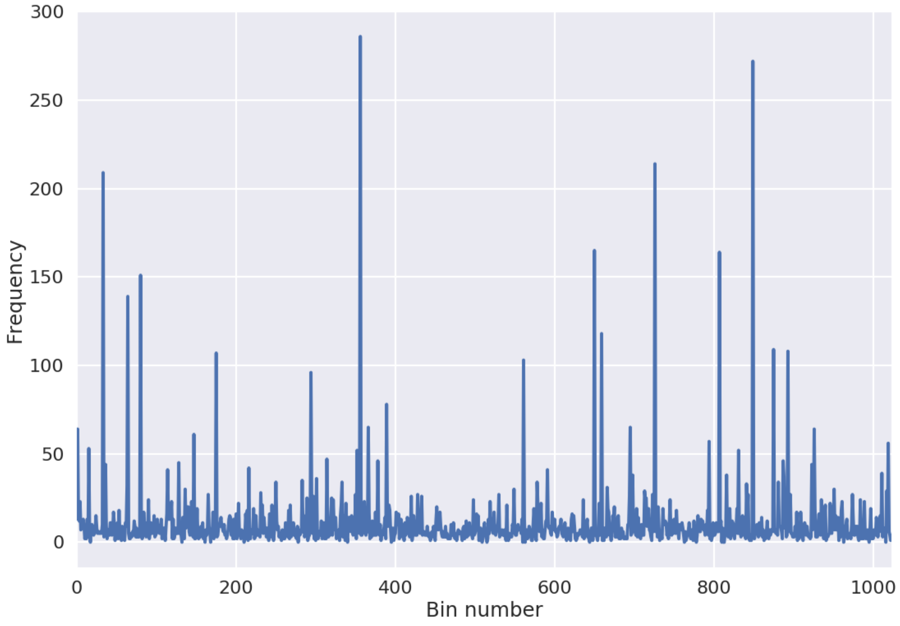
Machine Learning Performance Report

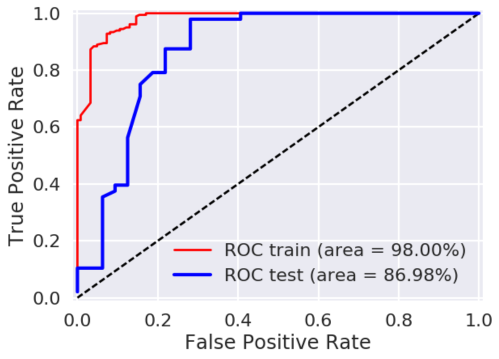
data_solubility

```
                          NaiveBayes
test  ACC                  81.250000
      AUC                  86.979167
      Cohen_Kappa           0.603175
      Matthews_corr         0.605083
      Precision             0.823529
      Recall                0.875000
      f1-score             84.848485
train ACC                  91.776316
      AUC                  97.998922
      Cohen_Kappa           0.827743
      Matthews_corr         0.828707
      Precision             0.914894
      Recall                0.950276
      f1-score             93.224932
```

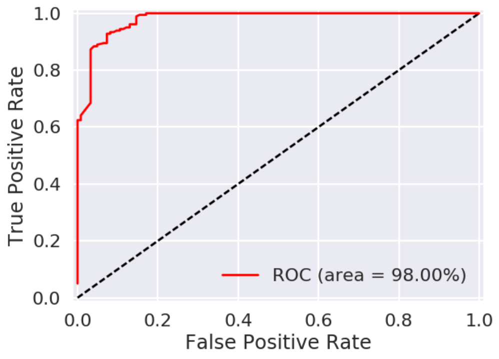
Frequency of fingerprints occurance in the bins for entire dataset



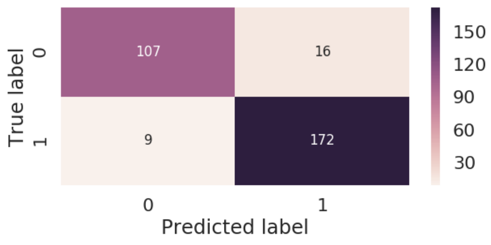
Receiver operating characteristic for Bernoulli Naive Bayes with isotonic class weights classifier



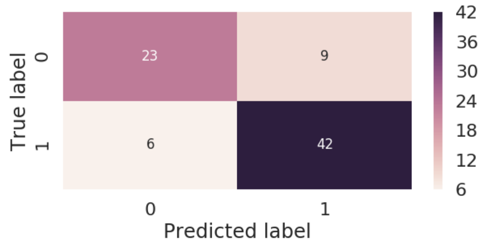
Receiver operating characteristic for Bernoulli Naive Bayes with isotonic class weights classifier



