distribution and characteristics of each variable

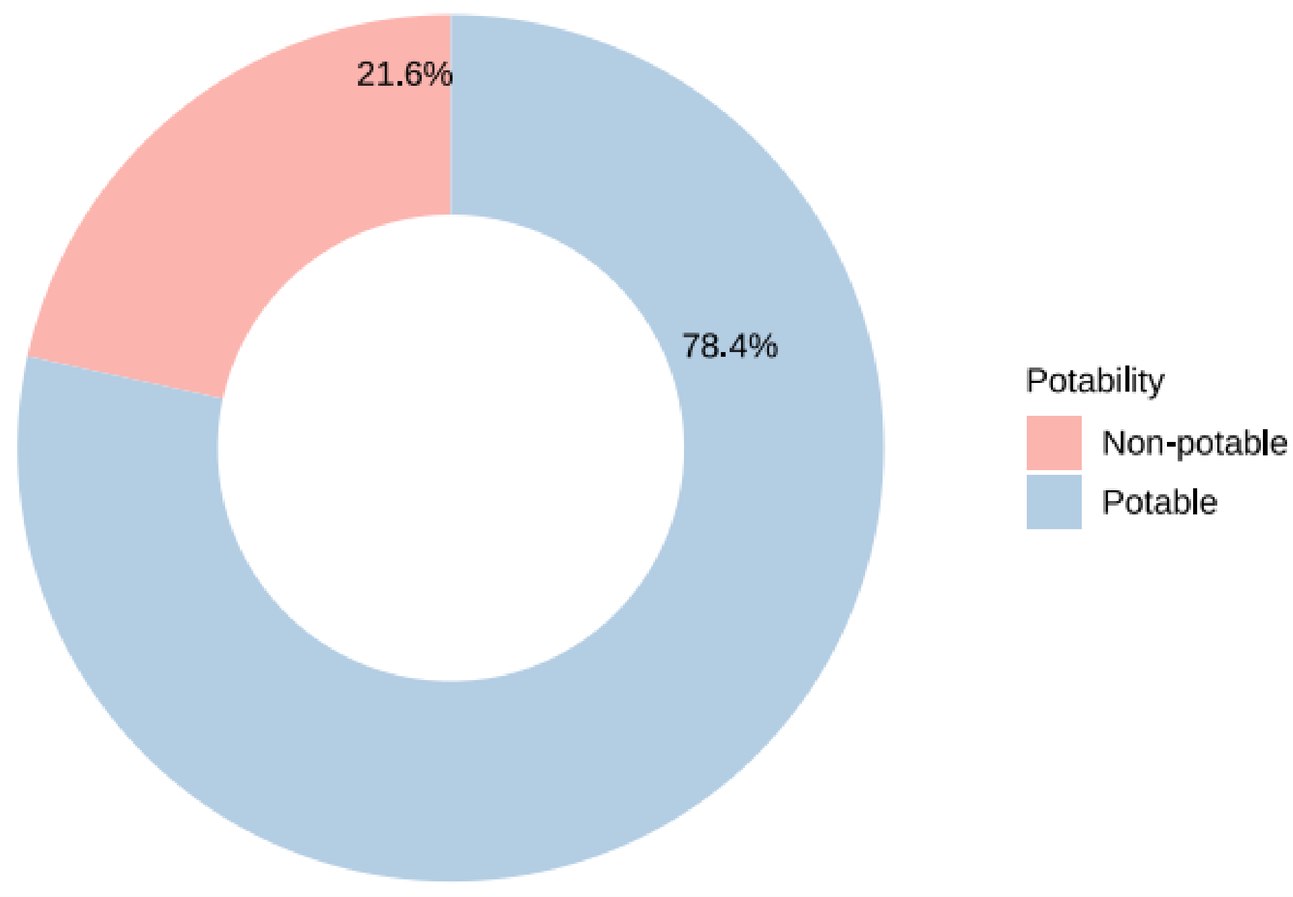
Behaviors which are worth mentioning on predictors are:

