基于校赛题目：

心血管疾病（CVD）是心脏和血管疾病的总称。常见的心血管疾病包括高血压（血压升高）、冠心病（心脏病发作）、脑血管疾病（中风）、外周血管疾病、心力衰竭、风湿性心脏病、先天性心脏病和心肌病。据世界卫生组织统计，2012 年约有 1750 万人死于心血管疾病，占全球死亡人数的 31%。由于心血管疾病患者人数不断增加，心血管疾病的诊断和治疗已成为医疗行业的主要问题。
本问题基于 Kaggle 公开的心血管疾病患者诊断数据。
(详见 https://www.kaggle.com/datasets/pirogovskiy/cardio-train/data)
问题

1. 首先，对数据进行预处理和探索性分析（见附件 "文件 cardio_train.csv "和 "data_dictionary.xlsx"）。
2. 根据患者提供的生理指标、医学检测指标和主观信息，使用机器学习中的分类方法（或变量方法）预测患者是否患有心血管疾病。
3. 最后，比较不同分类器下的预测性能。并得出结论。

宇哲嘟论文分析：

建模方面：

分类明确

问题一 分为 数据质量分析、数据源、数据预处理（数据清理、数据转换、数据简化）、探索性数据分析（相关系数的定义、相关分析结果）

问题二 通过六个模型建模，通过分类分成两块

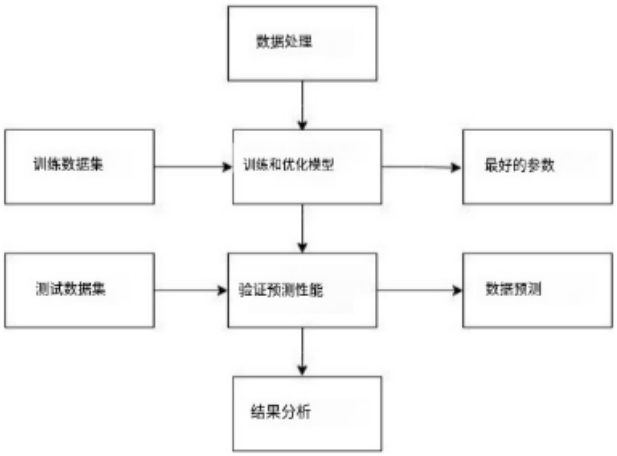
单一模型：Logistic回归模型、BP神经网络模型、支持向量机模型

集成模型：随机森林预测模型、XGBoost预测模型、LightGBM预测模型

问题三 结果比较分析

- 简单问题：分**定义**和**分析结果**
- **机器学习问题思路：**

  对于机器学习题目，可以借鉴这个思路，并改进、替换作图



- **BP神经网络**没有使用过，基本原理分为：输入层->隐藏层->输出层
- **支持向量机模型**：SVM，寻找最优决策平面

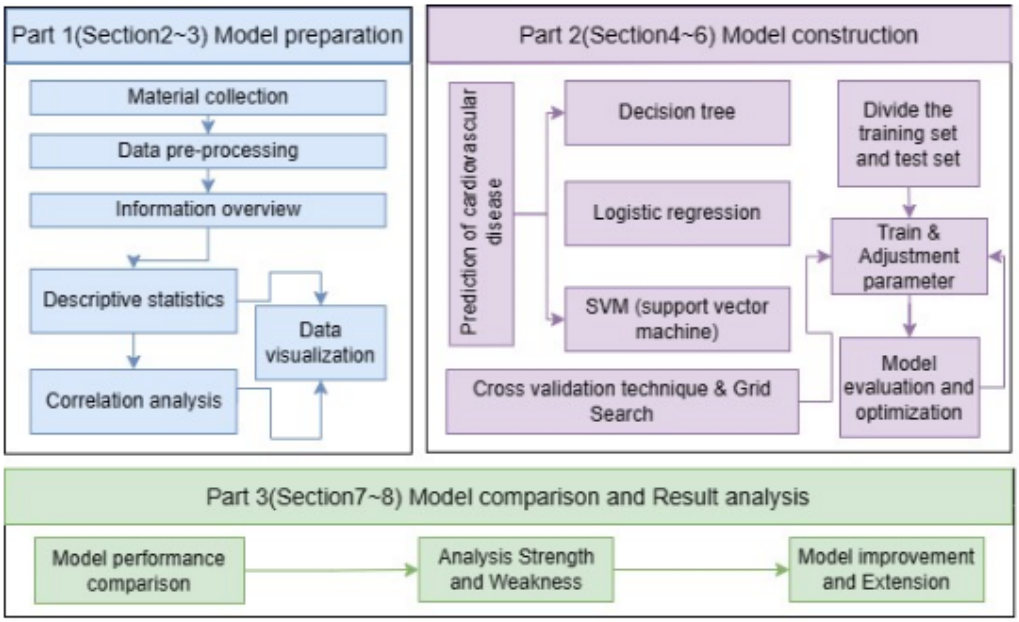
总结：

- 宇哲这篇文章有个特点，”简练“。
  在问题二对于问题的计算的时候只给出了关键的中间量和最后的结果。
  可以学习的是中间量的处理，但是他这篇文章还可以加强可视化、对比分析等。
- 六种机器学习的基本方法，在下一次使用机器学习拟合的时候可以借鉴它的文字、公式描述。
- 能得校一有两个原因：①文章结构清晰 ②方法用的够充足且有足够的中间量支持

蓉蓉姐嘟论文分析：

可以学习的点：

- 首先这个图画的好：



我们这部分不够细致

- 第二个优点：
  数据清理：
  数据清理给清理数量（增加可信度）还行，但主要学习的是后面的这个**箱线图**的绘制，根据删除数据的箱线图得到异常值数据的减少和保留的数据是否处于合理范围内。

Table 2: Number of Deletion

index	amount
ap_hi	255
ap_lo	1023
aphi < aplo	10
height	1
totol	1415

t of the deleted data shows that the outlier data is significantly
l outliers are within a reasonable range and evenly distributed

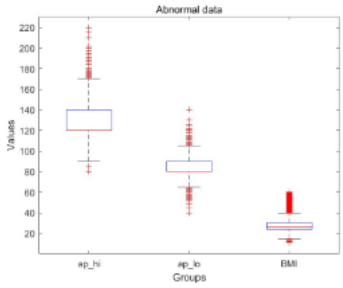


Figure 4: Box Plot of Processed Data

总结：

- 图画的好哎，加上思路清晰，很多准备和附件工作做的很足，建模思路很好
- 三个方法：决策树、逻辑回归、支持向量机

自己嘟论文分析：

姐姐写的第三个模型和前两个我写的模型风格差异太大了，加上乔乔处理论文时间不够，导致整体文章有点乱。
倒也是情有可原，姐姐没怎么参与我们的论文。导致两个人工作量过大，完成不了。再加上第一次写机器学习的文章，挺需要改进的。

相比于校赛，美赛会更加难，更加复杂，数据要自己找，思路也不会这么简单。
对时间效率和时间累计的要求就更高了

努力方向：
①更好地分析题目的隐含工作，寻找作图契机。
②校赛的主要区分点在于思路的清晰度，但是美赛可能还需要问题处理思路和文章美化，可以再多分析分析美赛优秀论文。
③比赛的时候认真一些叭，校赛没有拿出自己的全部实力喽，静下心来，把文字工作做明白，毕竟代码只是辅助，最重要的还是展示给评委的文字部分。