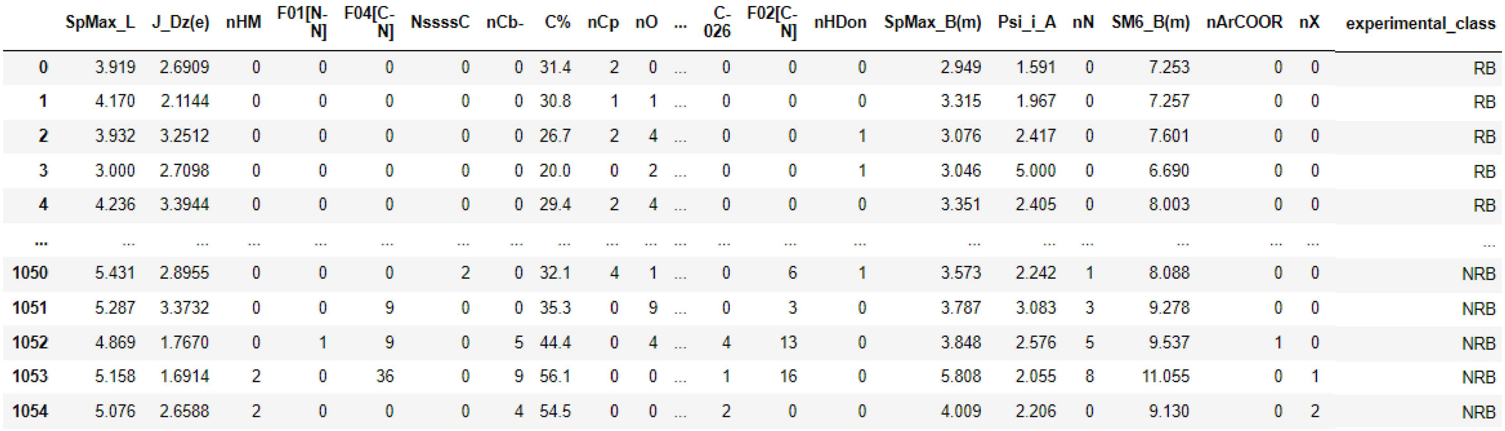
# CPC 251 Group Project Part 1 QSAR_6

NG WEI YI(147630),LEE JOE XUEN(154277), LUO YUJIE(154625),YAP YU XIAN(152795)

# QSAR-Quantitative Structure-Activity Relationship

## DATASET DESCRIPTION

The QSAR biodegradation dataset was obtained from the Milano Chemometrics and QSAR Research Group (UniversitÃ degli Studi Milano â€“ Bicocca, Milano, Italy). The dataset contained 1055 samples of chemicals, each with 41 inputs and one of them is a binary target which is named as "experimental_class" in which the output is RB(ready biodegradable) and NRB(not ready biodegradable).



| | SpMax_L | J_Dz(e) | nHM | F01[N-N] | F04[C-N] | NssssC | nCb- | C% | nCp | nO | ... | C-026 | F02[C-N] | nHDon | SpMax_B(m) | Psi_i_A | nN | SM6_B(m) | nArCOOR | nX | experimental_class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 3.919 | 2.6909 | 0 | 0 | 0 | 0 | 0 | 31.4 | 2 | 0 | ... | 0 | 0 | 0 | 2.949 | 1.591 | 0 | 7.253 | 0 | 0 | RB |
| **1** | 4.170 | 2.1144 | 0 | 0 | 0 | 0 | 0 | 30.8 | 1 | 1 | ... | 0 | 0 | 0 | 3.315 | 1.967 | 0 | 7.257 | 0 | 0 | RB |
| **2** | 3.932 | 3.2512 | 0 | 0 | 0 | 0 | 0 | 26.7 | 2 | 4 | ... | 0 | 0 | 1 | 3.076 | 2.417 | 0 | 7.601 | 0 | 0 | RB |
| **3** | 3.000 | 2.7098 | 0 | 0 | 0 | 0 | 0 | 20.0 | 0 | 2 | ... | 0 | 0 | 1 | 3.046 | 5.000 | 0 | 6.690 | 0 | 0 | RB |
| **4** | 4.236 | 3.3944 | 0 | 0 | 0 | 0 | 0 | 29.4 | 2 | 4 | ... | 0 | 0 | 0 | 3.351 | 2.405 | 0 | 8.003 | 0 | 0 | RB |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **1050** | 5.431 | 2.8955 | 0 | 0 | 0 | 2 | 0 | 32.1 | 4 | 1 | ... | 0 | 6 | 1 | 3.573 | 2.242 | 1 | 8.088 | 0 | 0 | NRB |
| **1051** | 5.287 | 3.3732 | 0 | 0 | 9 | 0 | 0 | 35.3 | 0 | 9 | ... | 0 | 3 | 0 | 3.787 | 3.083 | 3 | 9.278 | 0 | 0 | NRB |
| **1052** | 4.869 | 1.7670 | 0 | 1 | 9 | 0 | 5 | 44.4 | 0 | 4 | ... | 4 | 13 | 0 | 3.848 | 2.576 | 5 | 9.537 | 1 | 0 | NRB |
| **1053** | 5.158 | 1.6914 | 2 | 0 | 36 | 0 | 9 | 56.1 | 0 | 0 | ... | 1 | 16 | 0 | 5.808 | 2.055 | 8 | 11.055 | 0 | 1 | NRB |
| **1054** | 5.076 | 2.6588 | 2 | 0 | 0 | 0 | 4 | 54.5 | 0 | 0 | ... | 2 | 0 | 0 | 4.009 | 2.206 | 0 | 9.130 | 0 | 2 | NRB |

