机器学习：回归问题 实验报告

2021219105班 李梓轩 2021213519

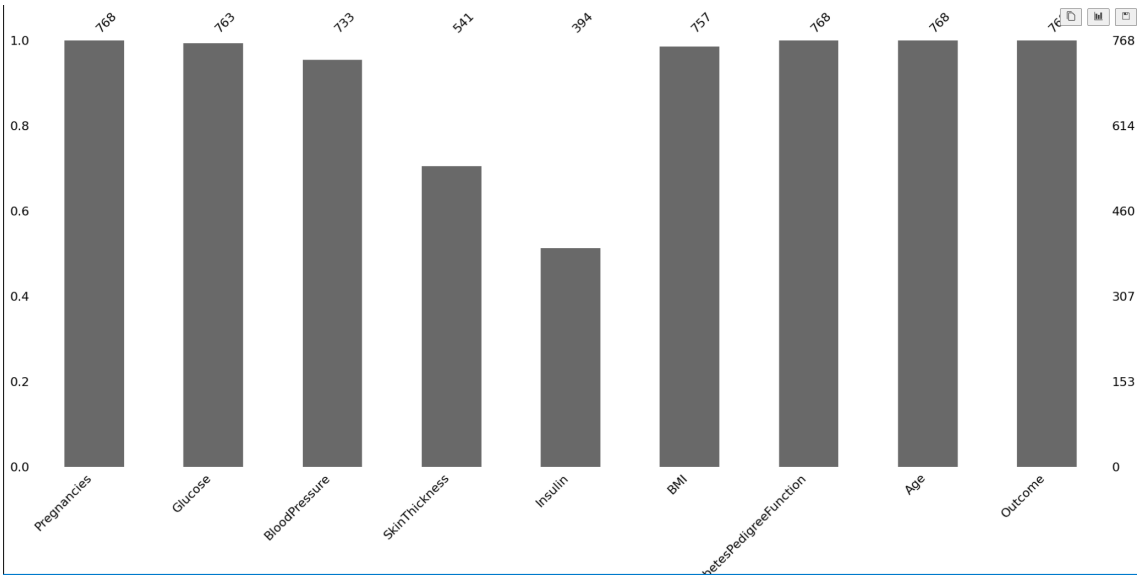
1. 使用特征工程方法对数据集中的特征变量进行处理

- **数据读取**

  使用pandas读取数据，并显示数据信息

- **缺省值处理**

  使用missingno库查看数据的缺失情况

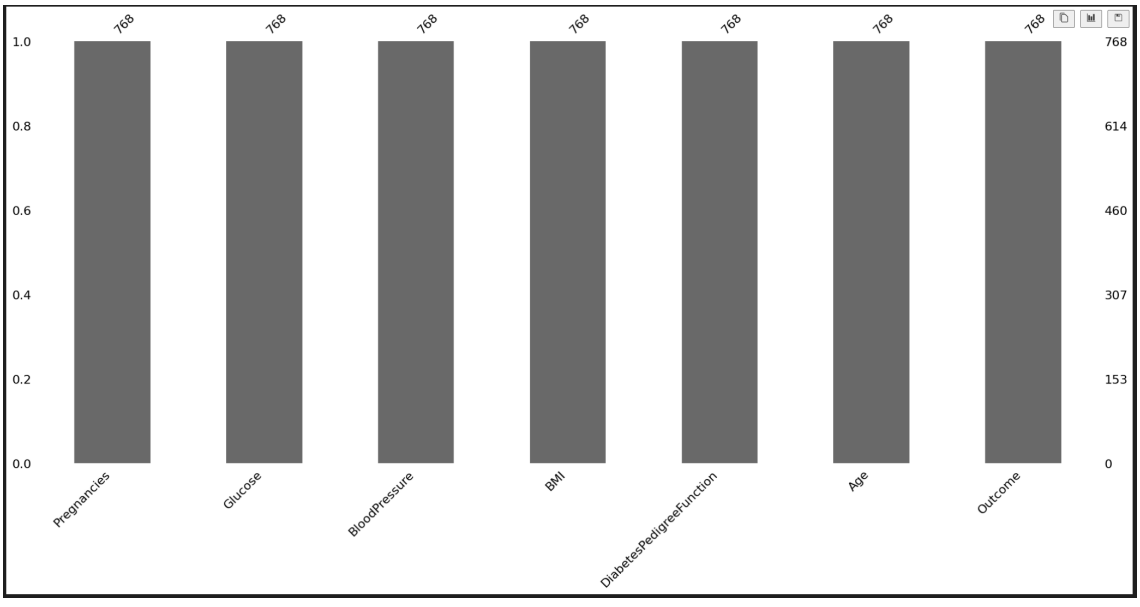


