阿布都赛米·阿布都外力　　学号：2020182631　　考试号：180150129

实验名称：基于机器学习的判别模型构建

实验目的：

1. 了解数据集中正负样本处理方法。

2. 掌握模型构建与结果分析。

3. 掌握未知化合物活性预测。

实验原理：

使用 Discovery Studio 软件进行，以朴素贝叶斯为例对 FXR 活性剂与非活性剂进行机器学习判别模型构建。

本实验所用软件环境：

DS Version：19.1.0.18287

PP Version：19.1.0.1963

DS Client Version：19.1.0.18287

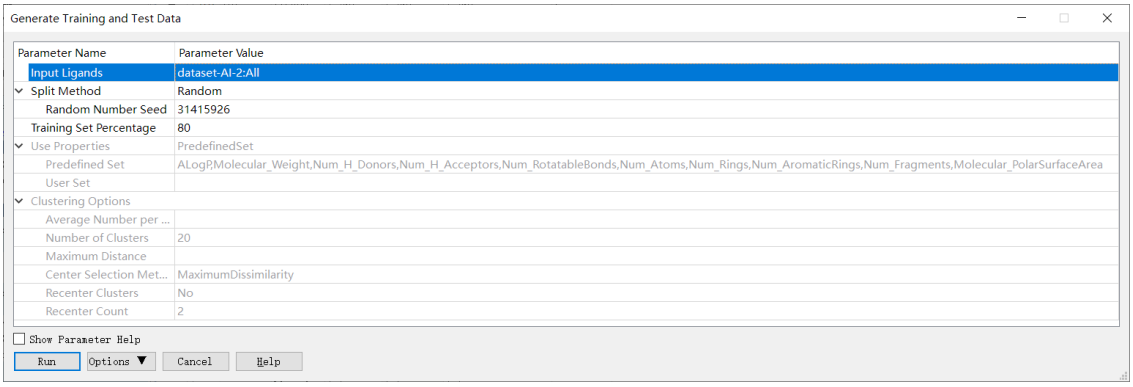
OS Distribution：Windows

OS Version：10.0.19044

实验步骤：

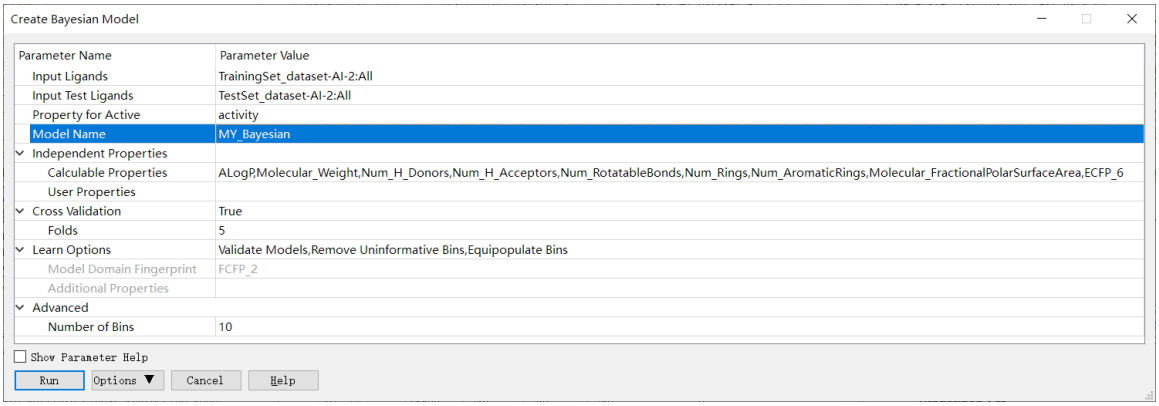
1. 已知活性数据收集：本实验使用指导老师提供的 dataset-AI-2.sdf 数据集。

2. 数据集预处理（正样本/负样本、训练集/测试集准备等）：本实验中，指导老师已经做好了正样本和负样本的分类。训练集/测试集的准备：点击 Discovery Studio 软件上的 Small Molecules→Create QSAR Model→Generate Training and Test Data 进行训练集与测试集拆分。设置参数如下：