- Positive skewness: Conductivity, BOD, Nitrate
- Strong Positive Skewness: Fecalcoliform, Totalcoliform

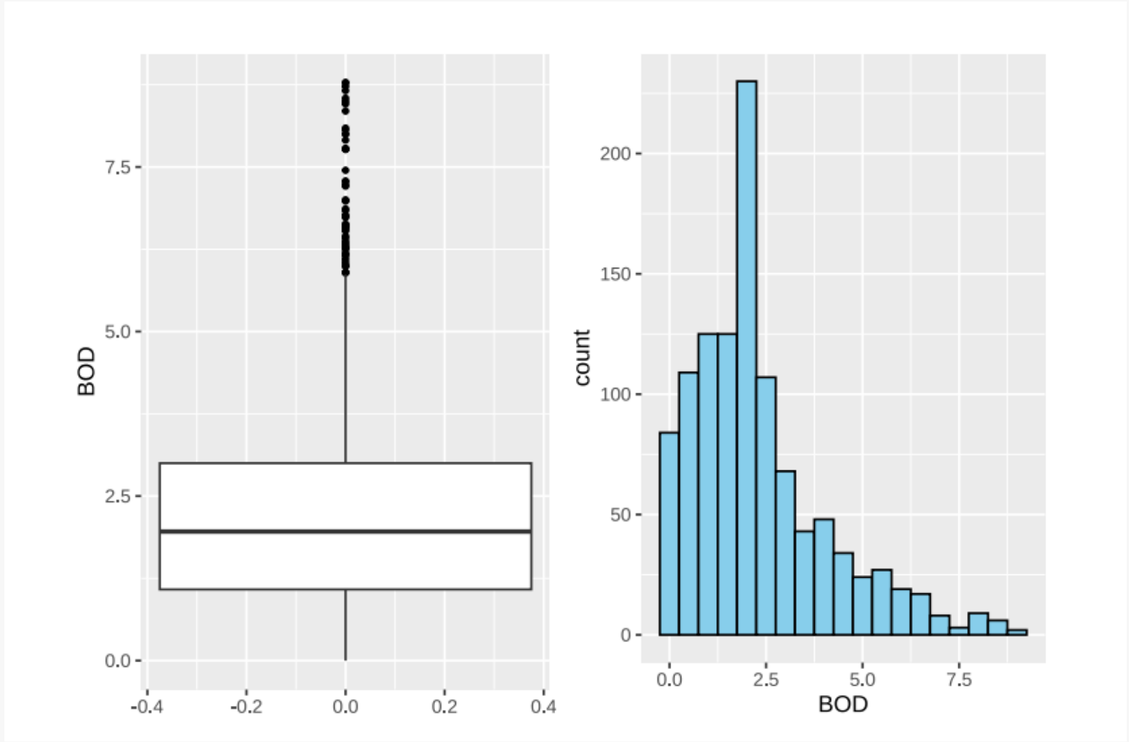
A first look to the Potability variable

