# 基于 UCI 心脏病数据的分析与建模

**摘要：**本文对 UCI 心脏病数据集进行探索性数据分析，以可视化、特征相关程度分析等方式对不同的特征属性进行分析，并尝试建立二分类模型。本文尝试了多种模型进行建模，并进行了大量可视化的实验。

## 一、 数据集与数据探索

### 1.1 数据集介绍

心脏病是一种成因复杂，表现形式多样，对于人类健康造成巨大影响的疾病。UCI 数据集是现在网络上唯一可被允许自由分析处理的数据集，数据的实例不大，仅有 300 多条，原数据集内有 76 个属性，最后发行的数据集只有 14 个。所有的数据已被脱敏，隐去了患者的个人信息。数据属性如下

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholestoral in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

Table 1: Data attributions

我们可以用 pandas profiling report 来预览一下数据集的概况

Dataset statistics

| | |
|---|---|
| **Number of variables** | 14 |
| **Number of observations** | 303 |
| **Missing cells** | 0 |
| **Missing cells (%)** | 0.0% |
| **Duplicate rows** | 1 |
| **Duplicate rows (%)** | 0.3% |
| **Total size in memory** | 33.3 KiB |
| **Average record size in memory** | 112.4 B |

| Variable types | |
|---|---|
| **NUM** | 6 |
| **BOOL** | 4 |
| **CAT** | 4 |

Table 2: Dataset statistics and Variable types

数据样本(sample)如下

**First rows**

| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 63 | 1 | 3 | 145 | 233 | 1 | 0 | 150 | 0 | 2.3 | 0 | 0 | 1 | 1 |
| **1** | 37 | 1 | 2 | 130 | 250 | 0 | 1 | 187 | 0 | 3.5 | 0 | 0 | 2 | 1 |
| **2** | 41 | 0 | 1 | 130 | 204 | 0 | 0 | 172 | 0 | 1.4 | 2 | 0 | 2 | 1 |
| **3** | 56 | 1 | 1 | 120 | 236 | 0 | 1 | 178 | 0 | 0.8 | 2 | 0 | 2 | 1 |
| **4** | 57 | 0 | 0 | 120 | 354 | 0 | 1 | 163 | 1 | 0.6 | 2 | 0 | 2 | 1 |
| **5** | 57 | 1 | 0 | 140 | 192 | 0 | 1 | 148 | 0 | 0.4 | 1 | 0 | 1 | 1 |
| **6** | 56 | 0 | 1 | 140 | 294 | 0 | 0 | 153 | 0 | 1.3 | 1 | 0 | 2 | 1 |
| **7** | 44 | 1 | 1 | 120 | 263 | 0 | 1 | 173 | 0 | 0.0 | 2 | 0 | 3 | 1 |
| **8** | 52 | 1 | 2 | 172 | 199 | 1 | 1 | 162 | 0 | 0.5 | 2 | 0 | 3 | 1 |
| **9** | 57 | 1 | 2 | 150 | 168 | 0 | 1 | 174 | 0 | 1.6 | 2 | 0 | 2 | 1 |

Table 3: First 10 rows of the dataset

当然，该报告还报告了数据的分布，交互(interaction)，相关性分析(correlations) 和缺失数据（missing values）。

## 1.2 数据探索 [1]

### 1.2.1 使用 seaborn 进行数据可视化

首先，我先完成了数据的特征组合，发现由于维数较大，pairplot 的结果其实意义不是很大，因为根本看不清楚两两数据的关联。于是，我直接生成了**数据的关联性矩阵，将其绘制成热点图**。

