Water Quality

Data Mining

Project Documentation

Building of Classifier Models

To Predict the Water Potability & Calculate the Error Percentage

- TA. **Nourhan Bahnasy**

Name	Section	University ID
أحمد ناصر أحمد حسن	**1**	**20191701016**
يوسف عصام فؤاد محمد	**9**	**20191701269**
مريم عبدالهادي محمد عبدالغفار	**9**	**20191701195**
أمانى جمال رسلان حسن	**2**	**20191701033**
جنى هانى أحمد صادق	**3**	**20191701059**
عبدالرحمن يسري إبراهيم البابلى	**5**	**20191701124**

Used Models

2	Naive-Bayes Classifier
3	KNN (ver1) Classifier
4	KNN (ver2) Classifier
5	Random-Forest Classifier
7	Gradient-Boosting Classifier
8	SGD Classifier
9	ID3 Classifier
10	Logistic Classifier
11	SVM (ver1) Classifier
12	SVM (ver2) Classifier
13	SVM (ver3) Classifier
14	SVM (new dataset) Classifier
15	XG-Boost Classifier

For the first dataset "waterQuality1", the best model that succeeded to get the highest testing accuracy is:

Random-Forest Classifier

---

1) Naive-Bayes *Classifier*

a) **Dataset**

i) **Used the first *"waterQuality1"* dataset**

b) **Data Preprocessing**

i) **Sort values**

ii) **Drop duplicates**

iii) **Fill data missing with mean to improve accuracy**

iv) **Shuffling**

v) **Stratify**

vi) **Label Encoder**

vii) **Standard Scaler**

c) **Testing Accuracy**

i) **63.262195121951216 %**

d) **Data Separation**

i) **80% training : 20% testing**

e) **Result & Output**

i) **Screen Shot**

```
Confusion matrix
 [[361 202]
 [ 39  54]]
Accuracy ----> 63.262195121951216
Mean square error -----> 0.3673780487804878

 Cross validation
 [0.61259542 0.63931298 0.63167939 0.60496183 0.60877863]
```

2) KNN (ver1) Classifier

a) Dataset

i) Used the first *"waterQuality1"* dataset

b) Data Preprocessing

i) Drop NULL values
ii) Drop the duplicates
iii) Normalization
iv) Selection of the important features
v) Stratified Sampling

c) Testing Accuracy

i) 62.03473945409429 %

d) Data Separation

i) 80% training : 20% testing

e) Result & Output

i) Screen Shot

```
Feautres:  Index(['Solids', 'Chloramines', 'Sulfate', 'Conductivity', 'Organic_carbon',
       'Trihalomethanes', 'Turbidity'],
      dtype='object')
Result is:  [0 0 1 0 1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 0 0 1 0
 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 1 0 0 1 0 0 0 1 0
 1 0 0 1 0 0 0 0 0 1 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 1 1 0 0
 0 1 0 0 0 1 0 0 0 1 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 1 1 0 0 0 0 1 0 1 0 1 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0
 0 1 1 1 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 1 1 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0
 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0
 1 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 1 0 1 1 0 1
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 0
 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1]

   Prediction Score is:  62.03473945409429

   Mean Square Error  0.37965260545905705
```

3) KNN (ver2) Classifier

a) Dataset

i) Used the first *"waterQuality1"* dataset

b) Data Preprocessing

i) Dropping NULL values
ii) Dropping Duplicates
iii) Extraction of the first 5 features
iv) Set the neighbors to 5
v) Normalization
vi) Shuffling

