问题列表：

1. 为什么选用决策树模型

决策树算法是一种基本的分类方法，通过决策树算法训练生成的模型可读性强，在处理案件时办案人员可以很直观地看到模型的构成，熟悉其工作方式和判定为嫌疑人的依据。并且决策树的时间复杂度低，在进行预测时速度很快，使办案人员在办案过程中能够快速锁定嫌疑人范围，最大程度上加快办案速度，提高办案效率。另外，多个（几十甚至上百）单树模型可以根据实际需要进行组合，构成随机森林，从而适应更多更广的业务场景。

1. 如何进行采样的

关联数据库中嫌疑人信息表和案件信息表，得到2017年所有隐性涉毒人员相关数据，并将该数据与旅馆住宿信息、手机通讯录信息、涉警行为信息等数据关联，作为正样本；对未登记在嫌疑人信息表和违法信息犯罪表中的正常人员数据进行聚类，再从每个类别当中随机抽取部分数据组成负样本。最终采集正负样本各3600个左右，总计7000多个样本，样本比例1:1。

3）根据什么/为什么选择这些特征

分别对涉毒人员和正常人的基本属性特征、社会属性特征、行为轨迹特征做统计性分析，选择两者分析结果差异较大的特征作为主要特征，这些特征对两者的区分作用更显著。

4）如何构建决策树模型以及如何评估模型

我们使用的是C4.5算法，即 通过信息增益率来选择当前最合适的节点，我们从7000多个样本中抽取60%作为训练集训练模型，使用信息增益率来选择当前最合适的节点，使用剩余40%作为测试集评估模型，计算其精确率和召回率，以判断模型的可用性。