[1] 注：后续的详细内容请见代码提交，文章内仅有 demo 演示和重要的图表展示。

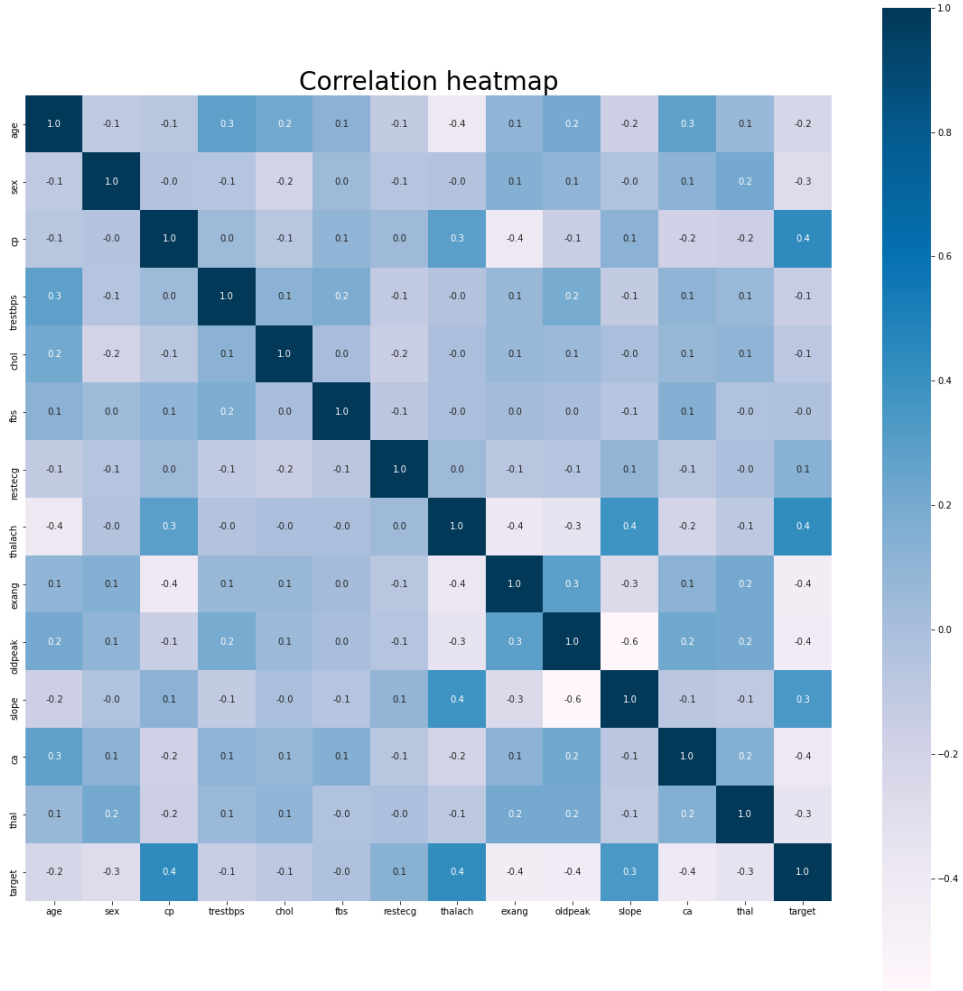


FIG1 Correlation heatmap

其次，我对**单变量的数据**画出了其频率分布直方图和函数拟合。这部分的工作可以轻松扩展到其他不同属性的数据中。

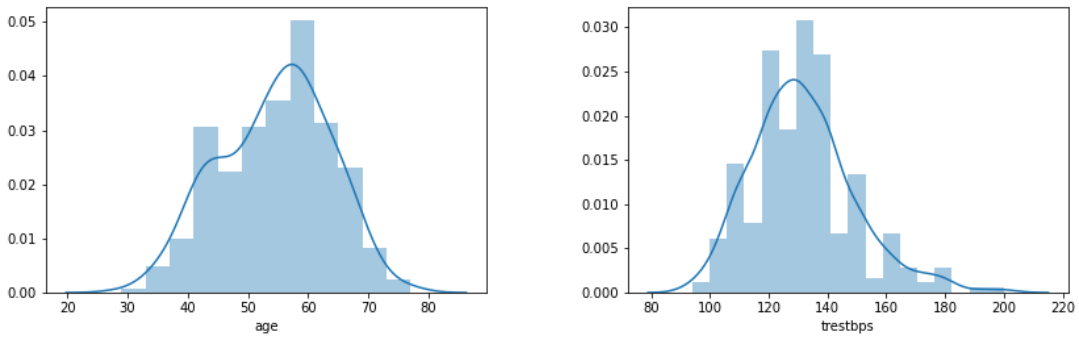


FIG2 age and max-heart-rate distributions

最后，我尝试了一些比较有意思的可视化分布展示，我们在学习地理的时候学过用小提琴图来展示人口的年龄在某一时间点的分布。这里，我们也可以用同样的思路展示**不同性别的病人的年龄分布（特征与标签关系）**。

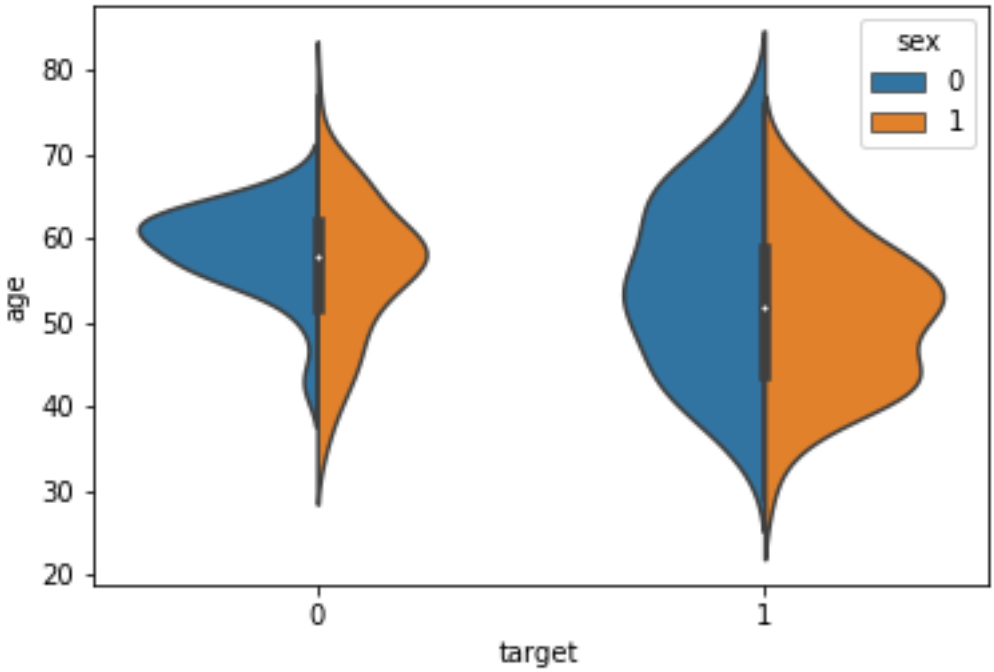


FIG3 age distributions based on target and sex

### 1.2.2 使用 PDPbox 进行进阶的数据可视化

PDPbox 是 Python 部分依赖图工具箱，这是用来可视化某些特征对于任意监督学习的算法所给出的模型的影响。这对于了解不同数据特征之间的行为很有帮助，为之后的预测算法提供了一些直观的印象。在之后的模型部分会用比较详细的展示与说明。

