**【参考答案】**

做描述性统计分析，结果显示，无缺失值（2分）和异常值（2分），不需要处理

离散变量的描述性统计（2分），样本较均衡（275：108）（3分）（处理均衡性也得分）



连续变量的描述性统计（2分），数据年龄有一定程度的右偏（2分），暂不处理，可以先查看结果再考虑要不要处理（box-cox处理也得分）



标签编码：（2分）

将2分类文本型字段进行标签编码处理，包括性别、随访期间是否吸烟、确诊前是否吸烟、是否放射治疗、颈部淋巴是否肿大、癌症灶性和是否复发7个字段



独热编码：（2分）

将多分类文本型字段进行独热编码（或哑变量）处理，最好选择哑变量，或者是独热编码后去掉最后一个字段进入模型，包括除年龄和上述7个字段以外的其他6个字段



**（2）**

原假设：确诊之前是否吸烟和是否复发没有显著关联；

备择假设：确诊之前是否吸烟和是否复发有显著关联。

卡方检验分析结果：



P值为 0.014，小于0.05，拒绝原假设。即确诊之前是否吸烟和是否复发有显著关联。

（原假设备择假设合理4分，p值2分，结果分析合理并且结论正确2分）

原假设：患者年龄和是否复发没有显著关联；

备择假设：患者年龄和是否复发有显著关联。

单因素方差分析结果：



P值为2.78\*10-7，小于0.05，拒绝原假设。患者年龄和是否复发有显著关联。

（原假设备择假设合理4分，p值2分，结果分析合理并且结论正确2分）



**（3）**

以是否复发作为因变量，其他变量作为自变量，构造分类模型。

逻辑回归结果（参数默认）：（选择一种分类算法2分+模型结果展示2分+模型结果分析4分）

训练集：



测试集：



逻辑回归结果既没有欠拟合也没有过拟合

决策树算法：深度为6，其他默认（选择一种分类算法2分+模型结果展示2分+模型结果分析4分）

训练集：



测试集：



决策树结果既没有欠拟合也没有过拟合

综合来看，逻辑回归和决策树的结果都比较令人满意，两种方法在解释力上各有优势，逻辑回归通过回归系数进行解释，而决策树可以通过树的形状来解释分类规则。（比较2类不同模型的分析结论4分）



**（4）**

解释患者是否复发的影响因素：（3分，言之有理即可）

治疗反应部分成功的人易复发

年龄越大越容易复发

男性复发的概率越大

吸烟的患者复发概率更大

接受过放疗的患者复发概率更大



建议：（2分，言之有理即可）

对治疗反应部分成功、年龄大、男性、吸烟和接受过放疗等情况的患者给与更密切的复查和关注