Table 1: Dataset samples

# DATA ANALYSIS

The dataset is analyzed in order to gain insight from the dataset. Data visualization such as pie chart is used to show the ratio of samples in terms of ready for biodegradation(RB) and not ready for biodegradation(NRB).

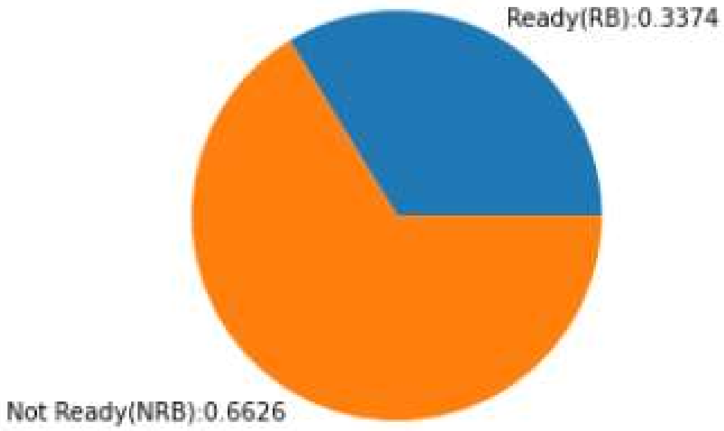
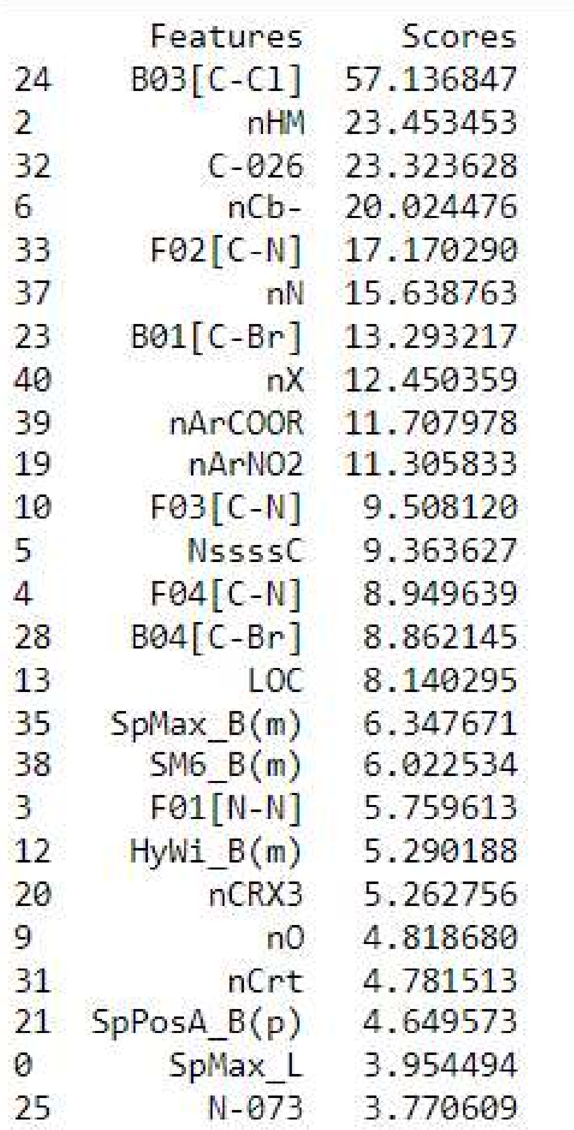