首先是**单一因子**的分析（仅举一例）：

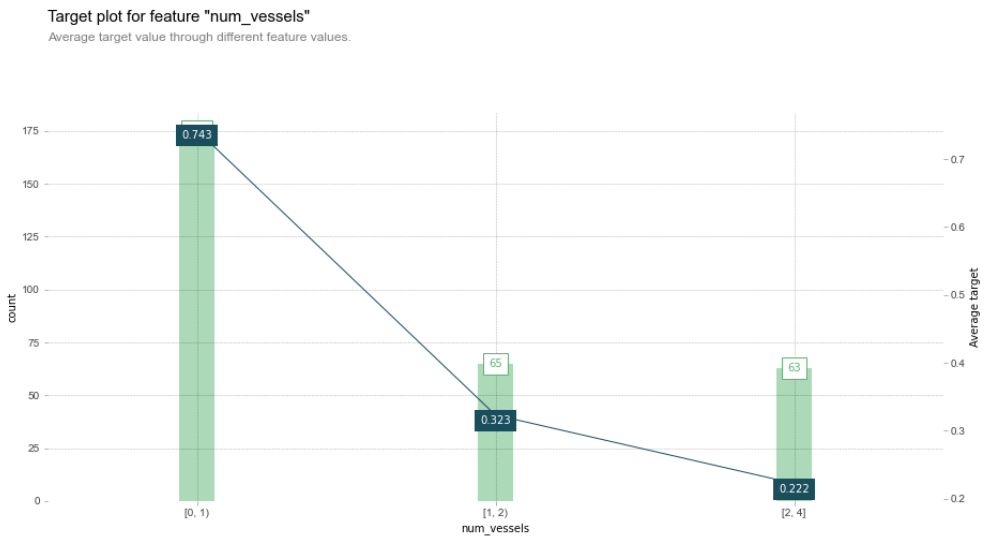


FIG4 impact of num_vessels towards the likelihoods of contracting heart disease

然后是**双因子**的分析（仅举一例）：

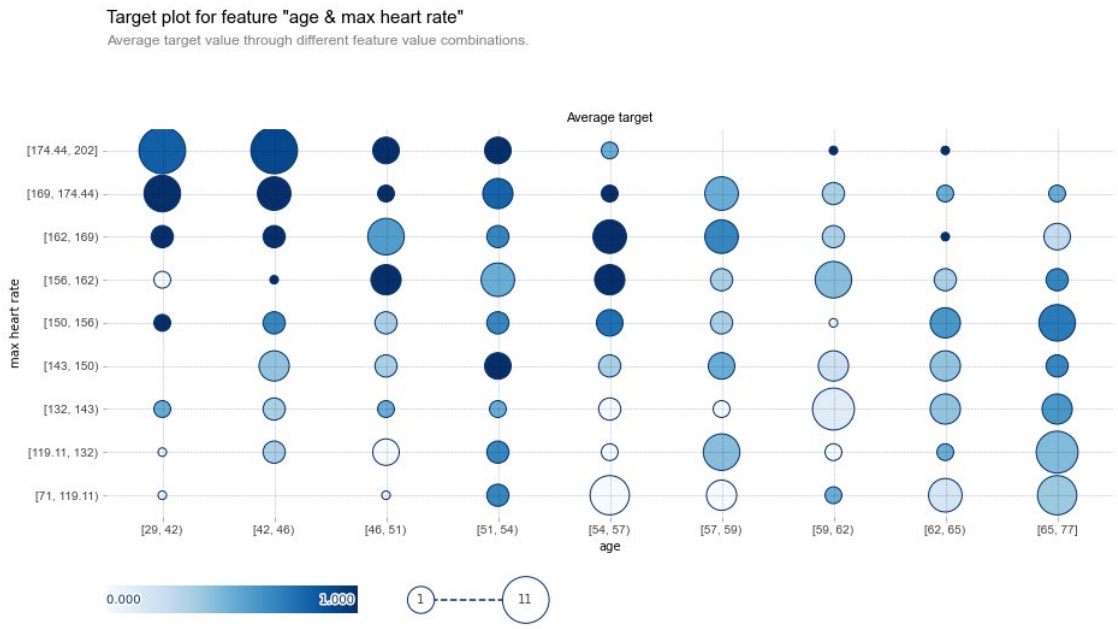


FIG5 factors: max_heart_rate and age towards the likelihoods of contracting heart disease

## 二、 数据建模

### 2.1 分类 (classifications)

分类是机器学习中重要的问题。在这里，我们先将问题定义为**二分类问题**，即患者是否患有心脏病。实验的操作顺序与导论课的实验类似，但新增了几个模型用来比较性能差异，分别是：1. Logistic 回归，2. K-NN，3. SVM，4. Naive Byte，5. Decision Tree，6. Random Forest。我们的评价主要是由**准确率**决定的，但为了完整性与科学性，我同时会给出每一种模型所生成的混淆矩阵。

