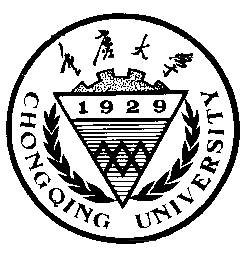
大数据分析实验报告

实验四



学 生：李嘉迪\_20191737

年 级：2019级

专 业：数据科学与大数据技术

重庆大学大数据与软件学院

2022年 5 月 31 日

## 实验目的：

理解因果分析原理；理解疾病数据集；掌握数据预处理方法；学会建立因果

模型；通过因果模型构建找出属性之间的因果关系及图谱。

## 实验要求：

通过公共疾病数据集 https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset 进行因果分析，用 PC 算法（PC 系列算法）、GES 算法等中的一种算法发现疾病数据的因果关系。

## 开发环境：

编程软件：anaconda/spyder/pycharm

环境：python3.6 以上、numpy、pandas、sklearn、Jupyter Notebook 等

## 实验内容：

具体要求步骤如下：

（1）加载并读取所分析的疾病数据集

（2）分析是否需要数据预处理：1）缺失值处理，2）标准化处理

（3）先进行相关性分析，进行特征选择；

（4）再做因果分析：建立疾病因果分析模型;

（5）对因果模型构建找出属性之间的因果关系及画出因果图谱。

（6）总结分析

数据集说明

# age : Age of the person

# sex : Gender of the person 1 = male; 0 = female

# cp : Chest Pain type

# trtbps : resting blood pressure (in mm Hg)

# chol : cholestoral in mg/dl fetched via BMI sensor

# fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

#restecg : resting electrocardiographic results

# thalachh : maximum heart rate achieved

# exng : exercise induced angina (1 = yes; 0 = no)

#oldpeak : Previous peak "ST depression induced by exercise

relative to rest”

# slp : Slope

#caa : number of major vessels (0-3)

# thall : Thal rate

# output : Target variable，0= less chance of heart attack 1= more

chance of heart attac

## 实验结果：

（1）加载所需要的数据并进行数据预处理，而后进行相关性分析

df = pd.read\_csv('data/heart.csv')

df.dropna(inplace=True)

df.reset\_index(drop=True, inplace=True)

scaler = MinMaxScaler()

df=pd.DataFrame(scaler.fit\_transform(df),columns=df.columns)

print(df.corr())

sns.heatmap(df.corr(), annot=True, vmax=1, square=True, cmap="Greens", fmt='.2g')

plt.show()

图形用户界面, 文本

描述已自动生成

树状图

描述已自动生成

（2）因果分析

labels = ["age", "sex", "cp", "trtbps", "chol","fbs","restecg","thalachh","exng","oldpeak","slp","caa","thall","output"]

row\_count = sum(1 for row in data)

p = pc(

suffStat = {"C": data.corr().values, "n": data.values.shape[0]},

alpha = 0.10,

labels = [str(i) for i in range(row\_count)],

indepTest = gauss\_ci\_test,

verbose = True

)

图表

描述已自动生成

从结果图可以看出，sex、thall、oldpeak、thalachh、exng、cp和caa这些属性与output的相关性更高

## 问题和解决：