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. 模型构建与结果分析。

3. 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

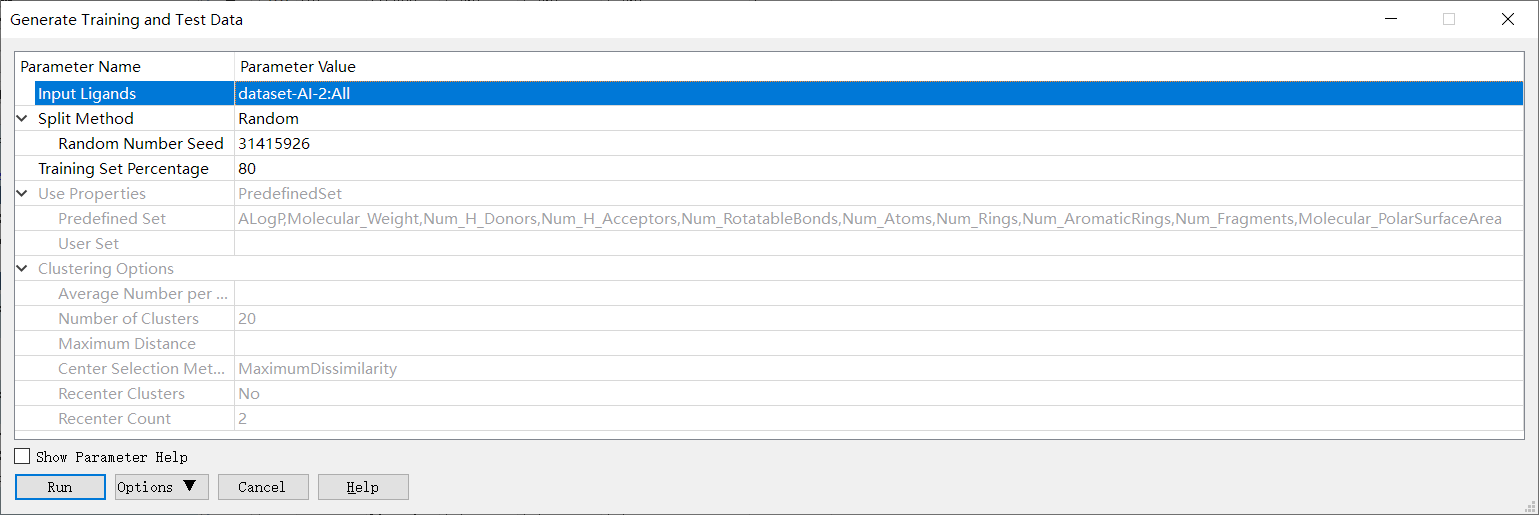
DS Client Version：19.1.0.18287

OS Distribution：Windows

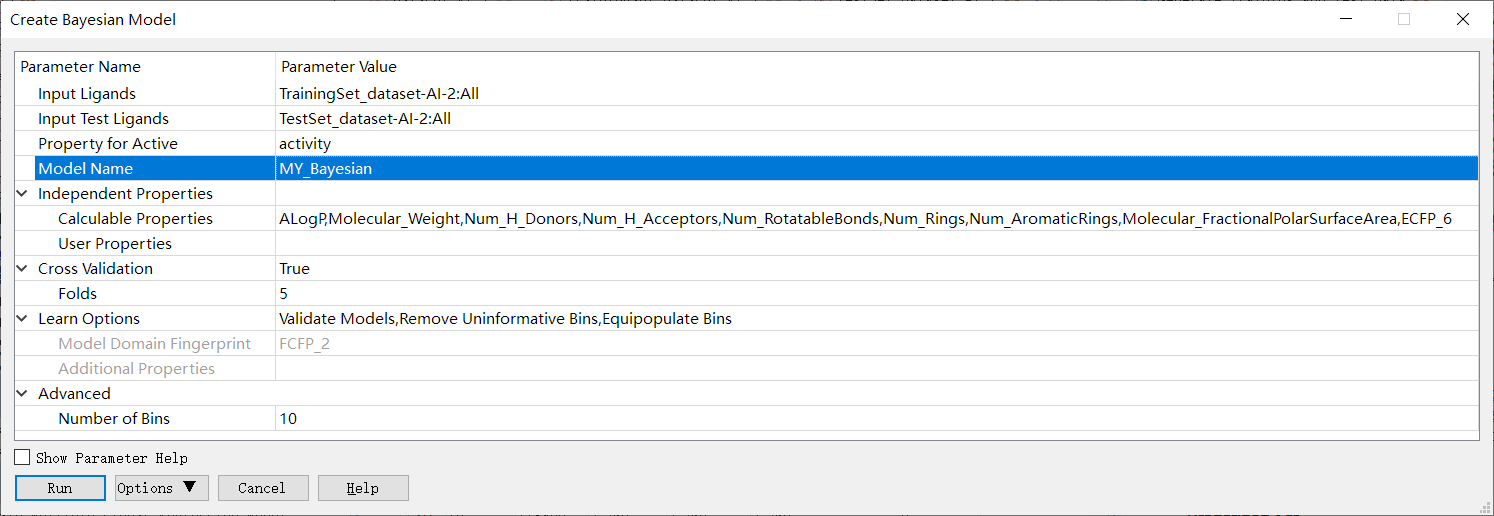