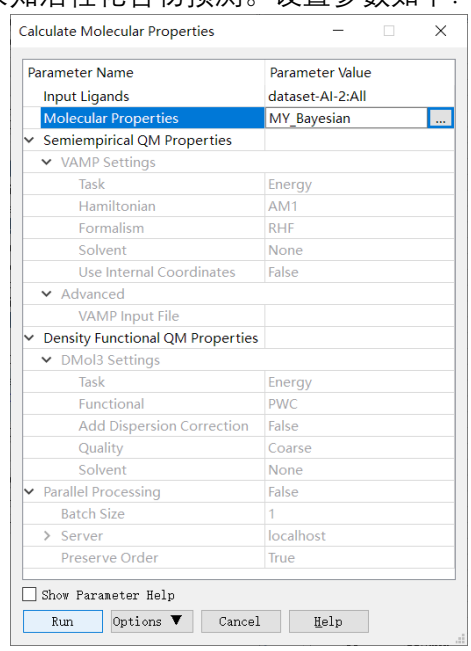


3. 分子描述属性计算（传统分子描述符、分子指纹等）：Discovery Studio 会在模型的构建中自动计算。在构建模型时，只需在 Calculable Properties 中挑选要计算的描述符。

4. 模型的构建与内外部验证：点击 Discovery Studio 软件上的 Small Molecules→ Create QSAR Model → Create Bayesian Model 进行朴素贝叶斯模型的构建。设置参数如下：



6. 未知活性化合物预测：未知活性化合物数据集用的是已知活性数据收集，点击 Discovery Studio 软件上的 Small Molecules→ Calculate Molecular Properties → Calculate Molecular Properties 进行未知活性化合物预测。设置参数如下：

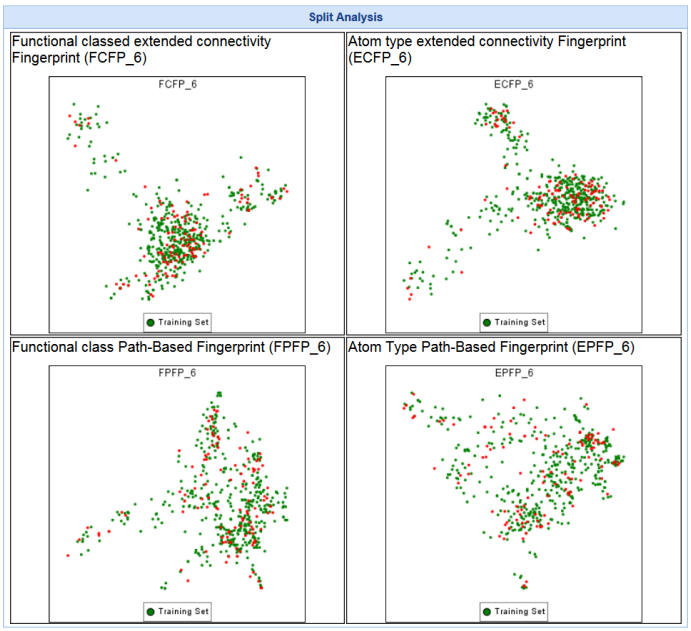


实验结果：

1. 数据集准备的结果：

Status：Success　　Elapsed Time: 00:00:22

Summary：Data split: 457 in training set, 114 in test set.



2. 模型的构建与内外部验证的结果:

Status: Success Elapsed Time: 00:00:13

Summary:

ROC score is 0.992 (leave-one-out).

Best cutoff for this model is -6.017.

See the Model Description results for more detailed information about this model.

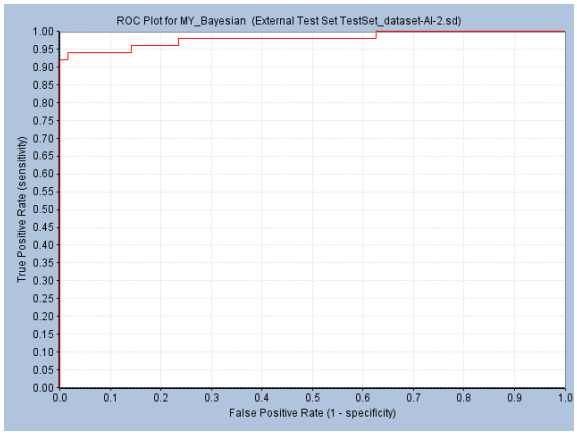
5-Fold Cross-Validation Result									
Model Name	ROC Score	ROC Rating	True Positive	False Negative	False Positive	True Negative	Sensitivity	Specificity	Concordance
MY_Bayesian	0.992	Excellent	219	2	0	236	0.991	1.000	0.996

