**抗乳腺癌候选药物治疗效果预测**

本课题研究不同药物对乳腺癌的治疗效果，即由不同特征组合而成的化合物对乳腺癌治疗靶标ERα的生物活性。

1. 课题采集了1974种化合物对ERα的生物活性数据，存放在“ERα\_activity.xlsx”文件的training表（训练集）中。该表的第一列提供了1974种化合物的结构式，第二列为化合物对ERα的生物活性值pIC50，值越大表明生物活性越高。该文件另有一个test表（测试集），里面列出了50个化合物，用于验证。
2. 上述1974个化合物的729个分子描述符信息（即自变量）存放于“Molecular\_Descriptor.xlsx”文件的training表（训练集）中（编号顺序与前面一样），每列代表化合物的一个分子描述符。该文件同样也有一个test表给出了上述50个测试集化合物的729个分子描述符。
3. 上述1974个化合物的ADMET性质（Absorption吸收、Distribution分布、Metabolism代谢、Excretion排泄、Toxicity毒性）数据存放于“ADMET.xlsx”文件的training表（训练集）中。该表第一列为化合物的结构式（编号顺序与前面一样），其后5列分别对应每个化合物的ADMET性质，均为布尔值。同样地，该文件也有一个test表，里面提供有上述50个化合物的结构式（编号顺序同上）。

**问题1.** 根据文件“Molecular\_Descriptor.xlsx”和“ERα\_activity.xlsx”提供的数据，针对1974个化合物的729个分子描述符进行变量选择，根据变量对生物活性影响的重要性进行排序，并给出前30个对生物活性最具有显著影响的分子描述符（即自变量），并请详细说明分子描述符筛选过程及其合理性。

**问题2.** 请结合问题1，选择不超过30个分子描述符变量，构建化合物对ERα生物活性的定量预测模型，请叙述建模过程。然后使用构建的预测模型，对文件“ERα\_activity.xlsx”的test表中的50个化合物的pIC50值进行预测，并给出预测精度。

**问题3.** 请利用文件“Molecular\_Descriptor.xlsx”提供的729个分子描述符，针对文件“ADMET.xlsx”中提供的1974个化合物的ADMET数据，分别构建化合物的Caco-2、CYP3A4、hERG、HOB、MN的分类预测模型，并简要叙述建模过程。然后使用所构建的5个分类预测模型，对文件“ADMET.xlsx”的test表中的50个化合物进行相应的预测，并给出预测结果的混淆矩阵。