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：

2. 特征工程设计

(1) 基因座层面的特征（每个基因座提取以下统计量）

- **等位基因数量**：每个基因座的有效等位基因数（排除空值和低峰）。
- **峰高统计量**：总和、均值、方差、峰高比例（最高峰/次高峰）。
- **Size跨度**：最大值与最小值的差值。
- **OL标记数**：标记为OL的等位基因数量。

(2) 样本层面的全局特征

- **跨基因座统计量**：
  - 所有基因座等位基因数量的均值、方差、最大值。
  - 峰高总和的均值、方差。
  - 高贡献基因座比例（等位基因数≥3的基因座占比）。
- **特殊标记统计**：OL标记总数、低峰（如Height<50 RFU）总数。

(3) 示例特征维度

若每个基因座提取5个特征（如等位基因数、峰高总和、峰高方差、Size跨度、OL数），则每个样本的特征维度为：

$$16 \text{ 基因座} \times 5 \text{ 特征} + 5 \text{ 全局特征} = 85 \text{ 特征}$$

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。