Figure 1: Ratio of output (RB & NRB)

Based on the pie chart, the ratio of samples that are ready to be biodegradable is 0.3374, which is 356 of 1055 samples. The ratio of samples that are not ready to be biodegradable is 0.6626 which is 699 of 1055 samples.



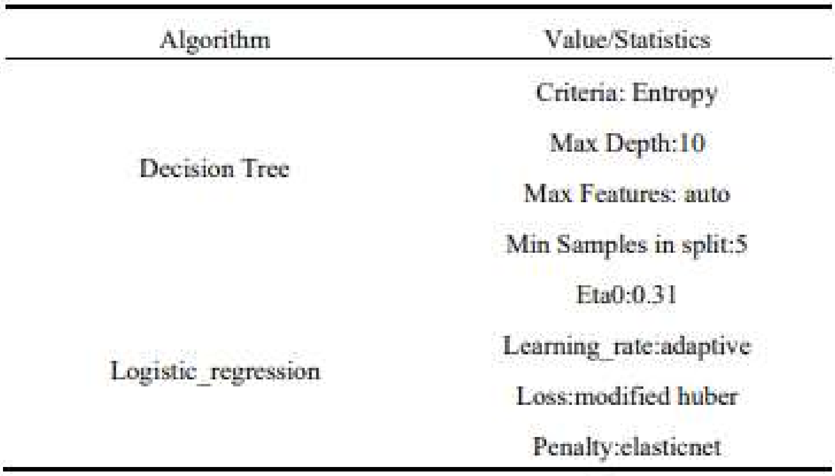
```
        Features     Scores
24   B03[C-Cl]  57.136847
2          nHM  23.453453
32       C-026  23.323628
6         nCb-  20.024476
33     F02[C-N]  17.170290
37          nN  15.638763
23   B01[C-Br]  13.293217
40          nX  12.450359
39     nArCOOR  11.707978
19      nArNO2  11.305833
10    F03[C-N]   9.508120
5       NssssC   9.363627
4     F04[C-N]   8.949639
28   B04[C-Br]   8.862145
13         LOC   8.140295
35  SpMax_B(m)   6.347671
38    SM6_B(m)   6.022534
3     F01[N-N]   5.759613
12   HyWi_B(m)   5.290188
20       nCRX3   5.262756
9           nO   4.818680
31        nCrt   4.781513
21 SpPosA_B(p)   4.649573
0      SpMax_L   3.954494
25       N-073   3.770609
```

Figure 2: 25 relevant features selected.

In the construction of the machine learning model, 25 relevant features are selected since the number of features is tested to be the optimum number.

# DATA MODELING

Two predictive models are built using Decision Tree and Logistic Regression algorithms. The models are evaluated using the hold-out method. The ratio of the split is 80% training set and 20% test set. 20% of the training set is split to be the validation set. The parameters of the predictive models are given in Table 2.