c) Testing Accuracy

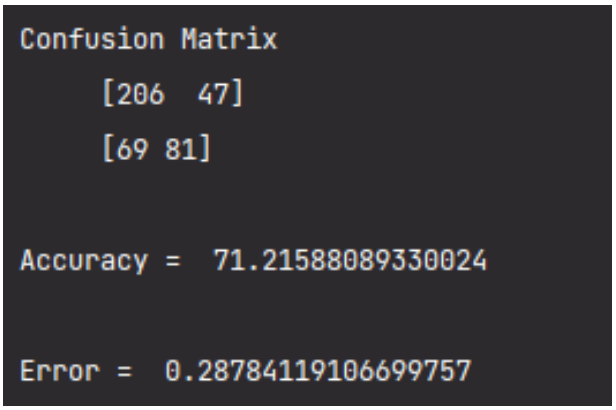
i) 71.21588089330024 %

d) Data Separation

i) 80% training : 20% testing

e) Result & Output

i) Screen Shot



4) Random-Forest Classifier

a) **Dataset**

i) **Used the first *"waterQuality1"* dataset**

b) **Data Preprocessing**

i) **Sort values**

ii) **Drop Duplicates**

iii) **Fill data missing with mean to improve accuracy**

iv) **Shuffle**

v) **Stratify**

vi) **Label Encoder**

vii) **Standard Scaler**

c) **Testing Accuracy**

i) **73.10469314079422 %**

d) **Data Separation**

i) **80% training : 20% testing**

e) **Result & Output**

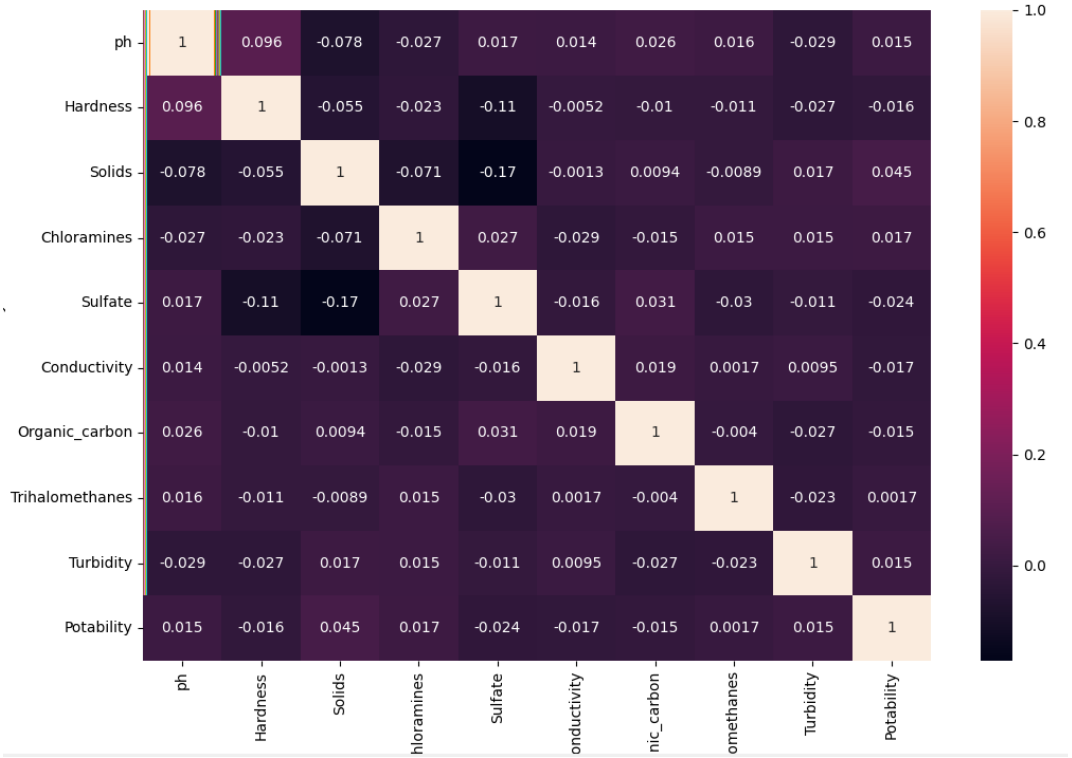
i) **Screen Shot**

```
GaussianNB Classifier
     Confusion Matrix
         [305 166]
         [30 53]
     Accuracy =  64.62093862815884
     Mean Squared Error =  0.35379061371841153

cross_val_score GaussianNB
 [0.57948718 0.66494845 0.61340206 0.59793814 0.59793814 0.63402062
 0.62371134 0.59278351 0.60824742 0.63402062]

RandomForest Classifier
     Confusion Matrix
         [317 131]
         [18 88]
     Accuracy =  73.10469314079423
     Mean Squared Error =  0.26895306859205775
```