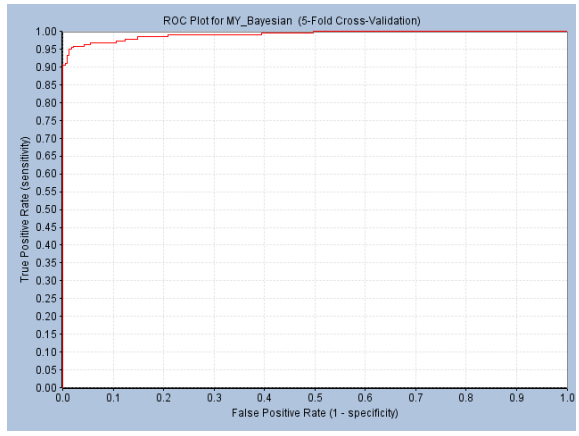
Test set validation: ROC score = 0.9796875.

Model Rating: Quality 0.980: Excellent

Confusion Matrix: True Positives = 46, False Negatives = 4, False Positives = 0, True Negatives = 64

Validation Result Using External Test Set TestSet_dataset-AI-2.sd									
Model Name	ROC Score	ROC Rating	True Positive	False Negative	False Positive	True Negative	Sensitivity	Specificity	Concordance
MY_Bayesian	0.980	Excellent	46	4	0	64	0.920	1.000	0.965





3. 未知活性化合物预测的结果：

Status： Success　　Elapsed Time: 00:00:02

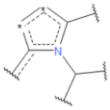
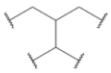
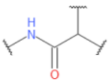
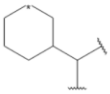
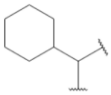
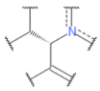
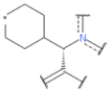
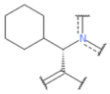
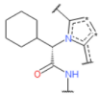
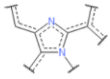
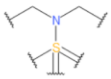
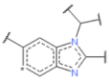
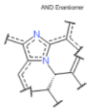
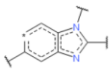
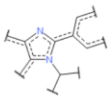
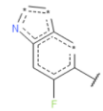
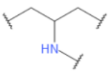
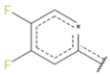
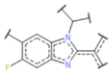
Summary: The following property has been added: MY_Bayesian

Index	Name	activity	MY_Bayesian	MY_Bayesian#Enrichment	MY_Bayesian#EstPGood	MY_Bayesian#Prediction
1	ligand_1	1	-13.2471	0.084979	0.0410949	false
2	ligand_2	1	-19.8647	0.0133365	0.00644937	false
3	ligand_3	1	-0.111291	1.65126	0.798531	true
4	ligand_4	1	8.54223	2.0534	0.993003	true
5	ligand_5	1	8.54223	2.0534	0.993003	true
6	ligand_6	1	9.4822	2.0583	0.995369	true
7	ligand_7	1	2.60008	1.90305	0.920291	true
8	ligand_8	1	36.2237	2.06787	1	true
9	ligand_9	1	4.93617	2.00138	0.967844	true
10	ligand_10	1	-11.9993	0.122351	0.0591677	false
11	ligand_11	1	5.25086	2.01036	0.972185	true
12	ligand_12	1	-4.89375	0.831676	0.402189	true
13	ligand_13	1	11.3311	2.0637	0.997983	true
14	ligand_14	1	38.7683	2.06787	1	true
15	ligand_15	1	-0.418819	1.61049	0.778816	true
16	ligand_16	1	31.4315	2.06787	1	true
17	ligand_17	1	32.7393	2.06787	1	true
18	ligand_18	1	21.563	2.06785	0.999987	true
19	ligand_19	1	38.3238	2.06787	1	true
20	ligand_20	1	37.6277	2.06787	1	true
21	ligand_21	1	37.2628	2.06787	1	true
22	ligand_22	1	37.4212	2.06787	1	true
23	ligand_23	1	28.9501	2.06787	1	true
24	ligand_24	1	36.1017	2.06787	1	true
25	ligand_25	1	19.2249	2.06778	0.999955	true
26	ligand_26	1	30.5767	2.06787	1	true
27	ligand_27	1	29.4841	2.06787	1	true
28	ligand_28	1	33.751	2.06787	1	true
29	ligand_29	1	34.5626	2.06787	1	true
30	ligand_30	1	33.2007	2.06787	1	true
31	ligand_31	1	35.2435	2.06787	1	true
32	ligand_32	1	32.2539	2.06787	1	true