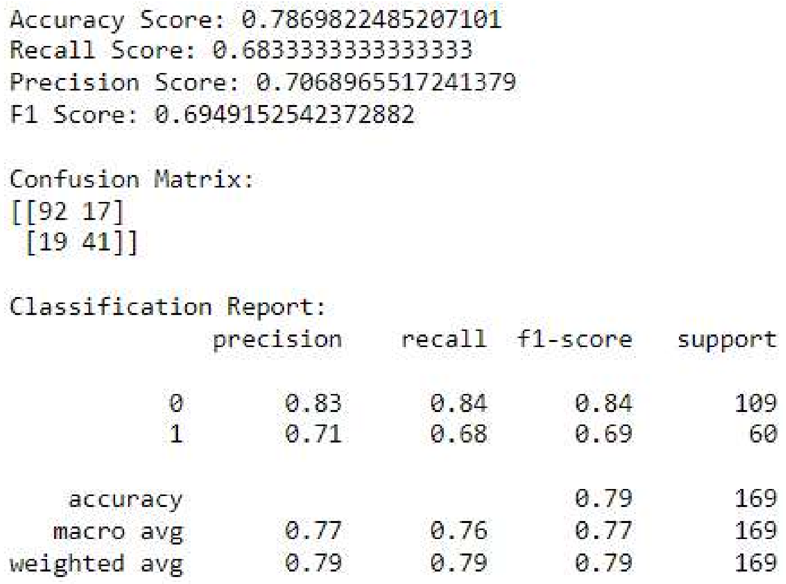


| Algorithm | Value/Statistics |
|---|---|
| Decision Tree | Criteria: Entropy<br>Max Depth:10<br>Max Features: auto<br>Min Samples in split:5 |
| Logistic_regression | Eta0:0.31<br>Learning_rate:adaptive<br>Loss:modified huber<br>Penalty:elasticnet |

Table 2: Parameters of the predictive models.

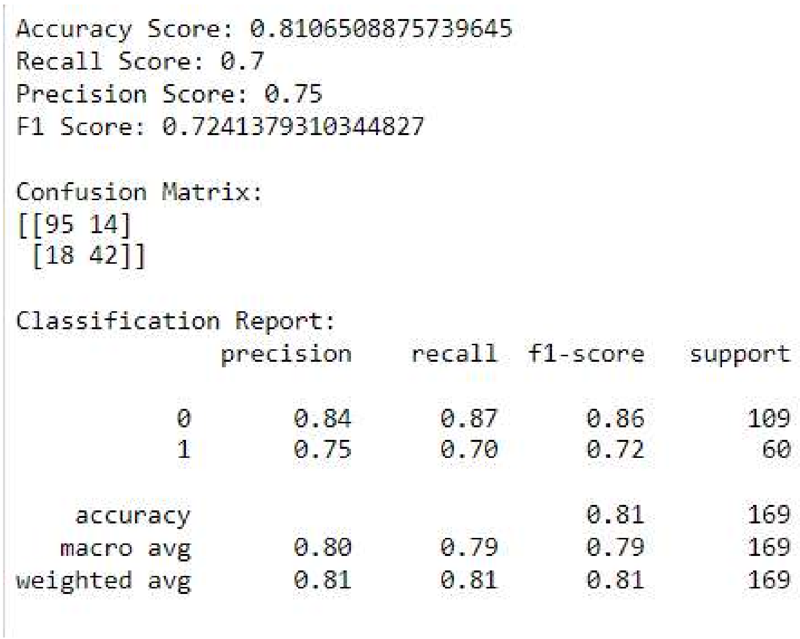
## Model Evaluation

The results of the classification of each evaluation are shown below.



```
Accuracy Score: 0.7869822485207101
Recall Score: 0.6833333333333333
Precision Score: 0.7068965517241379
F1 Score: 0.6949152542372882

Confusion Matrix:
[[92 17]
 [19 41]]

Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.84      0.84       109
           1       0.71      0.68      0.69        60

    accuracy                           0.79       169
   macro avg       0.77      0.76      0.77       169
weighted avg       0.79      0.79      0.79       169
```

Figure 3: Result of model evaluation of Decision Tree model



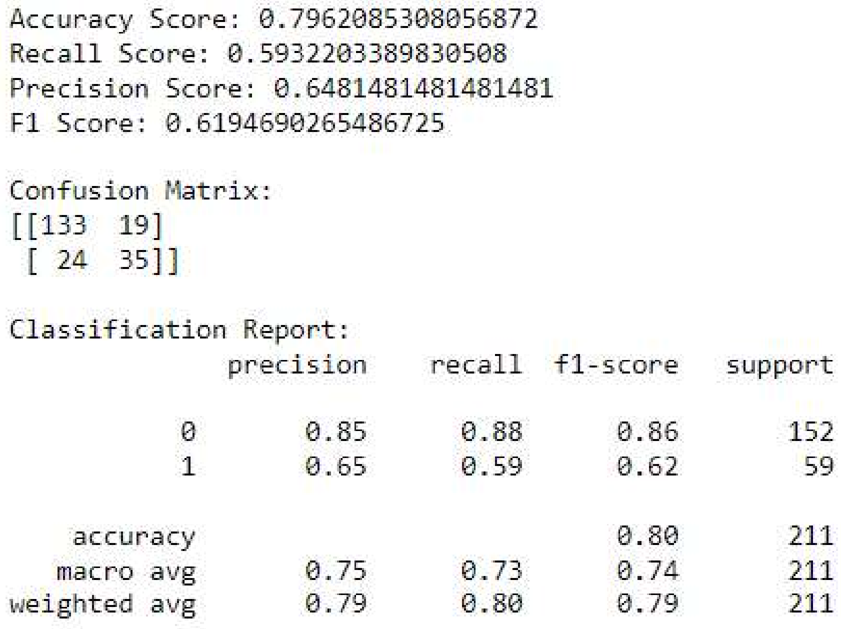
```
Accuracy Score: 0.8106508875739645
Recall Score: 0.7
Precision Score: 0.75
F1 Score: 0.7241379310344827

Confusion Matrix:
[[95 14]
 [18 42]]

Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.87      0.86       109
           1       0.75      0.70      0.72        60

    accuracy                           0.81       169
   macro avg       0.80      0.79      0.79       169
weighted avg       0.81      0.81      0.81       169
```