ii) Plotting


ph 1 0.096 -0.078 -0.027 0.017 0.014 0.026 0.016 -0.029 0.015
Hardness 0.096 1 -0.055 -0.023 -0.11 -0.0052 -0.01 -0.011 -0.027 -0.016
Solids -0.078 -0.055 1 -0.071 -0.17 -0.0013 0.0094 -0.0089 0.017 0.045
Chloramines -0.027 -0.023 -0.071 1 0.027 -0.029 -0.015 0.015 0.015 0.017
Sulfate 0.017 -0.11 -0.17 0.027 1 -0.016 0.031 -0.03 -0.011 -0.024
Conductivity 0.014 -0.0052 -0.0013 -0.029 -0.016 1 0.019 0.0017 0.0095 -0.017
Organic_carbon 0.026 -0.01 0.0094 -0.015 0.031 0.019 1 -0.004 -0.027 -0.015
Trihalomethanes 0.016 -0.011 -0.0089 0.015 -0.03 0.0017 -0.004 1 -0.023 0.0017
Turbidity -0.029 -0.027 0.017 0.015 -0.011 0.0095 -0.027 -0.023 1 0.015
Potability 0.015 -0.016 0.045 0.017 -0.024 -0.017 -0.015 0.0017 0.015 1
ph
Hardness
Solids
hloramines
Sulfate
onductivity
nic_carbon
omethanes
Turbidity
Potability
1.0
0.8
0.6
0.4
0.2
0.0

5) Gradient-Boosting Classifier

a) **Dataset**

i) **Used the first *"waterQuality1"* dataset**

b) **Data Preprocessing**

i) **Removing duplicates**

ii) **Replace outliers by NULLs**

iii) **Replace NULLs by mean value**

c) **Testing Accuracy**

i) **max: 65.110 %**

d) **Data Separation**

i) **80% training 20% testing**

e) **Result & Output**

i) **Screen Shot**

```
Learning rate:  0.1     Training Accuracy: 0.626    Validation Accuracy: 63.736 %

Learning rate:  0.2     Training Accuracy: 0.634    Validation Accuracy: 63.599 %

Learning rate:  0.3     Training Accuracy: 0.648    Validation Accuracy: 63.324 %

Learning rate:  0.4     Training Accuracy: 0.657    Validation Accuracy: 64.698 %

Learning rate:  0.5     Training Accuracy: 0.664    Validation Accuracy: 65.110 %

Learning rate:  0.6     Training Accuracy: 0.673    Validation Accuracy: 62.500 %

Learning rate:  0.7     Training Accuracy: 0.673    Validation Accuracy: 64.148 %

Learning rate:  0.8     Training Accuracy: 0.672    Validation Accuracy: 63.324 %

Learning rate:  0.9     Training Accuracy: 0.672    Validation Accuracy: 64.560 %

Learning rate:  1    Training Accuracy: 0.680    Validation Accuracy: 63.324 %
```

6) SGD Classifier

a) **Dataset**

i) **Used the first *"waterQuality1"* dataset**

b) **Data Preprocessing**

i) **Removing duplicates**

ii) **Replace outliers by nulls**

iii) **Replace nulls by mean value**

c) **Testing Accuracy**

i) **62.8%**

d) **Data Separation**

i) **80% training 20% testing**

e) **Result & Output**

i) **Screen Shot**

7) ID3 Classifier

a) **Dataset**

i) **Used the first *"waterQuality1"* dataset**

b) **Data Preprocessing**

i) **Dropping NULL values.**

ii) **Normalization.**

iii) **Shuffling data.**

iv) **Random Sampling.**

c) **Testing Accuracy**

i) **61.36645962732919 %**

d) **Data Separation**

i) **60% training : 40% testing**

e) **Result & Output**

i) **Screen Shot**

```
Test Accuracy:  61.49068322981367 %

Mean Square Error:  0.38509316770186336
```

8) Logistic Classifier

a) **Dataset**

i) **Used the first *"waterQuality1"* dataset**

b) **Data Preprocessing**

i) **Removing rows that has NULL values.**
ii) **Normalization.**
iii) **Shuffling data.**
iv) **Random Sampling**

c) **Testing Accuracy**

i) **60.12422360248447 %**

d) **Data Separation**

i) **60% training : 40% testing**

e) **Result & Output**

i) **Screen Shot**

```
Test Accuracy:  60.12422360248447 %
Mean Square Error:  0.3987577639751553


Report:
              precision    recall  f1-score   support

           0       0.61      0.97      0.75       492
           1       0.28      0.02      0.03       313

    accuracy                           0.60       805
   macro avg       0.44      0.49      0.39       805
weighted avg       0.48      0.60      0.47       805
```

9) SVM (ver1) Classifier

a) Dataset

i) Used the first *"waterQuality1"* dataset

b) Data Preprocessing

i) Scaling feature data using MINMAXScaler
ii) Handling NULL values with Median in the same column.
iii) Using a Simple Imputer.
iv) Shuffling data.
v) Random Sampling.