### 2.2 配置与模型解释 (configuration and model interpretation)

#### 2.2.1 测试集与训练集

同导论课程：80%训练，20%测试。

#### 2.2.2 二分类 logistic 回归模型

回归模型常用的有线性回归，logistic 回归，多项式回归等。这里我们采用了 logistic 回归，其原理如下：

a. Logistic 回归与线性回归的本质就是增加局部区间的敏感程度，使得算法比较健壮。因此我们首先定义了一个 sigmoid function：

$$sigmoid(x) = \frac{1}{1+e^{-x}}$$

这个函数的性质很好，适合二分类。

b. 我们还需要定义 cost function 来使得我们的模型输出值贴合真实值。但使用常用的损失函数如二次代价函数(quadratic cost)，我们无法得到一个非凸函数，因此无法使用梯度下降法。因此我们采用对数似然代价函数(log-likelihood cost)

c. 最后，我们采用梯度下降法求解此问题。在这里，可以利用反向传播(backward propagation)算法计算 loss 对各层参数的梯度。

### 2.2.3 其他模型

其余算法都是分类问题中非常常见的算法。这里直接采用 Sklearn 的机器学期包进行使用。这里的多数模型上课也有提及，在实验课中也已经有呈现。

## 2.3 Models comparison

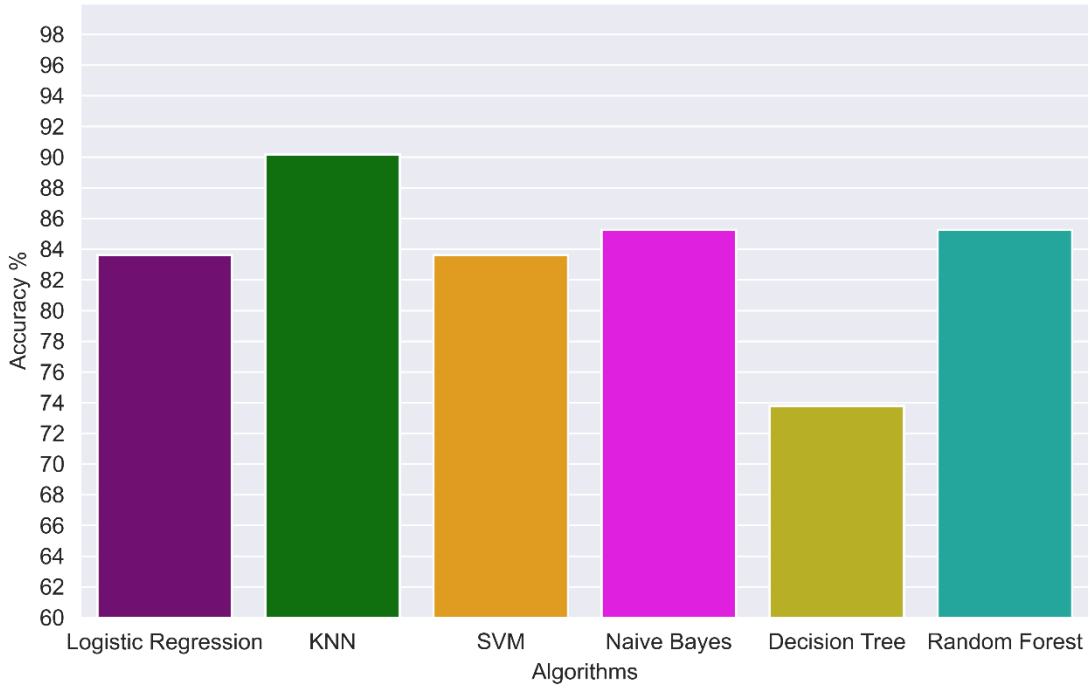


FIG6 models comparison based on 6 common classification algorithms

KNN 算法是解决二分类问题的出色选手。其以超过 90%的准确识别度位列六种算法的榜首，其次是回归模型，随机森林，支持向量机和朴素贝叶斯四种算法，准确度在85%附近，而 Decision Tree 表现比较差，仅有约 74%。

**参数设定：**

1. K-NN 的 k 值为 7。
2. Logistic regression 使用的是 sklearn 的预设，在手工搭建的 regression 中，经过 50 次迭代后的结果为 85.25%。
3. 其余模型均采用 sklearn 的预设。

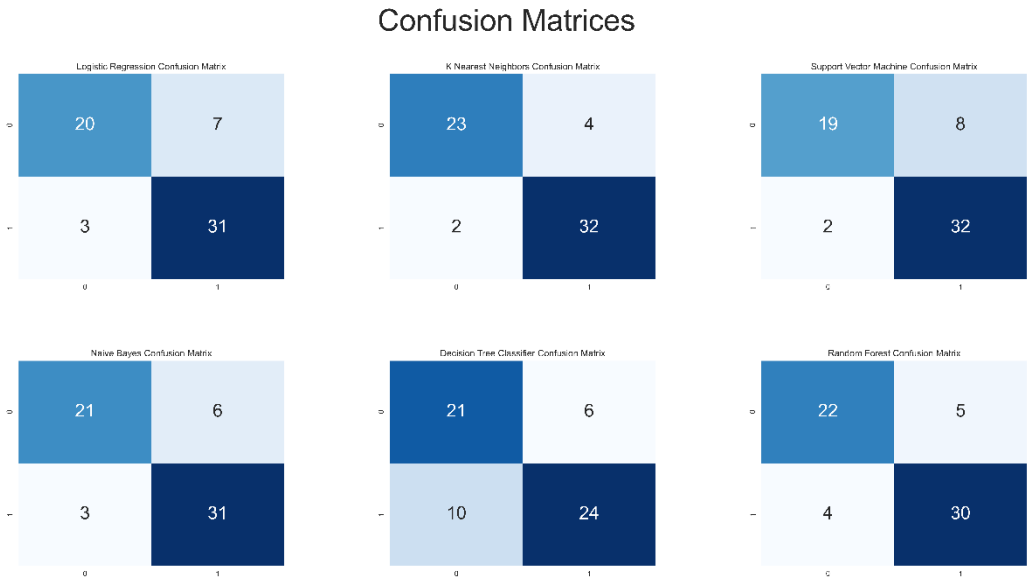


FIG7 confusion matrices

根据混淆矩阵，我们也可以计算出模型的 sensitivity, specificity, precision, false discovery rate 等相关参数。具体公式与定义如下：

| | | |
|---|---|---|
| True positive(**TP**) | upper left | eqv. with hit |
| True negative(**TN**) | upper right | eqv. with correct rejection |
| False positive(**FP**) | down left | eqv. with false alarm, Type I error |
| False negative(**FN**) | down right | eqv. with miss, Type II error |

