# Water Potability Prediction

By: Asmita Sharma & Pranshul Lakhanpal

# Project Goals

Predict if water is safe to drink

Binning the data to analyze prediction accuracy

Fine-tuning models to increase accuracy

# Dataset

| | ph | Hardness | Solids | Chloramines | Sulfate | Conductivity | Organic_carbon | Trihalomethanes | Turbidity | Potability |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0.277251 | 0.576927 | -0.558278 | 0.165643 | 0.071408 | 0.704187 | 0.048638 | -0.027069 | 0.409947 | 0 |
| **1** | 0.256573 | -0.323581 | -0.255784 | 0.644735 | 0.893819 | 0.599758 | -0.667969 | 2.399087 | 0.149515 | 0 |
| **2** | 0.327804 | 0.135948 | -0.303081 | 0.351662 | 0.457220 | 0.303033 | -2.978220 | 0.508450 | -0.552626 | 0 |
| **3** | 0.133429 | 0.174622 | 0.007963 | 0.300014 | 0.173823 | -0.586414 | -2.000227 | -0.152054 | 0.331065 | 0 |
| **4** | -0.075814 | -0.600905 | 0.123585 | -0.460244 | -0.428084 | 1.002543 | 0.273601 | -1.903931 | 1.070481 | 1 |
| **5** | -0.823529 | -1.199389 | -1.609874 | -0.302042 | -1.710241 | -0.605860 | 0.903045 | 0.185227 | -0.479140 | 1 |
| **6** | 2.846723 | -0.019913 | -0.890440 | 0.050243 | -0.509441 | -0.855290 | 1.475200 | 0.984400 | -1.073323 | 0 |
| **7** | -0.190597 | 0.020302 | 0.081847 | -1.262508 | 1.648857 | -0.052163 | 0.460845 | -1.149488 | 0.679718 | 1 |
| **8** | 0.833135 | 1.580147 | 0.697142 | -0.035253 | 0.253002 | 0.083438 | 1.277747 | -0.816606 | 2.194008 | 0 |
| **9** | 1.912965 | -1.957923 | -1.192612 | 0.885102 | -0.226390 | 0.198643 | -0.295934 | 1.593668 | 1.469213 | 1 |