c) Testing Accuracy

i) 61.35531135531136 %

d) Data Separation

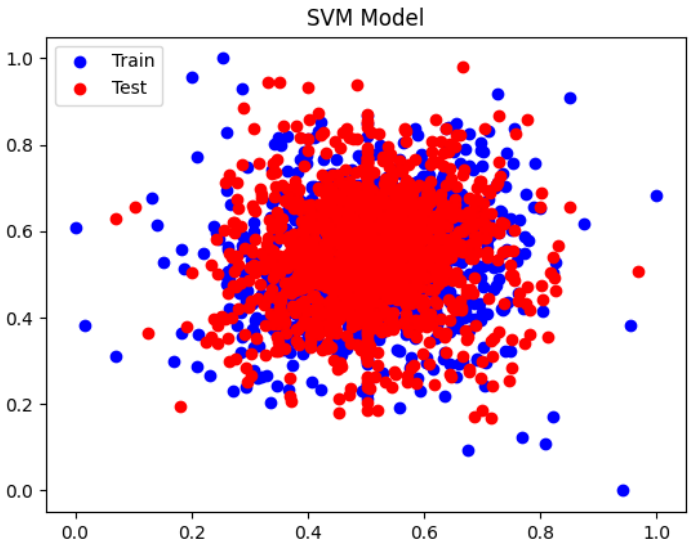
i) 50% training : 50% testing

e) Result & Output

i) Screen Shot

```
Train Accuracy :  60.62271062271062 %
Test Accuracy  :  61.35531135531136 %
Mean Square Error :  0.38644688644688646
```

ii) Plotting



10) SVM (ver2) Classifier

a) Dataset

i) Used the first *"waterQuality1"* dataset

b) Data Preprocessing

i) Remove null values

ii) Dropping for PH and Sulfate features

iii) Remove Duplicates

c) Testing Accuracy

i) 65.1603498542274 %

d) Data Separation

i) 80% training : 20% testing

e) Result & Output

i) Screen Shot

11) SVM (ver3) Classifier

a) Dataset

i) Used the first *"waterQuality1"* dataset

b) Data Preprocessing

i) Used Stratified-kFold, cross-validator

ii) Removing duplicates

iii) Replace NULLs by mean value

iv) Replace outliers by NULLs

c) Testing Accuracy

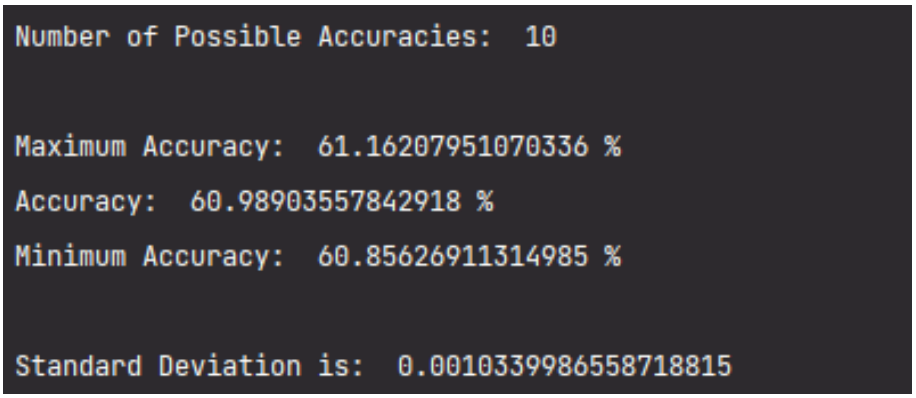
i) max: 61.16207951070336 %

d) Data Separation

i) 80% training 20% testing

e) Result & Output

i) Screen Shot



12) SVM (new dataset) Classifier

a) Dataset

i) Used the Second *"waterQuality2"* (new dataset)

b) Data Preprocessing

i) Convert objects data types into float & integers

ii) Normalization

c) Testing Accuracy

i) 90.0 %

d) Data Separation

i) 80% training : 20% testing

e) Result & Output

i) Screen Shot

```
Accuracy: 0.9
```

13) XG-Boost Classifier

i) **Data PreprocessingNormalization**
ii) **Remove duplicates**
iii) **Replace null values by mean**

b) Testing Accuracy

i) **65.88072122052705 %**

c) Data Separation

i) **80% training : 20% testing**

d) Result & Output

i) **Screen Shot**