Table 4: connotation of confusion matrix

| | |
|---|---|
| Sensitivity or true positive rate(**TPR**)<br>eqv. With hit rate, recall | $TPR = TP/P = TP/(TP + FN)$ |
| Specificity(**SPC**)or true negative rate(**TNR**) | $SPC = TN/N = TN/(FP + TN)$ |
| Precision or positive prediction value(**PPV**) | $PPV = TP/(TP + FP)$ |
| Negative predictive value(**NPV**) | $NPV = TN/(TN + FN)$ |
| Fall-out or false positive rate(**FPR**) | $FPR = FP/N = FP/(FP + TN)$ |
| False discovery rate(**FDR**) | $FDR = FP/(FP + TP) = 1 – PPV$ |
| Miss Rate or False Negative Rate(**FNR**) | $FNR = FN/P = FN/(FN + TP)$ |
| Accuracy(**ACC**) | $ACC = (TP + TN)/(P + N)$ |

Table 5: some useful formula about confusion matrix

有一个和混淆矩阵相关的曲线名为 ROC (Receiver Operating Characteristic Curve)。

**基本内容：**该曲线的横坐标为假阳性率（False Positive Rate, FPR），纵坐标为真阳性率（True Positive Rate, TPR）。其最直观的应用就是能反映模型在选取不同阈值的时候其敏感性（sensitivity, FPR）和其精确性（specificity, TPR）的趋势走向。

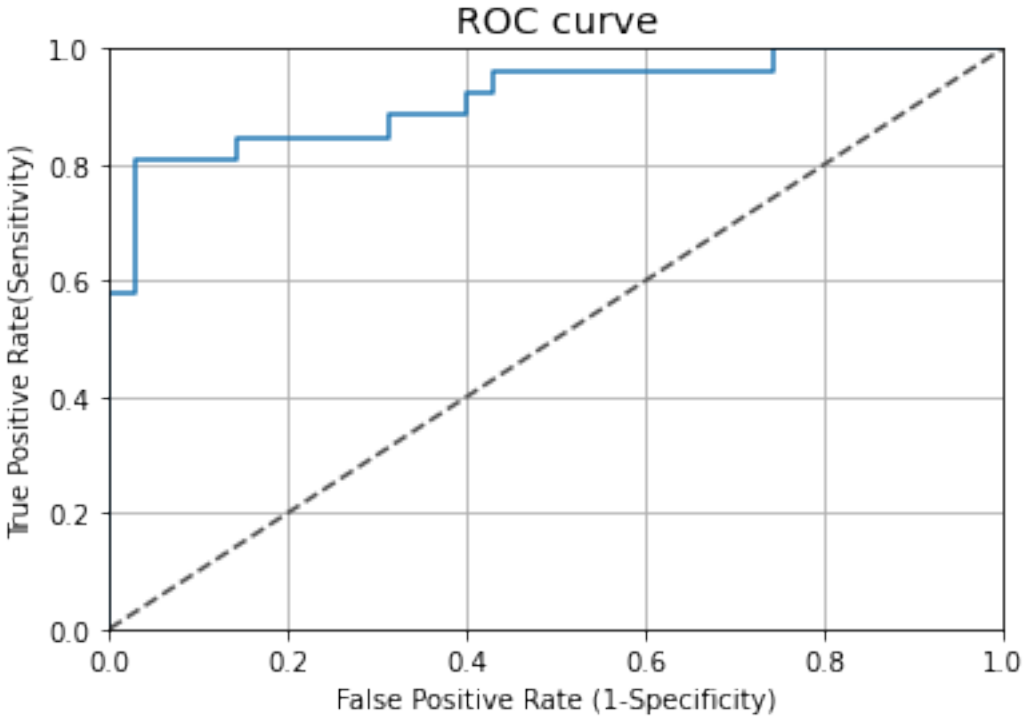


FIG8 ROC curve for naïve decision tree model

其曲线的积分 AUC 的值越高，说明该模型的性能越好。

## 2.4 模型可解释性

在上述的模型中，我们看到了计算机是如何通过受试者的 13 条信息就能以比较高的准确率判断受试者是否患有心脏病。但现实中，我们可能还要考虑这样一种情况，即是否某一特诊对于心脏病有关联。在 1.2.1 节中，我们已经给出了各个属性之间的关联度，但我们还未正式地给出各属性与心脏病之间的联系。通过可解释分析，可以指导特征工程。一般我们会根据一些专业知识和经验来做特征，同构分析特征重要性，可以挖掘更多有用的特征，尤其是在交互特征方面。当原始特征众多时，可解释性分析将特别重要。因此，这一节中我们就要讨论这一问题。