**Pre-processed Water Potability Dataset
from Kaggle**
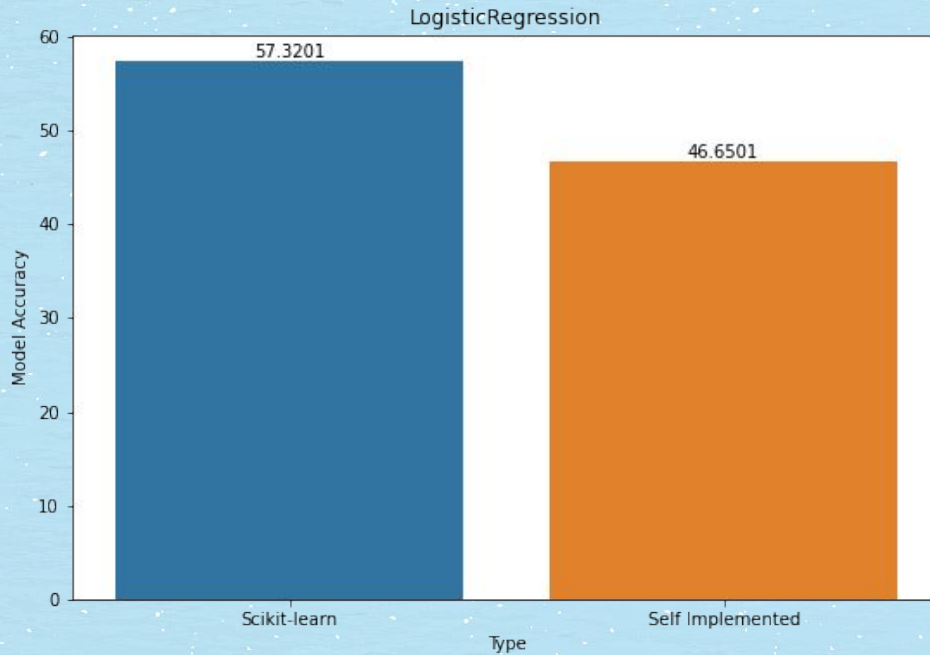
# Models

**01**

Logistic Regression

**02**

K-Nearest Neighbors
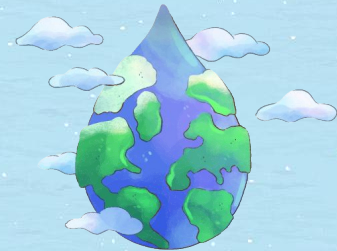
**03**

Gaussian Naive Bayes

# Logistic Regression
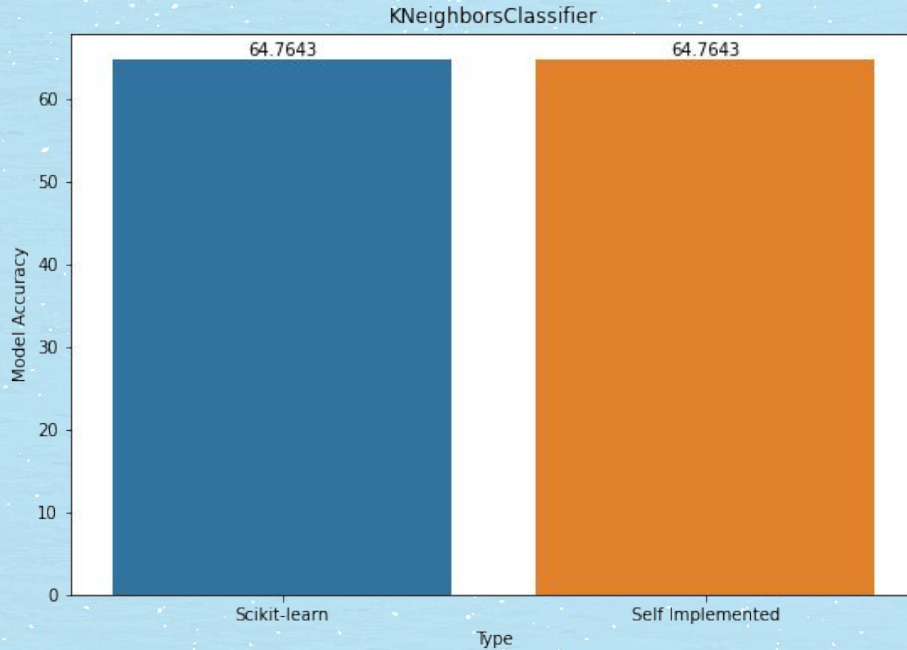


LogisticRegression

SciKit-learn accuracy: 57.32%

Self-implemented accuracy: 46.65%
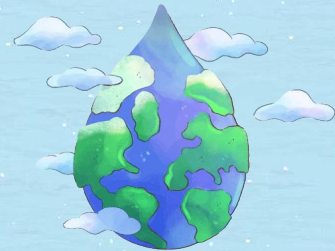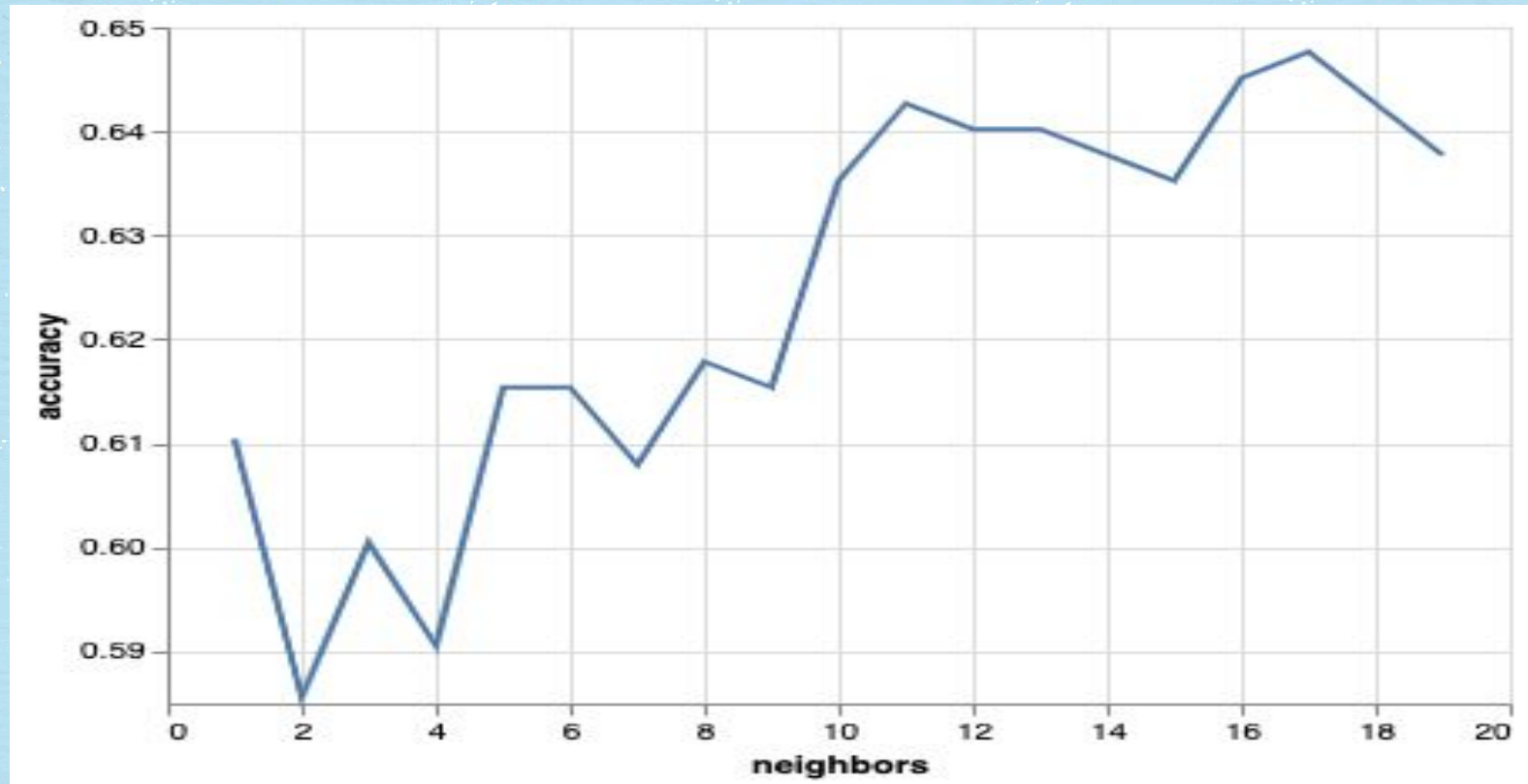
# K-Nearest Neighbors