Confusion matrix for train dataset



Confusion matrix for test dataset



```
Original dataset:

Major class is: 1 sample size: 1142

Minor class is: 0 sample size: 155

Original major class sample size is:1142

New major class sample size is: 229

New dataset:

Major class is: 1 sample size: 229

Minor class is: 0 sample size: 155

                0           1           2          3           4           5  \
count  384.000000  384.000000  384.000000  384.00000  384.000000  384.000000   
mean     0.033854    0.166667    0.036458    0.03125    0.059896    0.018229   
std      0.181090    0.373164    0.187672    0.17422    0.237603    0.133954   
min      0.000000    0.000000    0.000000    0.00000    0.000000    0.000000   
25%      0.000000    0.000000    0.000000    0.00000    0.000000    0.000000   
50%      0.000000    0.000000    0.000000    0.00000    0.000000    0.000000   
75%      0.000000    0.000000    0.000000    0.00000    0.000000    0.000000   
max      1.000000    1.000000    1.000000    1.00000    1.000000    1.000000   

                6           7           8           9  ...        1015  \
count  384.000000  384.000000  384.000000  384.000000  ...  384.000000   
mean     0.023438    0.028646    0.033854    0.023438  ...    0.020833   
std      0.151486    0.167027    0.181090    0.151486  ...    0.143012   
min      0.000000    0.000000    0.000000    0.000000  ...    0.000000   
25%      0.000000    0.000000    0.000000    0.000000  ...    0.000000   
50%      0.000000    0.000000    0.000000    0.000000  ...    0.000000   
75%      0.000000    0.000000    0.000000    0.000000  ...    0.000000   
max      1.000000    1.000000    1.000000    1.000000  ...    1.000000   

        1016        1017        1018        1019        1020        1021  \
count  384.0  384.000000  384.000000  384.000000  384.000000  384.000000   
mean     0.0    0.075521    0.018229    0.145833    0.015625    0.010417   
std      0.0    0.264575    0.133954    0.353400    0.124181    0.101662   
min      0.0    0.000000    0.000000    0.000000    0.000000    0.000000   
25%      0.0    0.000000    0.000000    0.000000    0.000000    0.000000   
50%      0.0    0.000000    0.000000    0.000000    0.000000    0.000000   
75%      0.0    0.000000    0.000000    0.000000    0.000000    0.000000   
max      0.0    1.000000    1.000000    1.000000    1.000000    1.000000   

             1022        1023     Soluble  
count  384.000000  384.000000  384.000000  
mean     0.002604    0.010417    0.596354  
std      0.051031    0.101662    0.491268  
min      0.000000    0.000000    0.000000  
25%      0.000000    0.000000    0.000000  
50%      0.000000    0.000000    1.000000  
75%      0.000000    0.000000    1.000000  
max      1.000000    1.000000    1.000000  

[8 rows x 1025 columns]

None

There are total 229 'true' labeled molecules out from 384 in the dataset

Baseline prediction (all 'true', metric - accuracy) is 59.64%

Baseline prediction (all 'false', metric - accuracy) is 40.36%

train size = 304, batch size = 76, test size = 80
```

```
Train features info:

None

              0           1           2           3           4           5     \
count  304.000000  304.000000  304.000000  304.000000  304.000000  304.000000   
mean     0.032895    0.164474    0.042763    0.036184    0.059211    0.016447   
std      0.178655    0.371316    0.202656    0.187056    0.236407    0.127398   
min      0.000000    0.000000    0.000000    0.000000    0.000000    0.000000   
25%      0.000000    0.000000    0.000000    0.000000    0.000000    0.000000   
50%      0.000000    0.000000    0.000000    0.000000    0.000000    0.000000   
75%      0.000000    0.000000    0.000000    0.000000    0.000000    0.000000   
max      1.000000    1.000000    1.000000    1.000000    1.000000    1.000000   

              6           7           8           9     ...        1014  \
count  304.000000  304.000000  304.000000  304.000000  ...  304.000000   
mean     0.019737    0.029605    0.032895    0.013158  ...    0.009868   
std      0.139324    0.169775    0.178655    0.114139  ...    0.099012   
min      0.000000    0.000000    0.000000    0.000000  ...    0.000000   
25%      0.000000    0.000000    0.000000    0.000000  ...    0.000000   