由于某些数据缺失过多，用均值等填充方法可能会造成较大误差，因此选择删除缺失20%以上的数据。

使用sklearn.impute库对数据进行填充，将Glucose、BloodPressure、 BMI的缺失值用均值填充。

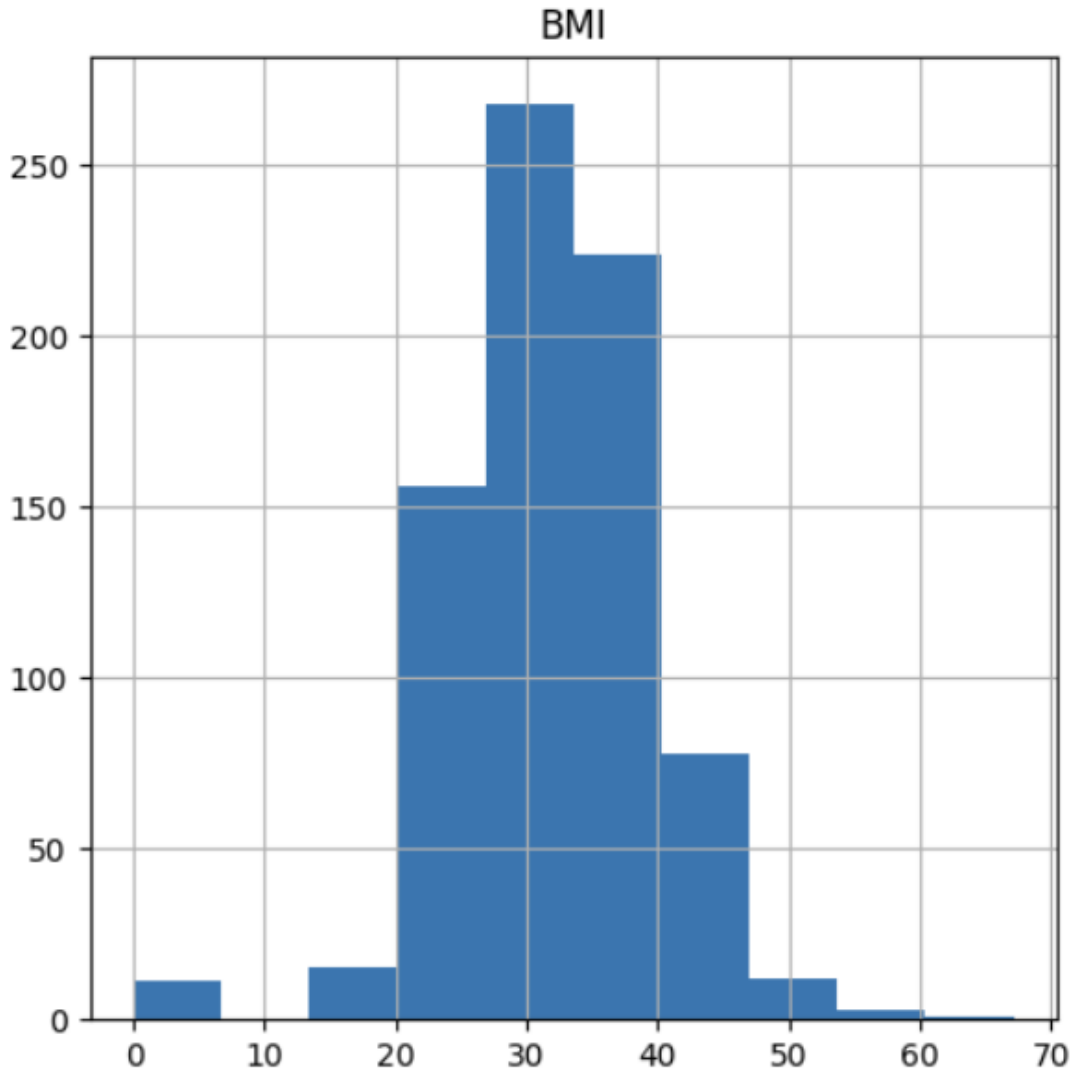
```
in_data[colume] = imp_mean.fit_transform(in_data[colume])
```