SciKit-learn accuracy: 64.76%

Self-implemented accuracy: 64.76%
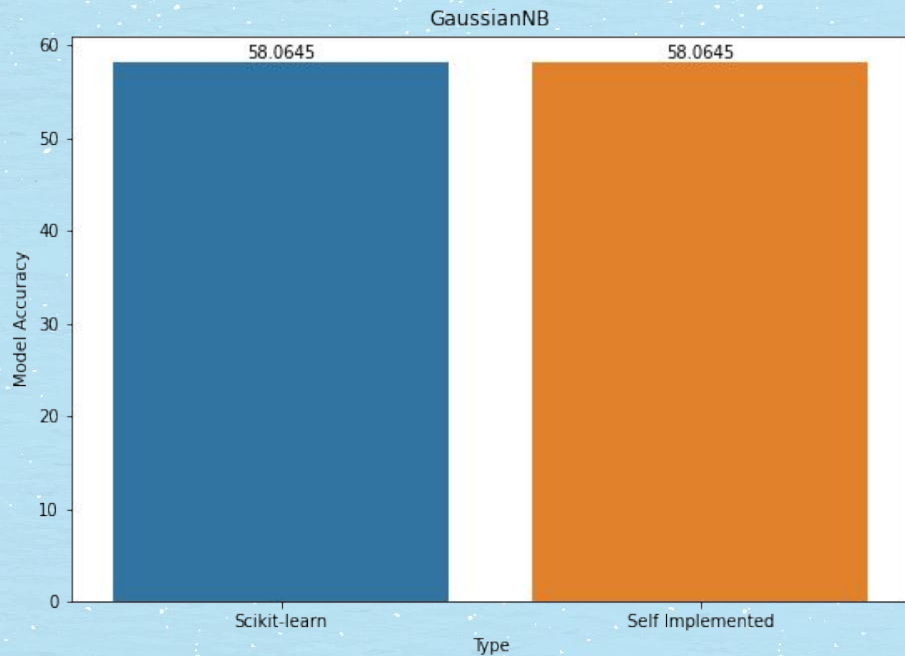
Distance metric: Euclidean

Neighbors = 17

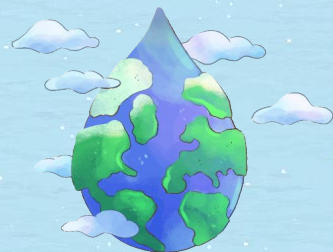**KNN Model accuracy based on different number of neighbors**

# Gaussian Naive Bayes



SciKit-learn accuracy: 58.0645%

Self-implemented accuracy: 58.0645%

# Possible reasons for low accuracy

Value sensitive data

Too many dimensions

Feature potability range

# Future Work

Predict accuracy using Decision Tree, PCA

Binning the data to analyze prediction accuracy

Fine-tuning models to increase accuracy

# Thanks!