Most of the observations are Potable

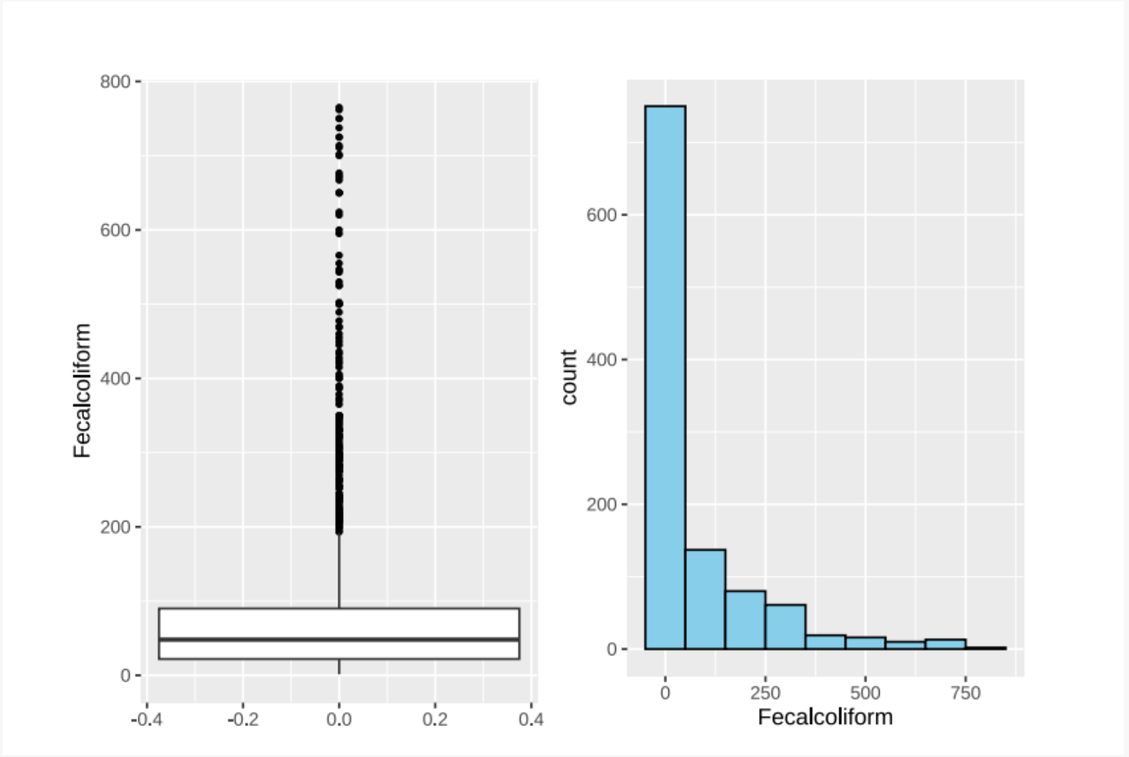
- Imbalanced class ⚠



Positive Skewness



Strong Positive Skewness



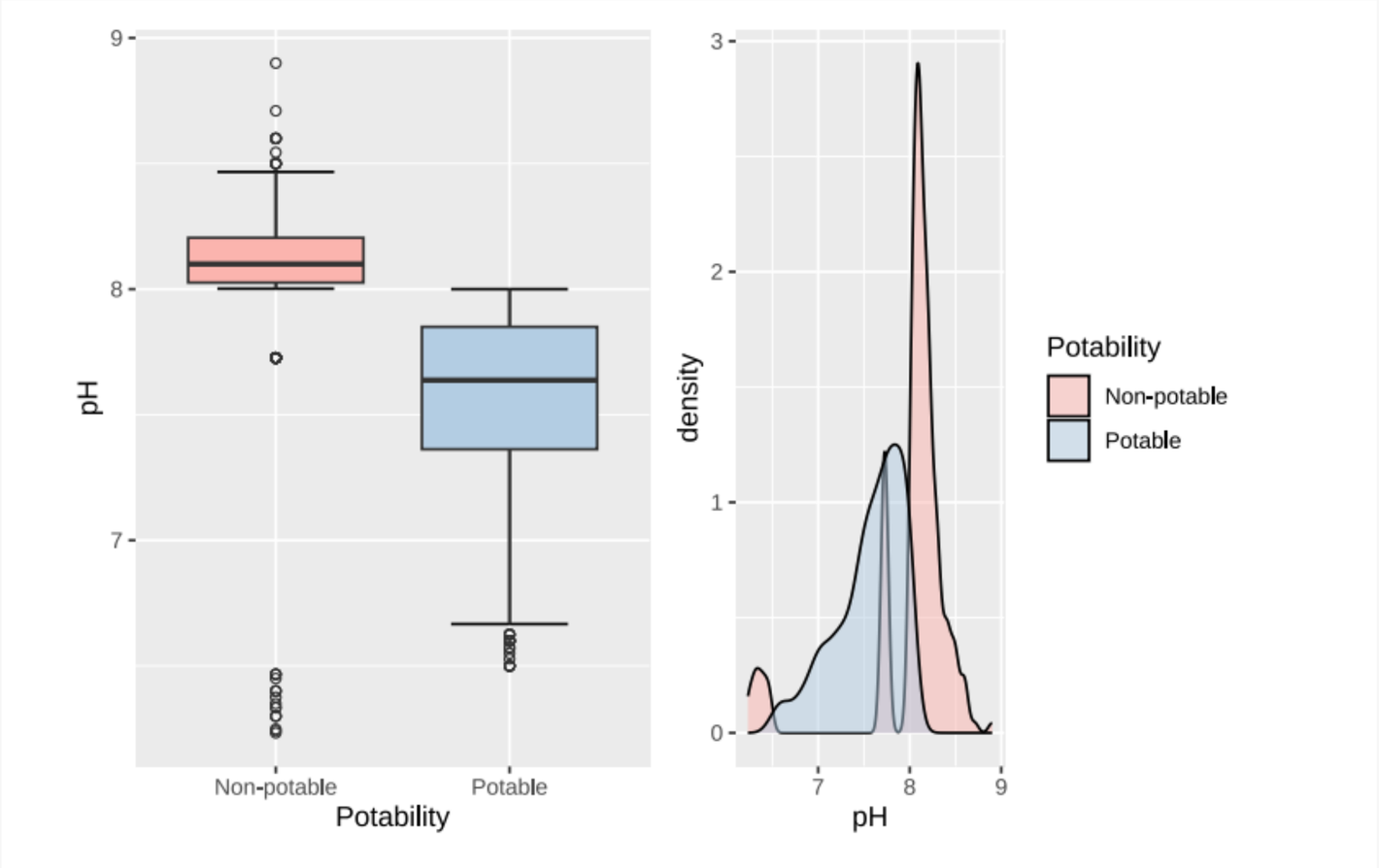