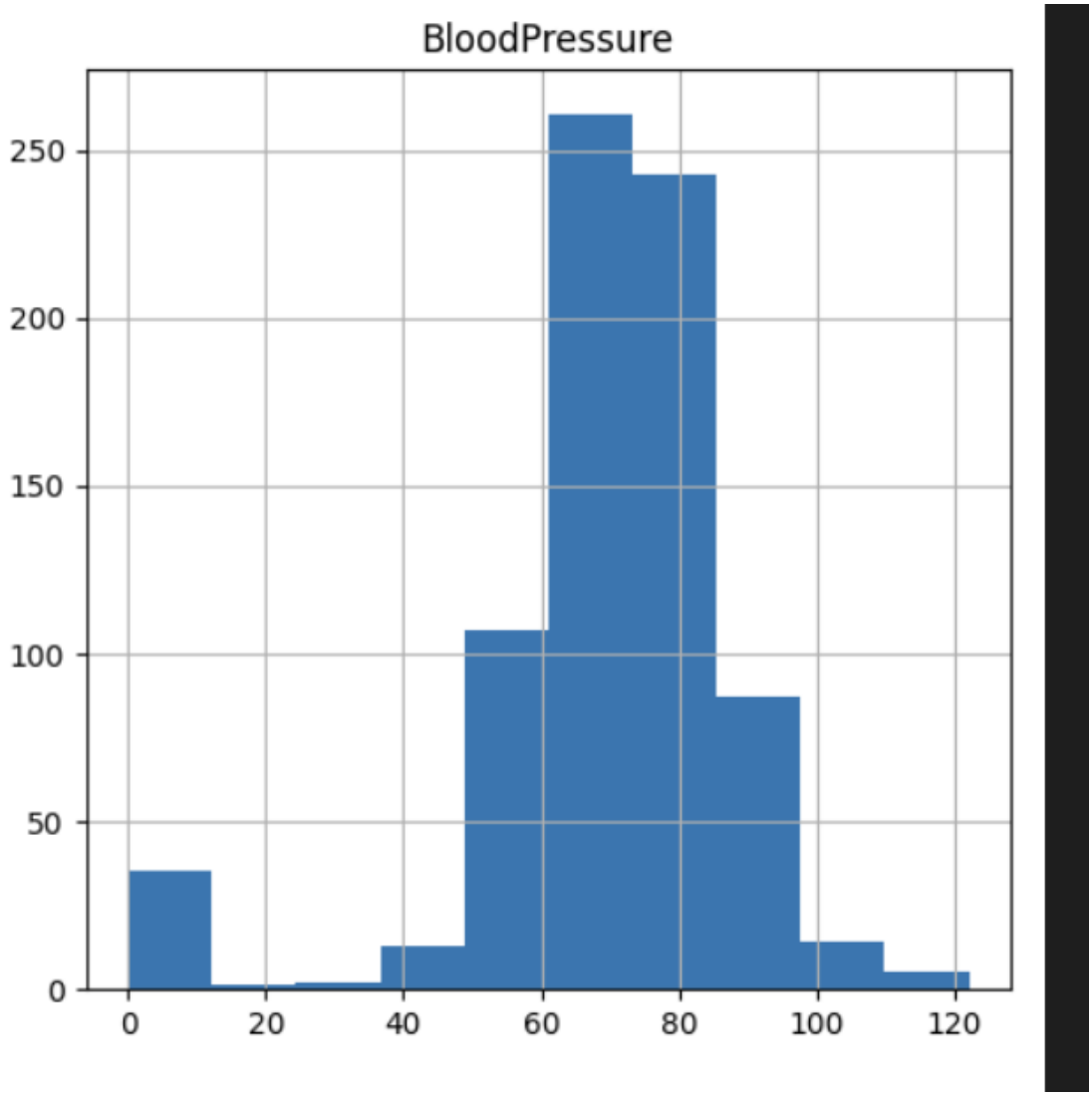
此时可以看到数据恢复完整



- **异常值处理**

  绘制各个特征值的柱状图


BMI
250
200
150
100
50
0
0
10
20
30
40
50
60
70



可以看出BloodPressure、Glucose等理论上应该是没有0值的，但是数据中存在0的异常值，将其设为NaN

```
colume = ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI']
```

```
in_data[colume] = in_data[colume].replace(0,np.nan)
```

- **数据标准化处理**

  使用StandarScaler库对数据进行标准化处理

```
X = pd.DataFrame(sc_X.fit_transform(in_data_copy.drop(["Outcome"],axis = 1),),columns=['Pregnancies', 'Glucose', 'BloodPressure','BMI', 'DiabetesPedigreeFunction', 'Age'])
```

2. 逻辑回归

- **拆分数据集**

  学号为奇数，测试集大小为0.3、0.25、0.2。

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

- **逻辑回归**

  当测试集为0.2时，初始化逻辑回归模型

```
lr1 = LogisticRegression(penalty="l2",solver="sag", C=0.5,max_iter=1000, random_state=2, multi_class='multinomial')
```