50%      0.000000    0.000000    0.000000    0.000000  ...    0.000000   
75%      0.000000    0.000000    0.000000    0.000000  ...    0.000000   
max      1.000000    1.000000    1.000000    1.000000  ...    1.000000   

             1015   1016        1017        1018        1019        1020  \
count  304.000000  304.0  304.000000  304.000000  304.000000  304.000000   
mean     0.023026    0.0    0.085526    0.019737    0.154605    0.016447   
std      0.150234    0.0    0.280124    0.139324    0.362124    0.127398   
min      0.000000    0.0    0.000000    0.000000    0.000000    0.000000   
25%      0.000000    0.0    0.000000    0.000000    0.000000    0.000000   
50%      0.000000    0.0    0.000000    0.000000    0.000000    0.000000   
75%      0.000000    0.0    0.000000    0.000000    0.000000    0.000000   
max      1.000000    0.0    1.000000    1.000000    1.000000    1.000000   

             1021        1022        1023  
count  304.000000  304.000000  304.000000  
mean     0.013158    0.003289    0.009868  
std      0.114139    0.057354    0.099012  
min      0.000000    0.000000    0.000000  
25%      0.000000    0.000000    0.000000  
50%      0.000000    0.000000    0.000000  
75%      0.000000    0.000000    0.000000  
max      1.000000    1.000000    1.000000  

[8 rows x 1024 columns]

Test features info:

None

             0          1          2          3          4         5     \
count  80.000000  80.000000  80.000000  80.000000  80.000000  80.00000   
mean    0.037500   0.175000   0.012500   0.012500   0.062500   0.02500   
std     0.191182   0.382364   0.111803   0.111803   0.243589   0.15711   
min     0.000000   0.000000   0.000000   0.000000   0.000000   0.00000   
25%     0.000000   0.000000   0.000000   0.000000   0.000000   0.00000   
50%     0.000000   0.000000   0.000000   0.000000   0.000000   0.00000   
75%     0.000000   0.000000   0.000000   0.000000   0.000000   0.00000   
max     1.000000   1.000000   1.000000   1.000000   1.000000   1.00000   

             6         7          8          9     ...       1014  \
count  80.000000  80.00000  80.000000  80.000000  ...  80.000000   
mean    0.037500   0.02500   0.037500   0.062500  ...   0.012500   
std     0.191182   0.15711   0.191182   0.243589  ...   0.111803   
min     0.000000   0.00000   0.000000   0.000000  ...   0.000000   
25%     0.000000   0.00000   0.000000   0.000000  ...   0.000000   
50%     0.000000   0.00000   0.000000   0.000000  ...   0.000000   
75%     0.000000   0.00000   0.000000   0.000000  ...   0.000000   
max     1.000000   1.00000   1.000000   1.000000  ...   1.000000   

            1015  1016       1017       1018       1019       1020  1021  \
count  80.000000  80.0  80.000000  80.000000  80.000000  80.000000  80.0   
mean    0.012500   0.0   0.037500   0.012500   0.112500   0.012500   0.0   
```

```
std     0.111803   0.0   0.191182   0.111803   0.317974   0.111803   0.0
min     0.000000   0.0   0.000000   0.000000   0.000000   0.000000   0.0
25%     0.000000   0.0   0.000000   0.000000   0.000000   0.000000   0.0
50%     0.000000   0.0   0.000000   0.000000   0.000000   0.000000   0.0
75%     0.000000   0.0   0.000000   0.000000   0.000000   0.000000   0.0
max     1.000000   0.0   1.000000   1.000000   1.000000   1.000000   0.0

       1022       1023
count  80.0  80.000000
mean    0.0   0.012500
std     0.0   0.111803
min     0.0   0.000000
25%     0.0   0.000000
50%     0.0   0.000000
75%     0.0   0.000000
max     0.0   1.000000

[8 rows x 1024 columns]

Train class info:

count    304.000000
mean       0.595395
std        0.491625
min        0.000000
25%        0.000000
50%        1.000000
75%        1.000000
max        1.000000
Name: Soluble, dtype: float64

True train target fraction for batch 0 is 60.53%

True train target fraction for batch 1 is 57.89%

True train target fraction for batch 2 is 60.53%

True train target fraction for batch 3 is 59.21%

Test class info:

count    80.000000
mean      0.600000
std       0.492989
min       0.000000
25%       0.000000
50%       1.000000
75%       1.000000
max       1.000000
Name: Soluble, dtype: float64

Train class weights:
{0: 1.2357723577235773, 1: 0.83977900552486184}

Test class weights:
{0: 1.25, 1: 0.83333333333333337}
```