Bivariate analysis

<u>Goal</u>: analyze the relationship between the response variable and the explanatory variables

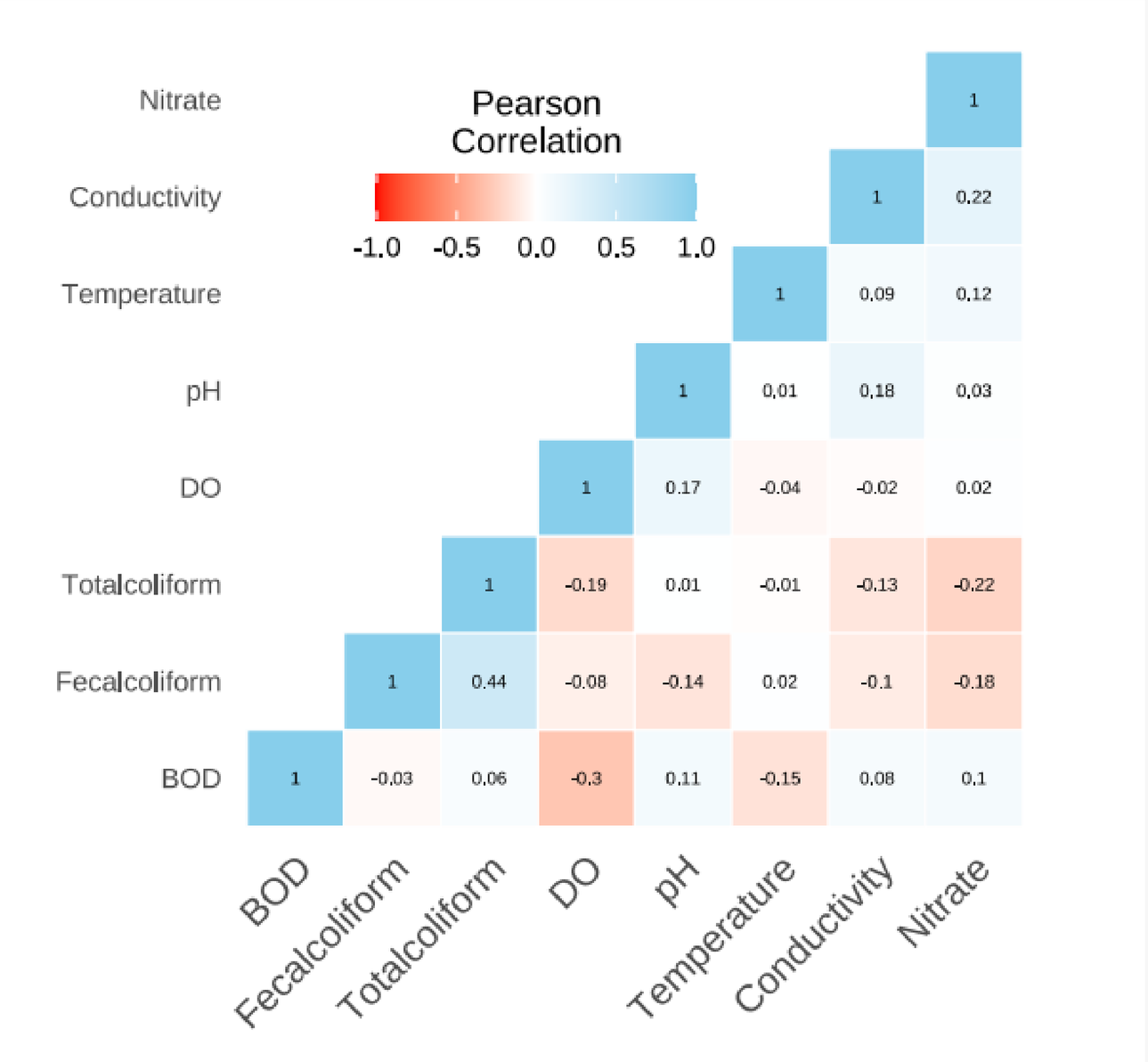
Behaviors which are worth mentioning are:

- pH is a significant variable
- All the other variables are not significant

A closer look: pH



Correlation matrix



Models

Models: methodology

Models are trained on a downsampled set having 470 observations due to imbalanced classes in the Train Set

Potable	Non- Potable
0.5	0.5

Evaluation metrics:

- Accuracy
- False Positive Rate
- AUC (Area Under the Curve)

Logistic Regression

- Binary Classification
- Linearly Separable Data

Features selected through stepwise selection with AIC criterion

	True Non-Potable	True Potable	Total
Pred. Non-Potable	48	48	96
Pred. Potable	15	162	177
Total	63	210	273

Accuracy: 0.77

False Positive rate: 0.24

Ridge Logistic Regression

- L2 Penalization
- Shrinks asymptotically to 0
- **Best λ = 0.021**

All features are kept.
The non-significant ones have coefficients values close to 0

	True Non-Potable	True Potable	Total
Pred. Non-Potable	43	35	78
Pred. Potable	20	175	195
Total	63	210	273

Accuracy: 0.80

False Positive rate: 0.32

Lasso Logistic Regression

- L1 Penalization
- Shrinks exactly to 0
- **Best λ = 0.00077**

Non-significant features have coefficients values equal to 0.

	True Non-Potable	True Potable	Total
Pred. Non-Potable	51	53	104
Pred. Potable	12	157	169
Total	63	210	273

Accuracy: 0.76

False Positive rate: 0.19

LDA

- Linear decision boundary
- Threshold at 0.5

All features are kept

	True Non-Potable	True Potable	Total
Pred. Non-Potable	42	40	82
Pred. Potable	21	170	191
Total	63	210	273

