大纲

1. data analysis文件：先加载数据，将所有文字属性替换成01二值，分析数据，例如中位数、平均值等等；检查缺失值；是否患病比例及数量；用雷达图展示是否患病的比例，据此可以推断哪些疾病重要哪些没那么重要；再用热力图分析，pearson相关性，判断不同属性及是否患病的内在联系，用0.4作为阈值屏蔽小于0.4的值，保证剩下的都是相关性较大的
2. 不同模型：首先使用baseline模型（logistics回归），用来作为其余模型的表现基准，得出一组数据；再用SVM模型，用GridSearchCV方法找出最佳组合，然后用10折交叉验证来保证模型稳定性和准确性，得出最优的数据；接着用decision tree方法，同样使用GridSearchCV和十折交叉验证；在decision tree基础上，使用ensemble method方法提升其性能，搜索发现最优情况在使用了160个基础估计器，因此得到了使用random forest的最优解；最后使用神经网络方法，该方法属于深度学习范畴，由于每个属性重要性相当（因为不是视觉领域），所以使用前向传播的全连接网络，因为是二值分类，最后使用sigmoid函数作为输出层，同样使用交叉验证保证稳定性，发现神经网络的性能是最好的。
3. 结合literature review部分，佐证各种属性的重要性和相关性的结论，该部分还没有写完，暂时先不写在这里。