《统计分析与建模》大作业

**合计30分，组队人数：1-2人**

**背景**

乳腺癌是目前世界上最常见，致死率较高的癌症之一。目前，抗激素治疗常用于ERα表达的乳腺癌患者，其通过调节雌激素受体活性来控制体内雌激素水平。因此，ERα被认为是治疗乳腺癌的重要靶标，能够拮抗ERα活性的化合物可能是治疗乳腺癌的候选药物。

**数据**

为了寻找抗乳腺癌候选药物，对若干化合物进行了实验，化合物的特征和实验数据见“附件一”。其中SMILES列表示化合物的分子式；之后的729列为化合物的分子描述符特征；**pIC50列是实验测量值，该值通常与生物活性具有正相关性，视为因变量**；Caco-2、CYP3A4、hERG、HOB、MN列为实验测得的化合物性质，性质的具体说明如下。化合物的分子描述符的详情可见“附件二”。

化合物性质说明：每个化合物的性质，采用二分类法提供相应的取值。Caco-2：‘1’代表该化合物的小肠上皮细胞渗透性较好，‘0’代表该化合物的小肠上皮细胞渗透性较差；CYP3A4：‘1’代表该化合物能够被CYP3A4代谢，‘0’代表该化合物不能被CYP3A4代谢；hERG：‘1’代表该化合物具有心脏毒性，‘0’代表该化合物不具有心脏毒性；HOB：‘1’代表该化合物的口服生物利用度较好，‘0’代表该化合物的口服生物利用度较差；MN：‘1’代表该化合物具有遗传毒性，‘0’代表该化合物不具有遗传毒性。

**问题**

请根据实验数据，完成以下问题分析与建模：

***问题1（10分）：***一般而言，无关特征的引入不仅会增加模型的复杂度，还可能由于引入干扰信息导致模型性能下降，因此在建模之前往往需要进行特征筛选。本题提供了729个特征，请你首先进行**数据预处理**（去除异常值、重复值等）**（4分）**，然后通过**相关性分析（2分）**，找出与pIC50相关性较强的特征（不超过20个）**（2分）**，以供后续建模使用，并将相关性前10的特征**填入下表（2分）**。

|  |  |
| --- | --- |
| 特征名 | 相关系数 |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

***问题2（10分）：***使用题（1）中筛选出的特征，按照4：1划分训练集和测试集**（2分）**，构建pIC50的**回归模型（2分）**，给出模型概览**截图（2分）。**在测试集上计算**拟合优度（2分）**，最后对回归系数进行**解读（2分）**；

***问题3（10分）：***使用题（1）中筛选出的特征，按照4：1划分训练集和测试集，从Caco-2、CYP3A4、hERG、HOB、MN五个性质中任选一个，构建对应性质的**分类模型**（即构建一个逻辑回归模型）**（2分）**，在测试数据集上绘制对应结果的**ROC曲线（2分）**，对模型进行**解读（4分）**及**质量评估（至少给出测试集上的F1，AUC结果）（2分）**；