对回归模型进行训练并预测，并输出其准确率

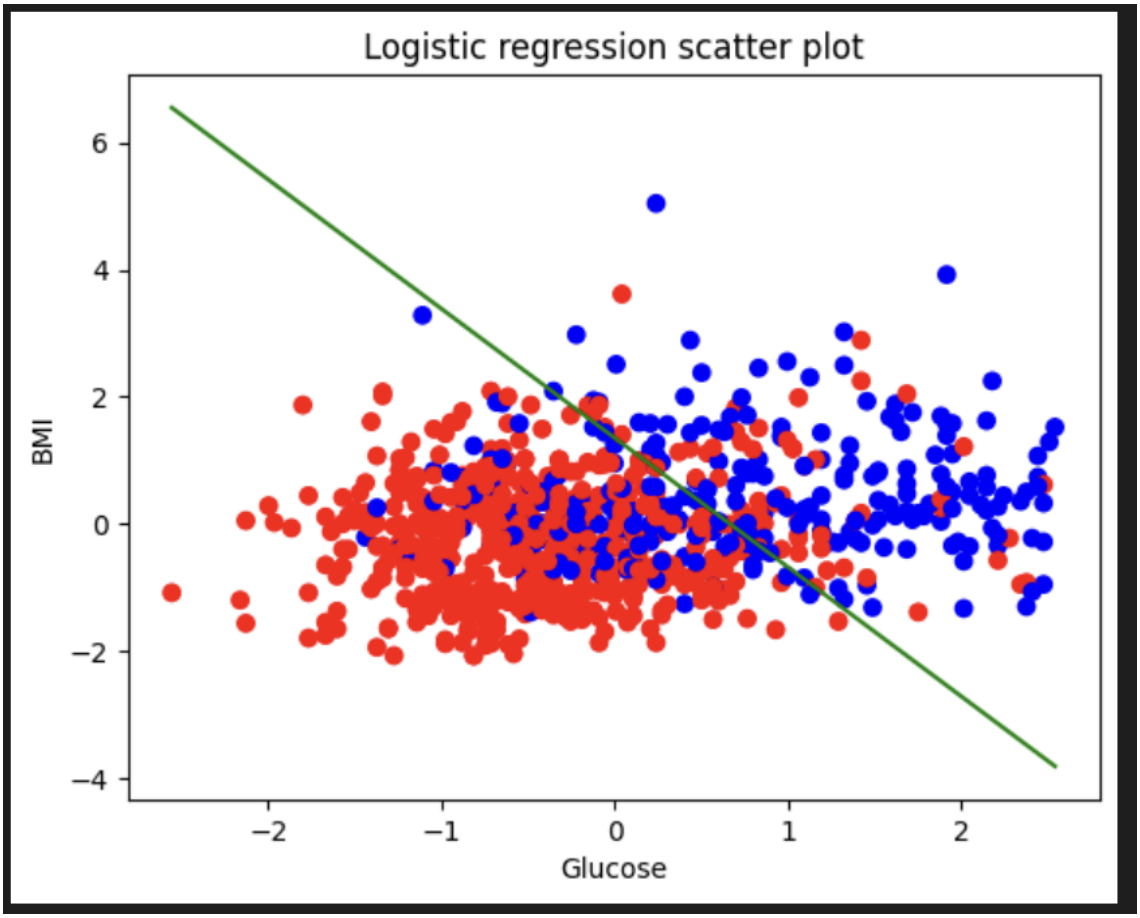
```
Accuracy1: 0.8051948051948052
```

- **绘制逻辑回归散点图**

  使用相关性（下文分析）最高的BMI和Gloucose作为自变量输入。与之前的步骤相同，将数据标准化，训练逻辑回归模型并预测，准确率如下

```
Accuracy: 0.7857142857142857
```