### 2.4.1 Permutation importance 排列重要性

Permutation Importance 是一种计算模型特征重要性的算法。

**基本思想：**假设要研究特征的重要性，那么将这列数据打乱，其他列的数据保持不变，然后观察预测的 metric(eg.准确率)或者 loss 变化了多少，根据变化量来决定特征重要性。

**实验结果：**每种模型的 permutation Importance 长得都不太一样，实验结果如图所示。

| Weight | Feature |
|---|---|
| 0.0656 ± 0.0464 | st depression |
| 0.0525 ± 0.0730 | chest pain type_typical angina |
| 0.0393 ± 0.0491 | max heart rate |
| 0.0361 ± 0.0787 | num major vessels |
| 0.0262 ± 0.0533 | thalassmia_fixed defect |
| 0.0230 ± 0.0334 | cholesterol |
| 0.0131 ± 0.0245 | age |
| 0.0131 ± 0.0131 | exercise induced angina_yes |
| 0.0131 ± 0.0131 | sex_male |
| 0.0066 ± 0.0572 | st slope_flat |
| 0.0033 ± 0.0131 | st slope_upsloping |
| 0.0033 ± 0.0131 | sex_female |
| 0 ± 0.0000 | chest pain type_non-anginal pain |
| 0 ± 0.0000 | chest pain type_asymptomatic |
| 0 ± 0.0000 | chest pain type_atypical angina |
| 0 ± 0.0000 | thalassmia_unknown |
| 0 ± 0.0000 | thalassmia_reversable defect |
| 0 ± 0.0000 | fasting blood sugar_lower than 120mg/L |
| 0 ± 0.0000 | rest ecg_ST-T wave abnormality |
| 0 ± 0.0000 | rest ecg_left ventricular hypertrophy |
| 0 ± 0.0000 | rest ecg_normal |
| 0 ± 0.0000 | exercise induced angina_no |
| 0 ± 0.0000 | thalassmia_normal |
| 0 ± 0.0000 | fasting blood sugar_greater than 120mg/L |
| -0.0164 ± 0.0000 | st slope_downsloping |
| -0.0328 ± 0.0207 | resting blood pressure |

| Weight | Feature |
|---|---|
| 0.0525 ± 0.0669 | st depression |
| 0.0393 ± 0.0736 | num major vessels |
| 0.0393 ± 0.0819 | chest pain type_typical angina |
| 0.0361 ± 0.0435 | max heart rate |
| 0.0230 ± 0.0262 | sex_female |
| 0.0164 ± 0.0415 | thalassmia_fixed defect |
| 0.0131 ± 0.0245 | cholesterol |
| 0.0098 ± 0.0334 | sex_male |
| 0.0066 ± 0.0608 | st slope_flat |
| 0 ± 0.0000 | thalassmia_unknown |
| 0 ± 0.0000 | chest pain type_asymptomatic |
| 0 ± 0.0000 | chest pain type_atypical angina |
| 0 ± 0.0000 | chest pain type_non-anginal pain |
| 0 ± 0.0000 | fasting blood sugar_greater than 120mg/L |
| 0 ± 0.0000 | thalassmia_reversable defect |
| 0 ± 0.0000 | fasting blood sugar_lower than 120mg/L |
| 0 ± 0.0000 | rest ecg_ST-T wave abnormality |
| 0 ± 0.0000 | rest ecg_left ventricular hypertrophy |
| 0 ± 0.0000 | rest ecg_normal |
| 0 ± 0.0000 | exercise induced angina_no |
| 0 ± 0.0000 | exercise induced angina_yes |
| 0 ± 0.0000 | st slope_downsloping |
| 0 ± 0.0000 | thalassmia_normal |
| -0.0066 ± 0.0161 | st slope_upsloping |
| -0.0066 ± 0.0161 | age |
| -0.0164 ± 0.0207 | resting blood pressure |

| Weight | Feature |
|---|---|
| 0.1705 ± 0.0736 | max heart rate |
| 0.1016 ± 0.1085 | age |
| 0.0623 ± 0.1143 | resting blood pressure |
| 0.0393 ± 0.0675 | cholesterol |
| 0 ± 0.0000 | thalassmia_reversable defect |
| 0 ± 0.0000 | st depression |
| 0 ± 0.0000 | num major vessels |
| 0 ± 0.0000 | sex_female |
| 0 ± 0.0000 | sex_male |
| 0 ± 0.0000 | chest pain type_asymptomatic |
| 0 ± 0.0000 | chest pain type_atypical angina |
| 0 ± 0.0000 | chest pain type_non-anginal pain |
| 0 ± 0.0000 | chest pain type_typical angina |
| 0 ± 0.0000 | thalassmia_unknown |
| 0 ± 0.0000 | fasting blood sugar_lower than 120mg/L |
| 0 ± 0.0000 | rest ecg_ST-T wave abnormality |
| 0 ± 0.0000 | rest ecg_left ventricular hypertrophy |
| 0 ± 0.0000 | rest ecg_normal |
| 0 ± 0.0000 | exercise induced angina_no |
| 0 ± 0.0000 | exercise induced angina_yes |
| 0 ± 0.0000 | st slope_downsloping |
| 0 ± 0.0000 | st slope_flat |
| 0 ± 0.0000 | st slope_upsloping |
| 0 ± 0.0000 | thalassmia_fixed defect |
| 0 ± 0.0000 | thalassmia_normal |
| 0 ± 0.0000 | fasting blood sugar_greater than 120mg/L |