OS Version：10.0.19044

**实验步骤：**

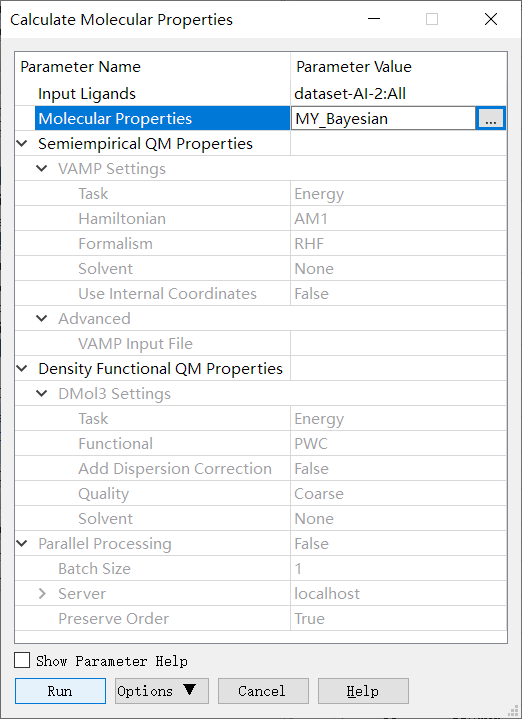
1. 已知活性数据收集：本实验使用指导老师提供的dataset-AI-2.sdf数据集。

 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）：本实验使用指导老师提供的dataset-qsar.sdf数据集中的pki-trypsin描述符。