  绘制逻辑回归散点图

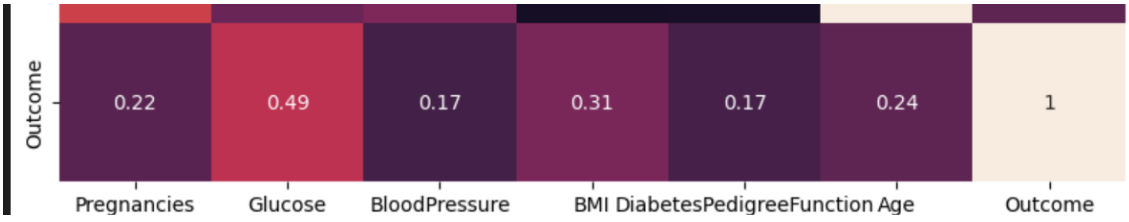


  可以看出，训练的模型可以较为精确地将糖尿病和非糖尿病分离

3. 分析特征值与变量的关系

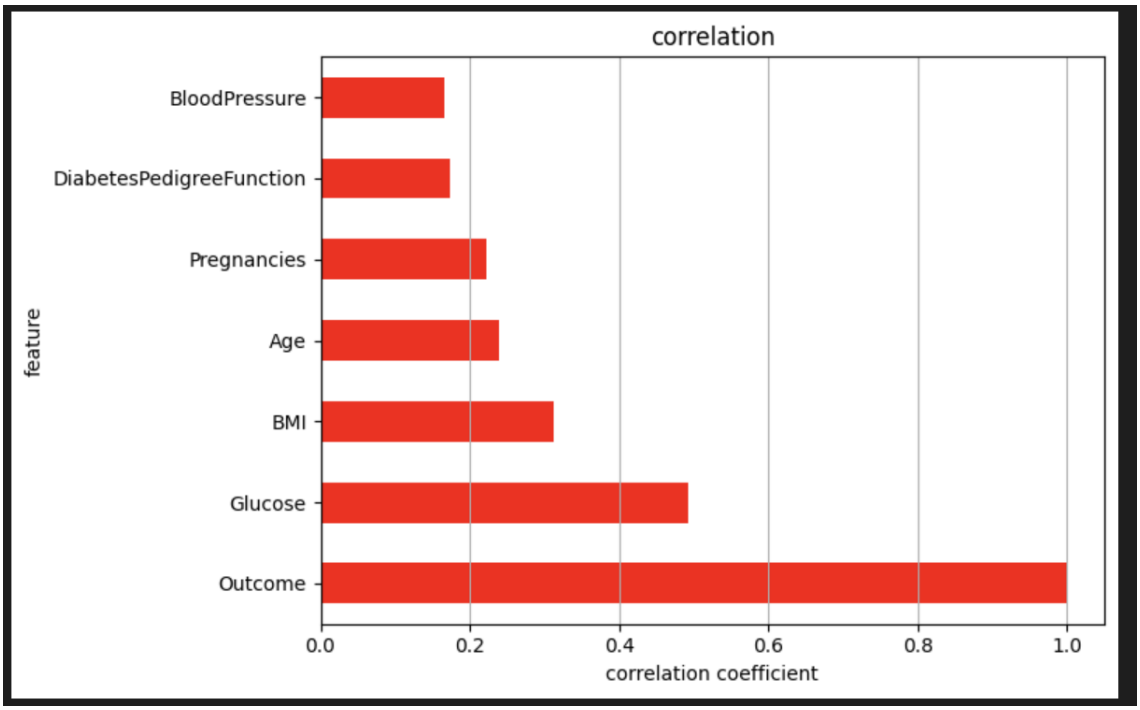
- 绘制热力图

  部分热力图如下，可以看到各个属性和Outcome之间的相关性大小，最大的为Glucose