Accuracy: 0.78

False Positive rate: 0.33

QDA

- Non-linear, quadratic decision boundary
- Threshold at 0.6

All features are kept

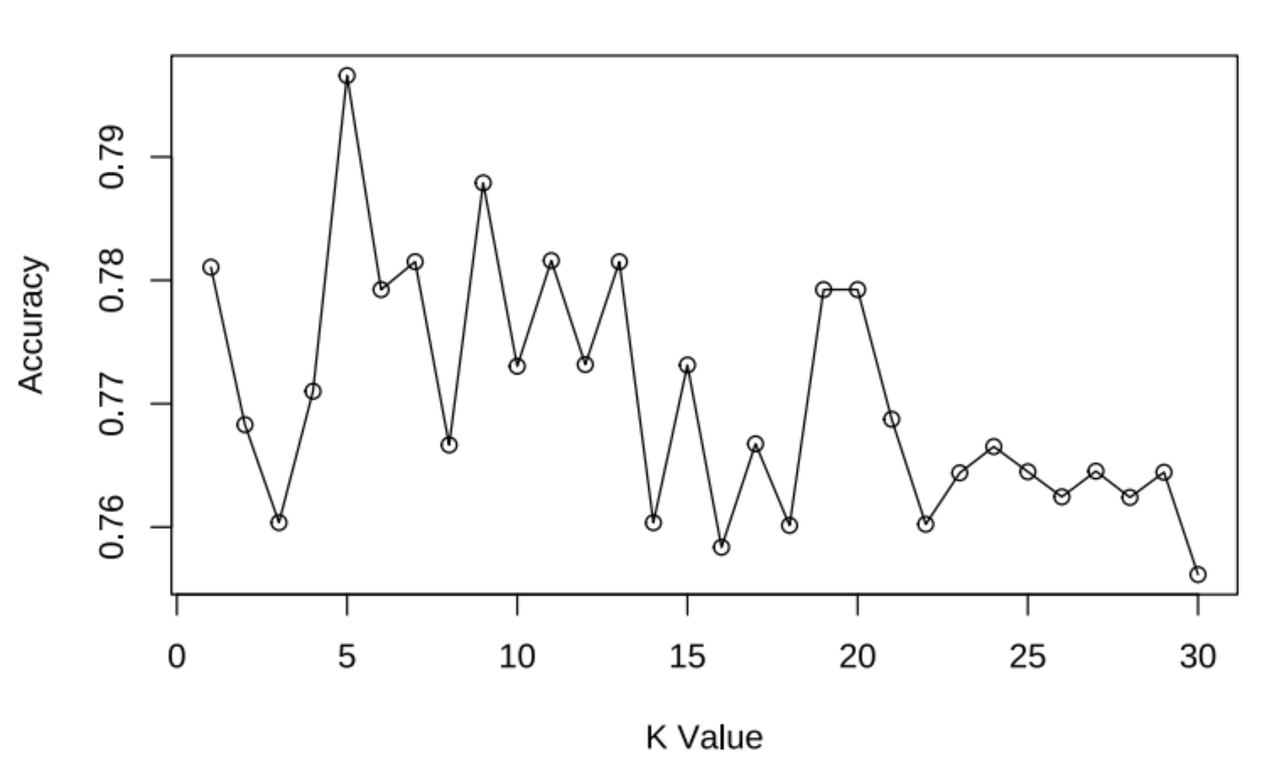
	True Non-Potable	True Potable	Total
Pred. Non-Potable	38	52	90
Pred. Potable	25	158	183
Total	63	210	273