 4. 模型的参数设置与构建：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：

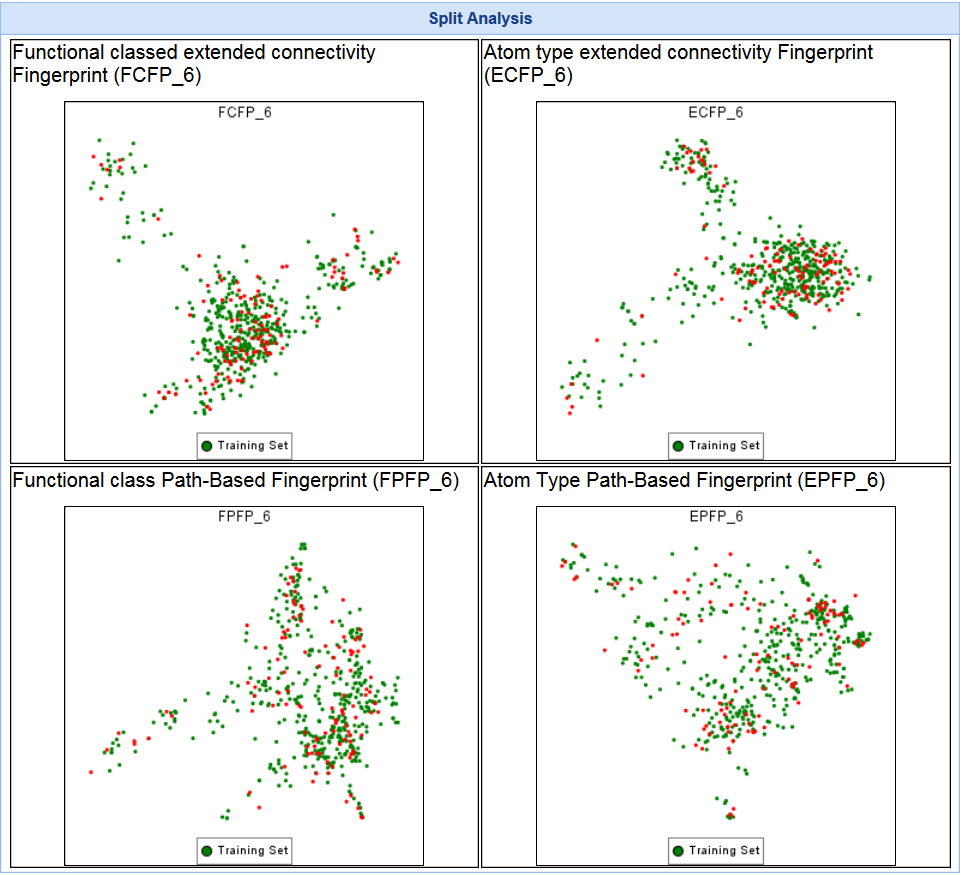
5. 外部数据集检测：（本实验没做）

 6. 未知活性化合物预测：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. 模型的参数设置与构建的结果：

Status：Success Elapsed Time: 00:00:13

Summary:

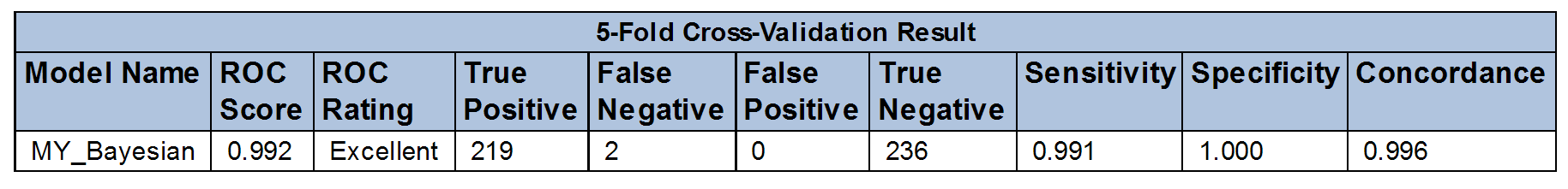
ROC score is 0.992 (leave-one-out).

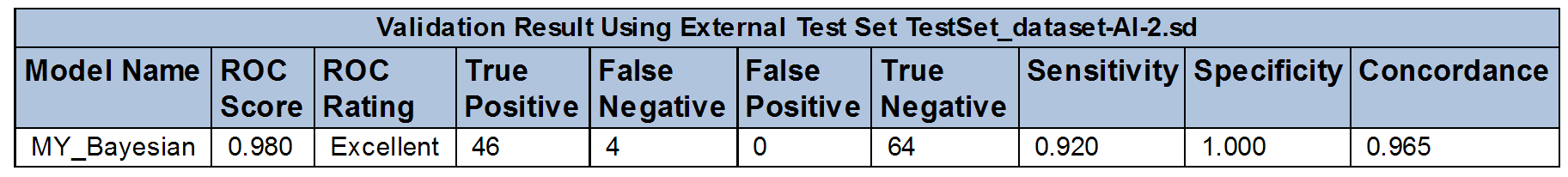
Best cutoff for this model is -6.017.

See the ModelDescription results for more detailed information about this model.

