**法医物证多人身份鉴定问题**

问题1 混合STR图谱分析的首要问题是判断贡献者人数。贡献者人数的正确与否决定着分析结果的准确率。依据附件1中混合STR图谱数据（如图1所示）设计算法或模型，用于识别某一混合样本中的贡献者人数，并评估其准确性。

**生物学背景知识：**

基因座是染色体上一个特定的物理位置，等位基因是同一基因座上可能出现的不同基因。个体从父母各继承一个等位基因，组合形成基因型（如“9/12”（杂合子））。因此一个人可以贡献2个等位基因（杂合子）或者1个（纯合子）。Allele检测出多少个等位基因对混合的人数判断具有重要作用（须结合Height排除干扰）。

**数据集摘要：**

行：816行，51份样本（2人混合到4人混合），每份均调查16个基因座。（相当于只有51个样本，只不过每个有16行的数据）

列：Allele N Size N Height N，N为[1,100]，31后无数据，此前，均存在缺失值。具体如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sample File | Marker（分类变量） | Dye | Allele N（分类变量） | Size N（数值型） | Height N（数值型） |
| 样本名，可提取人数和比例 | 基因座 | 染色（不管） | 数字/X/Y/OL代表为某等位基因（OL表示不知道什么基因） | 该等位基因的DNA片段长度。不同size对应不同的等位基因 | 峰高，反映该等位基因的DNA量，可用于判断样本是否为混合样本及混合比例。 |

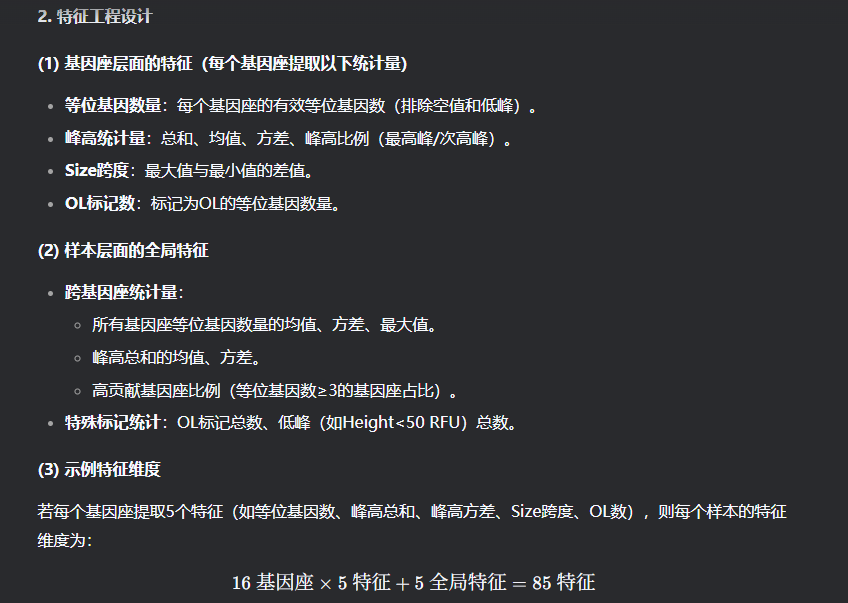
例如附件4第一行为44号贡献者和45号贡献者的1：1混合样本对D8S1179基因座进行测试的数据。测试出8个等位基因，其中3个的峰高>0，因此可以推测为一个杂合子和一个纯合子两人混合的样本（也可能多人，需要根据峰高进一步判断）。

注：附件4的数据排除了附件1的干扰，本题要求使用附件1，因此设计时须考虑排除干扰（如低height）。

**研究思路：**

提取特征构建随机森林模型，输出准确率等指标。

数据集：应该使用这816个行的数据对每行的结果进行预测（即使用单个基因座的测试结果进行预测），还是使用这51个样本，结合其16个基因座的检测结果进行预测？（后者吧）

特征工程：

问题2 在分析出贡献者人数后，还需要判断各贡献者的混合比例。当贡献者比例接近时，等位基因可能重叠，导致误判基因型。明确比例有助于更精准地分析混合图谱。依据附件2中混合ST图谱数据（如图2所示）设计算法或模型，用于识别某一混合样本中的贡献者比例，并评估其准确性。

问题3 根据附件1与附件2的混合STR图谱数据以及附件3中各个贡献者的基因型，设计算法或模型，用于推断某一混合STR图谱中各个贡献者对应的基因型，并评估其准确性。

问题4 依据附件4中混合STR图谱数据（如图3所示）设计算法或模型，用于减少混合样本中噪声的干扰，以提高混合样本分析的准确性。