Accuracy: 0.72

False Positive rate: 0.40

k-NN

- Non-parametric approach
- **Best k = 5**



Accuracy: 0.78

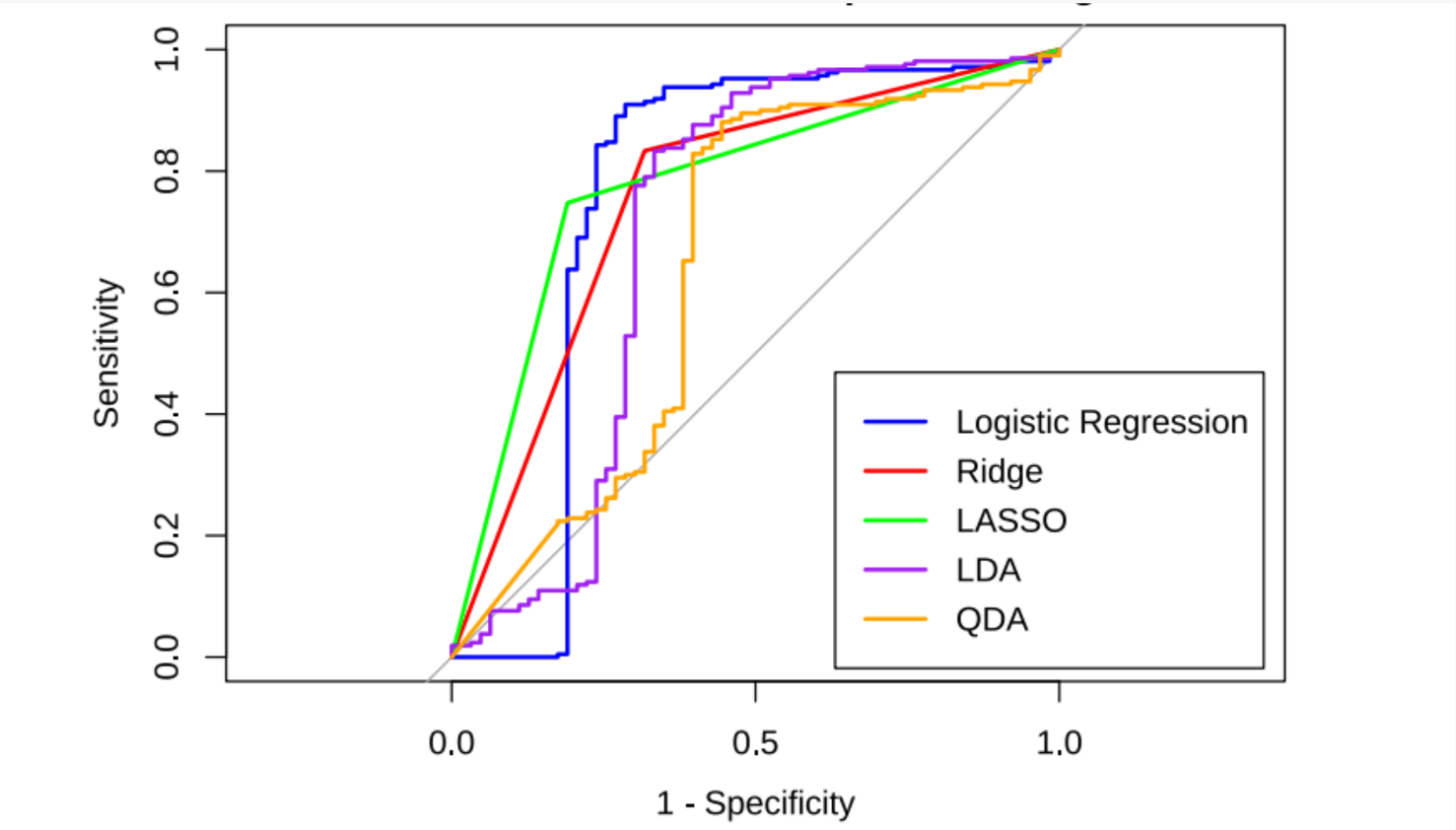
False Positive rate: 0.11

Conclusions

How to choose the best model?

- Accuracy does not take into account the context
- AUC is robust to:
  - different thresholds
  - how each model performed on False Positives and False Negatives
- K-NN can't help in determining the predictors impact on the response
- Alternative models trained on the original Train Set confirmed the necessity of fitting on downsampled set

ROC Curves and AUC



Model	AUC
Logistic Regression	0.758
Lasso	0.779
Ridge	0.758
LDA	0.701
QDA	0.644

GOAL 1

Best prediction method

Lasso model

Variables	Coefficients
Intercept	24.196
Temperature	-0.017
pH	-2.888

GOAL 2

Predictors impact on water Potability

- The higher the temperature, the lower the probability of being Potable

- The higher the pH, the lower the probability of being Potable