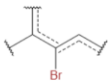
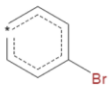
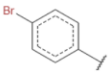
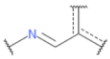
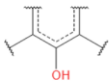
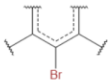
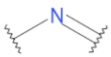
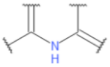
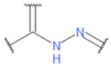
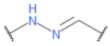
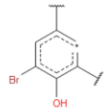
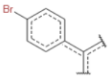
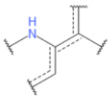
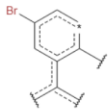
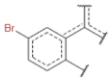
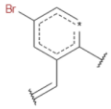
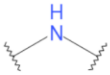
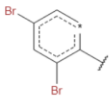
讨论：

可以从在内外部验证中所得出的模型精度指标看出，模型的敏感型很不错大于，0.9，特异性优异，等于 1，全局准确率也非常好，大于 0.9，ROC 分数也优良，表明所构建的模型可靠。（判断依据为全局准确率>0.8）。

Category MY_Bayesian: good features from ECFP_6

 G1: 79590563 69 out of 69 good Bayesian Score: 0.559	 G2: -858846751 67 out of 67 good Bayesian Score: 0.558	 G3: -649348348 66 out of 66 good Bayesian Score: 0.558	 G4: 1053438941 64 out of 64 good Bayesian Score: 0.558	 G5: 1986284472 64 out of 64 good Bayesian Score: 0.558
AND Enantiomer  G6: -1036816793 62 out of 62 good Bayesian Score: 0.558	AND Enantiomer  G7: -363996352 61 out of 61 good Bayesian Score: 0.557	AND Enantiomer  G8: 1281594252 61 out of 61 good Bayesian Score: 0.557	AND Enantiomer  G9: 1596422083 58 out of 58 good Bayesian Score: 0.557	AND Enantiomer  G10: -1470483086 58 out of 58 good Bayesian Score: 0.557
 G11: 295711361 57 out of 57 good Bayesian Score: 0.557	 G12: -868002145 55 out of 55 good Bayesian Score: 0.556	 G13: -2064047929 51 out of 51 good Bayesian Score: 0.555	AND Enantiomer  G14: 439894043 50 out of 50 good Bayesian Score: 0.555	 G15: 577496320 48 out of 48 good Bayesian Score: 0.554
 G16: -442520898 48 out of 48 good Bayesian Score: 0.554	 G17: 140099956 46 out of 46 good Bayesian Score: 0.553	 G18: -859078569 45 out of 45 good Bayesian Score: 0.553	 G19: 174740992 45 out of 45 good Bayesian Score: 0.553	 G20: 907447630 44 out of 44 good Bayesian Score: 0.553

Category MY_Bayesian: bad features from ECFP_6

 B1: 1334250623 1 out of 72 good Bayesian Score: -3.038	 B2: -1071952480 0 out of 34 good Bayesian Score: -3.007	 B3: -787327968 0 out of 34 good Bayesian Score: -3.007	 B4: -1832102709 0 out of 27 good Bayesian Score: -2.790	 B5: -1660913849 0 out of 26 good Bayesian Score: -2.754
 B6: -1661063237 0 out of 24 good Bayesian Score: -2.679	 B7: -1087070950 1 out of 46 good Bayesian Score: -2.604	 B8: -1699286547 0 out of 22 good Bayesian Score: -2.599	 B9: 1961034078 0 out of 18 good Bayesian Score: -2.414	 B10: 544048674 0 out of 17 good Bayesian Score: -2.362
 B11: 1814278164 0 out of 17 good Bayesian Score: -2.362	 B12: 1303458790 0 out of 16 good Bayesian Score: -2.308	 B13: -1665900899 0 out of 16 good Bayesian Score: -2.308	 B14: 1335833675 0 out of 15 good Bayesian Score: -2.250	 B15: -2067793897 0 out of 15 good Bayesian Score: -2.250
 B16: -742919872 0 out of 14 good Bayesian Score: -2.188	 B17: -2068518944 0 out of 13 good Bayesian Score: -2.123	 B18: -302078100 17 out of 253 good Bayesian Score: -2.080	 B19: -154530762 1 out of 26 good Bayesian Score: -2.061	 B20: -2063261623 0 out of 12 good Bayesian Score: -2.053