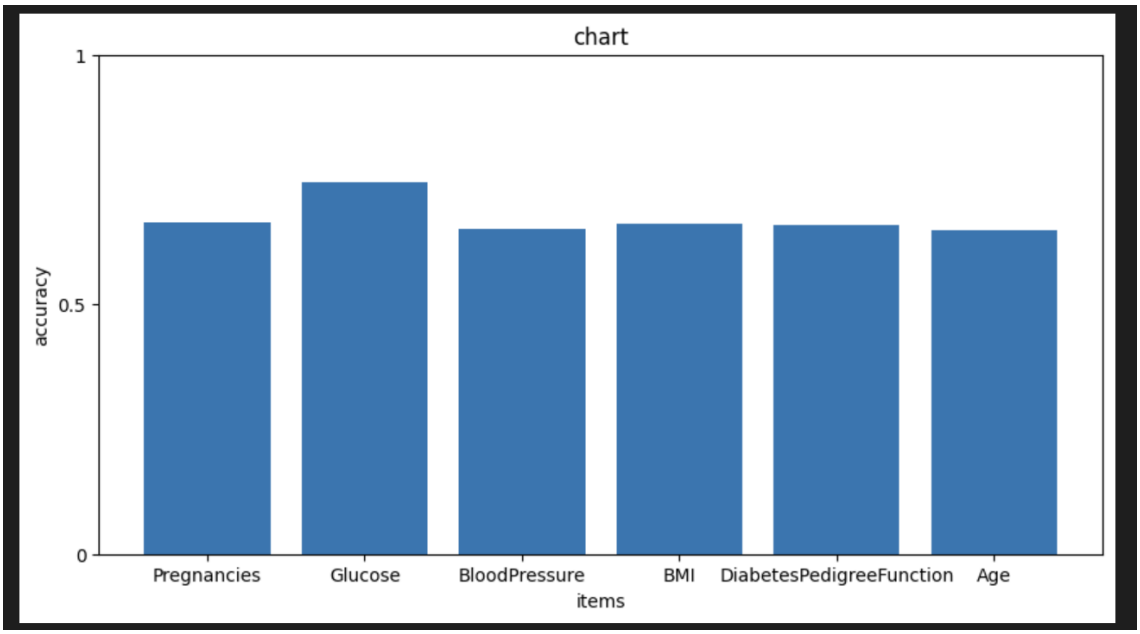
- 排序

  将各个属性的相关性大小进行排序



- 单个和多个自变量和输出的关系

  将每个自变量单独作为输入，训练逻辑回归模型并进行预测，得出的准确率如下



可以看到，准确率大小和相关性大小近似成正相关。Glucose与Outcome的相关性最大，其预测准确率也最高。而多个特征值预测时，由于不同特征值与结果的相关性不同，可以根据此调整不同特征值的占比，可以更高的预测。