| Weight | Feature |
|---|---|
| 0.0230 ± 0.0334 | thalassmia_fixed defect |
| 0.0197 ± 0.0131 | thalassmia_reversable defect |
| 0.0098 ± 0.0161 | chest pain type_atypical angina |
| 0.0098 ± 0.0262 | sex_male |
| 0.0033 ± 0.0245 | st depression |
| 0 ± 0.0000 | thalassmia_unknown |
| 0 ± 0.0000 | rest ecg_left ventricular hypertrophy |
| 0.0000 ± 0.0207 | chest pain type_typical angina |
| 0 ± 0.0000 | fasting blood sugar_lower than 120mg/L |
| 0 ± 0.0000 | fasting blood sugar_greater than 120mg/L |
| 0 ± 0.0000 | st slope_upsloping |
| 0 ± 0.0000 | thalassmia_normal |
| -0.0033 ± 0.0245 | age |
| -0.0066 ± 0.0161 | st slope_downsloping |
| -0.0066 ± 0.0262 | sex_female |
| -0.0066 ± 0.0393 | cholesterol |
| -0.0066 ± 0.0445 | num major vessels |
| -0.0098 ± 0.0161 | exercise induced angina_no |
| -0.0131 ± 0.0321 | exercise induced angina_yes |
| -0.0164 ± 0.0207 | rest ecg_normal |
| -0.0164 ± 0.0000 | st slope_flat |
| -0.0164 ± 0.0000 | chest pain type_asymptomatic |
| -0.0197 ± 0.0245 | max heart rate |
| -0.0262 ± 0.0161 | rest ecg_ST-T wave abnormality |
| -0.0262 ± 0.0334 | chest pain type_non-anginal pain |
| -0.0262 ± 0.0262 | resting blood pressure |

| Weight | Feature |
|---|---|
| 0.0525 ± 0.0245 | st depression |
| 0.0426 ± 0.0334 | thalassmia_fixed defect |
| 0.0426 ± 0.0393 | chest pain type_typical angina |
| 0.0328 ± 0.0207 | thalassmia_reversable defect |
| 0.0295 ± 0.0131 | chest pain type_asymptomatic |
| 0.0197 ± 0.0131 | cholesterol |
| 0.0197 ± 0.0245 | max heart rate |
| 0.0197 ± 0.0382 | sex_male |
| 0.0197 ± 0.0435 | exercise induced angina_no |
| 0.0164 ± 0.0000 | rest ecg_normal |
| 0.0164 ± 0.0207 | age |
| 0.0164 ± 0.0207 | st slope_flat |
| 0.0131 ± 0.0131 | st slope_downsloping |
| 0.0131 ± 0.0245 | exercise induced angina_yes |
| 0.0131 ± 0.0482 | chest pain type_non-anginal pain |
| 0.0098 ± 0.0161 | chest pain type_atypical angina |
| 0.0098 ± 0.0161 | thalassmia_normal |
| 0.0098 ± 0.0262 | sex_female |
| 0.0098 ± 0.0445 | num major vessels |
| 0.0066 ± 0.0161 | rest ecg_left ventricular hypertrophy |
| 0.0033 ± 0.0131 | st slope_upsloping |
| 0.0033 ± 0.0131 | rest ecg_ST-T wave abnormality |
| 0 ± 0.0000 | resting blood pressure |
| 0 ± 0.0000 | thalassmia_unknown |
| 0 ± 0.0000 | fasting blood sugar_lower than 120mg/L |
| 0 ± 0.0000 | fasting blood sugar_greater than 120mg/L |

| Weight | Feature |
|---|---|
| 0.1082 ± 0.0765 | max heart rate |
| 0 ± 0.0000 | thalassmia_unknown |
| 0 ± 0.0000 | fasting blood sugar_lower than 120mg/L |
| 0 ± 0.0000 | st depression |
| 0 ± 0.0000 | num major vessels |
| 0 ± 0.0000 | sex_female |
| 0 ± 0.0000 | sex_male |
| 0 ± 0.0000 | chest pain type_asymptomatic |
| 0 ± 0.0000 | chest pain type_atypical angina |
| 0 ± 0.0000 | chest pain type_non-anginal pain |
| 0 ± 0.0000 | chest pain type_typical angina |
| 0 ± 0.0000 | thalassmia_reversable defect |
| 0 ± 0.0000 | fasting blood sugar_greater than 120mg/L |
| 0 ± 0.0000 | rest ecg_ST-T wave abnormality |
| 0 ± 0.0000 | rest ecg_left ventricular hypertrophy |
| 0 ± 0.0000 | rest ecg_normal |
| 0 ± 0.0000 | exercise induced angina_no |
| 0 ± 0.0000 | exercise induced angina_yes |
| 0 ± 0.0000 | st slope_downsloping |
| 0 ± 0.0000 | st slope_flat |
| 0 ± 0.0000 | st slope_upsloping |
| 0 ± 0.0000 | thalassmia_fixed defect |
| 0 ± 0.0000 | thalassmia_normal |
| -0.0000 ± 0.0549 | resting blood pressure |
| -0.0066 ± 0.0161 | age |
| -0.0230 ± 0.0533 | cholesterol |

FIG9-14 permutation importance of different classification models `

(a. Random Forest b. Decision Tree c. Logistic Regression d. SVC e. GaussianNB f. KNN)

2.4.2 PDP (Partial Dependence Plot) 和 ICE (Individual Conditional Expectation Plot)

PDP 和 ICE 是很好的模型解释，模型 debug 和特征选择的工具。其公式如下

$$f^S(X^S) \approx \frac{1}{N}\sum_{1}^{N} f(X^S, X_i^C)$$

*其中：$f$为prediction model，$S$为自变量，$C$为其他自变量*

ICE 是右面和式中的任意一项组成的集合，而 PDP 则是其平均值。通过上文的 permutation importance 可以得到特征的重要性，但是不知道不同的特征值如何影响预测结果的。PDP 可以求得特征和预测结果的关系。