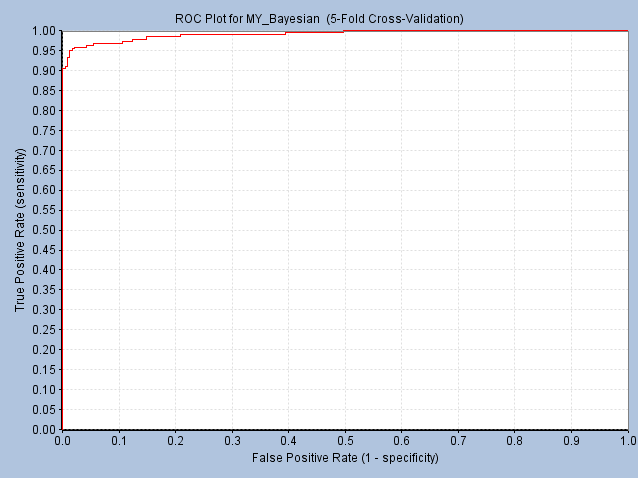
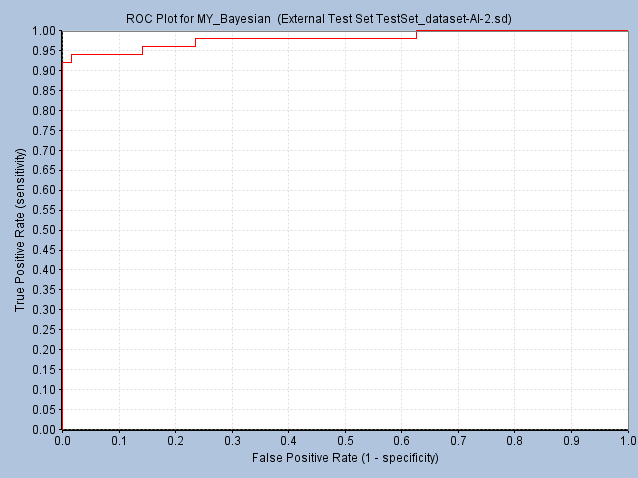
Test set validation: ROC score = 0.9796875.

Model Rating: Quality 0.980: Excellent

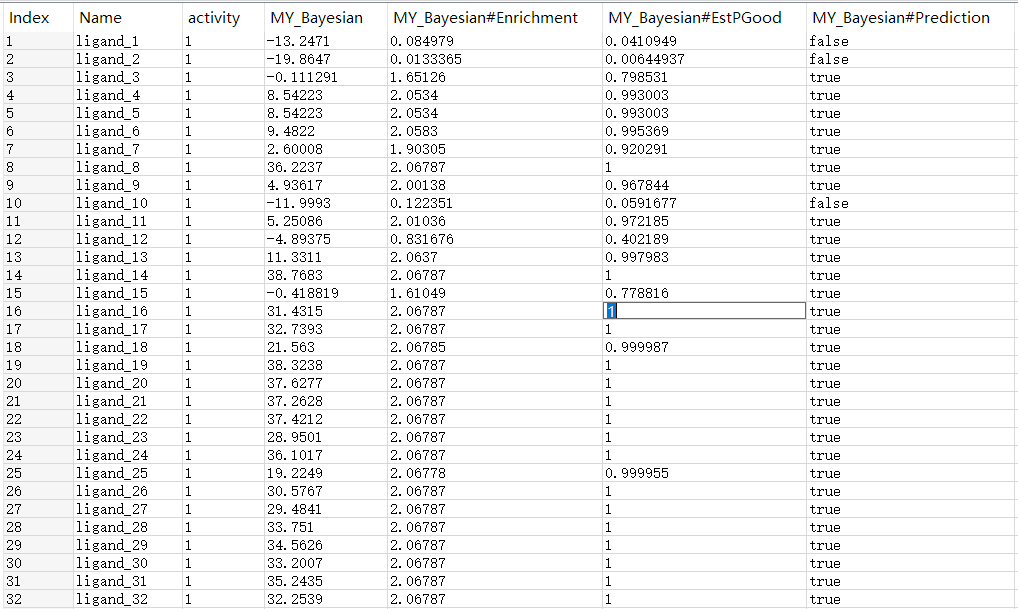
 Confusion Matrix: True Positives = 46, False Negatives = 4, False Positives = 0, True Negatives = 64



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

**讨论：**

模型精度的影响因素、化合物结构分析等.