Figure 4: Result of model evaluation of Logistic Regression model

The machine learning model which is the logistic regression model has a better performance than the decision tree model. This is because the accuracy score of the logistic regression model which is 81% is higher than the decision tree model with an accuracy score of 78.7%.

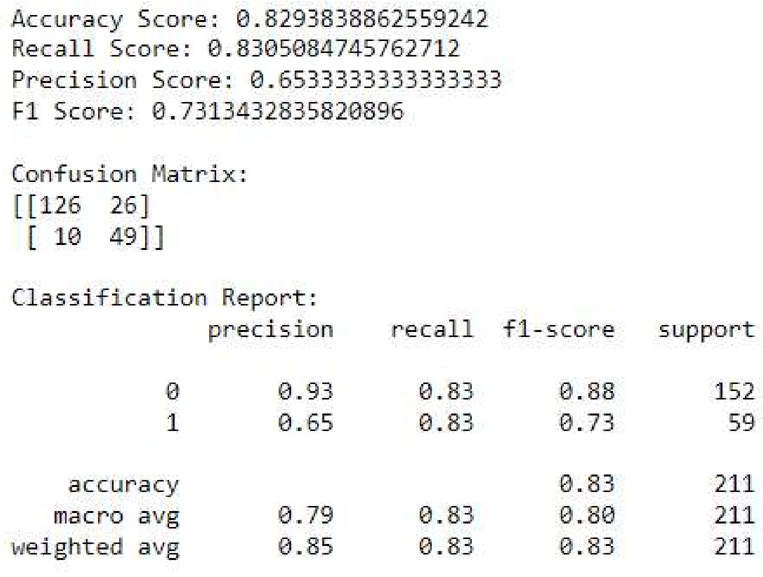
## Model Prediction

The results of the classification of each predictive model are given below.



```
Accuracy Score: 0.7962085308056872
Recall Score: 0.5932203389830508
Precision Score: 0.6481481481481481
F1 Score: 0.6194690265486725

Confusion Matrix:
[[133  19]
 [ 24  35]]

Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.88      0.86       152
           1       0.65      0.59      0.62        59

    accuracy                           0.80       211
   macro avg       0.75      0.73      0.74       211
weighted avg       0.79      0.80      0.79       211
```

Figure 5: Results of classification using the Decision Tree model.



```
Accuracy Score: 0.8293838862559242
Recall Score: 0.8305084745762712
Precision Score: 0.6533333333333333
F1 Score: 0.7313432835820896

Confusion Matrix:
[[126  26]
 [ 10  49]]

Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.83      0.88       152
           1       0.65      0.83      0.73        59

    accuracy                           0.83       211
   macro avg       0.79      0.83      0.80       211
weighted avg       0.85      0.83      0.83       211
```

Figure 6: Results of classification using Logistic Regression.

According to the prediction results of the decision tree and logistic regression models shown in Figures 5 and 6, the recall of logistic regression reaches 83.1% and the precision rate reaches 65.3%, while the recall rate and the precision rate of the decision tree are 59.3% and 64.8% respectively. This is due to the fact that logistic regression is a linear function, which is better for data with only one decision boundary, while a decision tree is a nonlinear function, which is better for data containing multiple decision boundaries. In a nutshell, the classification effect is better for this data set using the logistic regression model.

**Credits:**

Created with an image by Евгений Вершинин - "Small Plastic pellets. Micro plastic. air pollution."