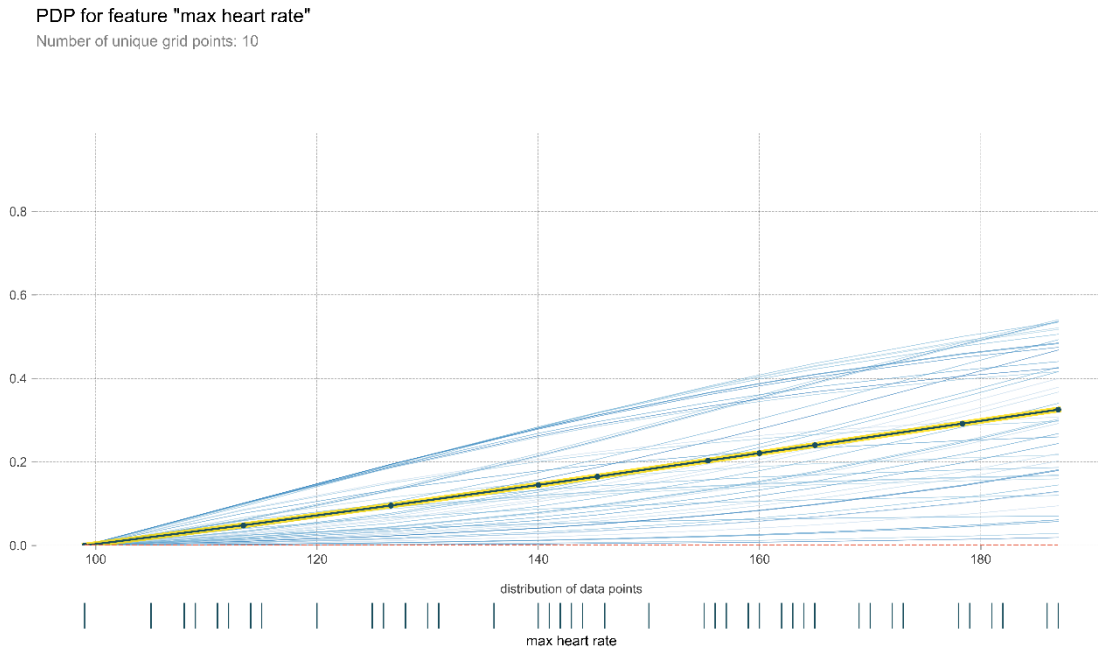


FIG15 PDP for feature max_heart_rate (model: logistic regression)

图中高亮的线即为 PDP，而其余蓝色的线即为 ICE。不同特征不同模型的 PDP 图均有所不同。图中采用的模型为 logistic regression，其与 random forest 模型较为相似。

### 2.4.3 SHAP (SHapley Additive exPlanation)

SHAP 所做的是量化每个特征对机器学习模型做出的预测的贡献。

**基本思想：**计算一个特征加入到模型时的边际贡献，然后考虑到该特征在所有的特征序列的情况下不同的边际贡献，取均值，即某该特征的 SHAP baselinevalue。

在此部分我们会给出两种解释的案例，第一种是整体的 SHAP 值，使用的是随机森林模型(max_depth=7,n_estimators=200)，第二种是对单个病人的分析。


FIG16 SHAP value (impact on RF model output)


FIG17 SHAP value of a specific patient

有了 shap 值，我们就可以比较直观的观察到模型背后的行为。对于整体，我们可以看到每一个因素对于整体的贡献；对于单一样本，我们更能看到每一部分对于其患病的贡献程度，因此我们可以结合医生的知识更有针对性的解决患者的问题。

## 三、 总结

本次实验大量参考了 kaggle 网站上的教学样例，b 站的教学视频，youtube 的案例以及 StackOverflow 和 CSDN 的 debug 解析。本次实验同样参考了很多 Wikipedia 和 sklearn 的 document，还有很多解释机器学习分类问题的 blog，学习了很多有关于分类问题的求解思路。

作为初学者，一开始会有些无所适从。仅仅知道最基本的概念和实验训练无法完整的支撑起整个数据科学的全过程。因此，在导论课上，要明确自己的定位就是理解数据是怎样获取，加工，建模，处理，可视化和展示的。

一开始我的想法是自己手动去爬虫数据，但是因为隐私问题和网站的权限问题作罢。然后又尝试了视觉的项目，但由于训练量很大，而且脱离这门课太多，因此也放弃了。最后在 kaggle 上找到了这个入门项目进行试验。本实验的源代码中混合了大量分类问题求解的“车轮子”。比如，使用 sklearn 快速进行二分类模型，使用 seaborn 进行可视化，使用 pandas 显示数据报告等等。同时，我也分析了 kaggle 高赞的 notebook，在提交的代码中有很多是我自己尝试复现他们结果的部分。因此，实际的代码量没有提交的代码量那么多。由于数据集的大小和实验数据集是同一量级的，处理这些数据并没有太大的难度，网上也有相当多处理的版本以供参考。

采用这个数据集的还有一个原因是其与我父母的工作高度近似，很多时候可以问问他们对于实验结果的体会，避免犯下原则性或低级的错误。

在进行实验的遇到过一些问题，有些是因为配置的关系，有些是因为版本的关系，有些则是由于 jupyter 本身的问题，有些则是自己 python 语法没学好造成的问题。经过这次实验，把之前实验课上后半段的数据建模过程好好的复习了一遍，也在相当程度上提高了自己 python 的实践能力。

最后，希望自己在之后的时间继续努力学习与数据相关的课程，早日做出有创新性的，较为前沿的内容。谢谢！