糖尿病预测及可视化

姓名： 李帅辉

学号： 4612210109

1. 数据预处理

**1.1 空缺值的检测与处理**

1.1.1 空缺值的检测

章节标题号都为Times New Roman。需顶格书写。

一级标题，使用四号、黑体；二级标题，使用小四、宋体、加粗；三级标题，使用小四、宋体。标题号与标题之间空一格，行间距为1.5倍。

每章另起一页。全文统一使用小四、宋体、两端对齐、行间距1.5倍；英文使用Times New Roman；图片居中，要有标题，标题字体为五号、宋体，示例如图1-1所示。



图1-1 四叶草

图号与图标题之间空一格，图号按照章开始编号，例如：第一章的第一个图，图号为图1-1，第二章的第一个图，图号为图2-1。

2. 数据可视化

文章语句要简洁，尽量使用书面语言，不要口语化的表述。

页码从第一章开始，起始页从1开始。

**2.1.1 第一个问题**

文档命名统一使用“学号+姓名+题目”的形式。

**2.1.2 论文内容要求**

要包含机器学习，包括机器学习的整个流程，数据导入，数据预处理，特征工程（根据实际情况进行标准化、特征变换、特征选择等），建模（分类、预测、聚类、推荐），模型训练，预测，性能评估;可视化部分，可以选择FineBI、Tableau工具或者echarts，最终完成一个数据大屏。数据大屏要求图形多样化，不能少于4种基本图形，不能少于6个图形，要求位置、配色美观，可交互。

1. 数据导入

**1.1 使用pandas对数据进行导入**



部分数据如下：

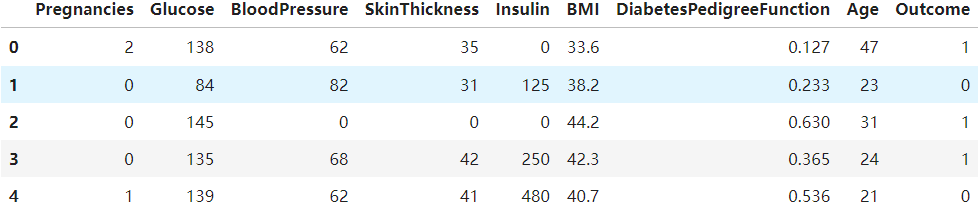


图 1-1部分数据

**1.2 变量解释**

1.Pregnancies: 怀孕次数

描述女性有多少次怀孕。

2.Glucose: 葡萄糖浓度

血浆葡萄糖浓度（口服葡萄糖耐量试验中的2小时值）。

3.BloodPressure: 血压

患者的舒张压（毫米汞柱）。

4.SkinThickness: 皮肤厚度

三轴测量的患者的皮肤褶皱厚度（毫米）。

5.Insulin: 胰岛素

两小时血清胰岛素（μU/毫升）。

6.BMI: 体质指数

体重指数（体重（公斤）/（身高（米）^2））。

7.DiabetesPedigreeFunction: 糖尿病谱系功能

血统中糖尿病的概率函数。

8.Age: 年龄

患者的年龄。

9.Outcome: 结果

二元变量，1 表示患者患有糖尿病，0 表示患者未患有糖尿病。

1. 数据探索性分析
   1. 数据详情

数据详情如下

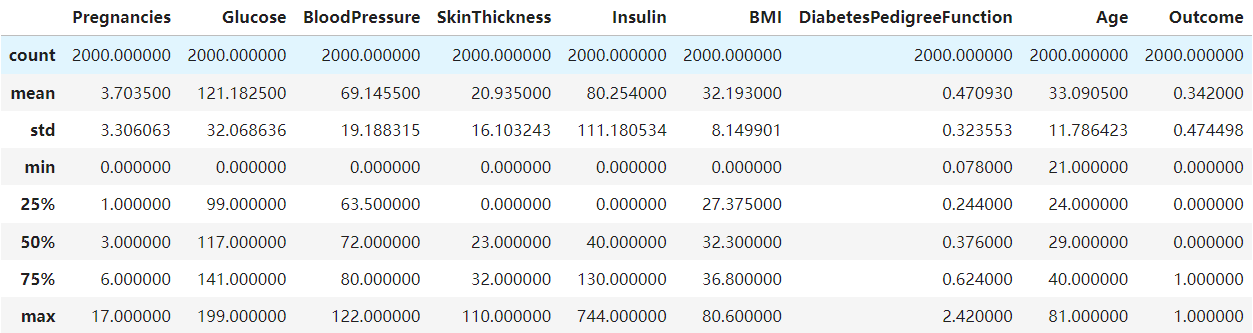


图 2-1数据详情

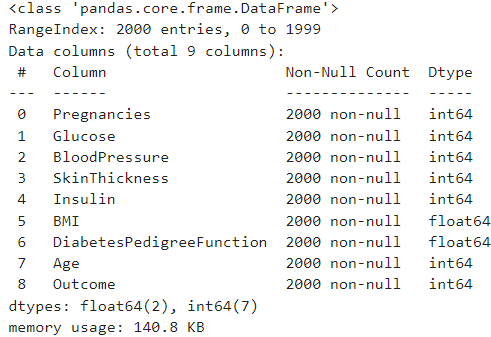


图2-2数据详情

* 1. **查看空缺值**

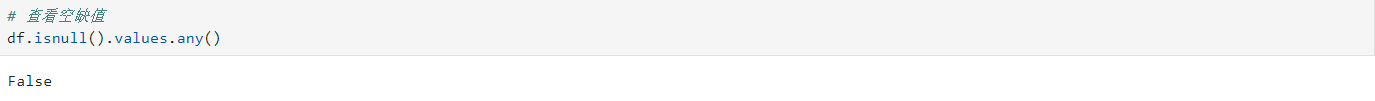
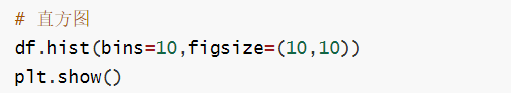


图2-3查看空缺值

1. 数据可视化及探索性分析
   1. **直方图**

绘制直方图的目的是可视化每个数值型变量的分布情况。直方图将显示数据在不同区间的频率，从而帮助我们了解数据的整体形状、集中趋势和分散程度。例如，我们可以看到每个变量的值分布情况，是否存在偏斜，是否有异常值等。



bins=10: 这指定直方图的箱子数量。直方图将数据范围划分为多个箱子，每个箱子显示数据的分布。在这里，选择了 10 个箱子，表示数据范围将被划分为 10 个等宽的区间。

figsize=(10, 10): 这定义了绘图的图形大小，宽度和高度均为 10 英寸。

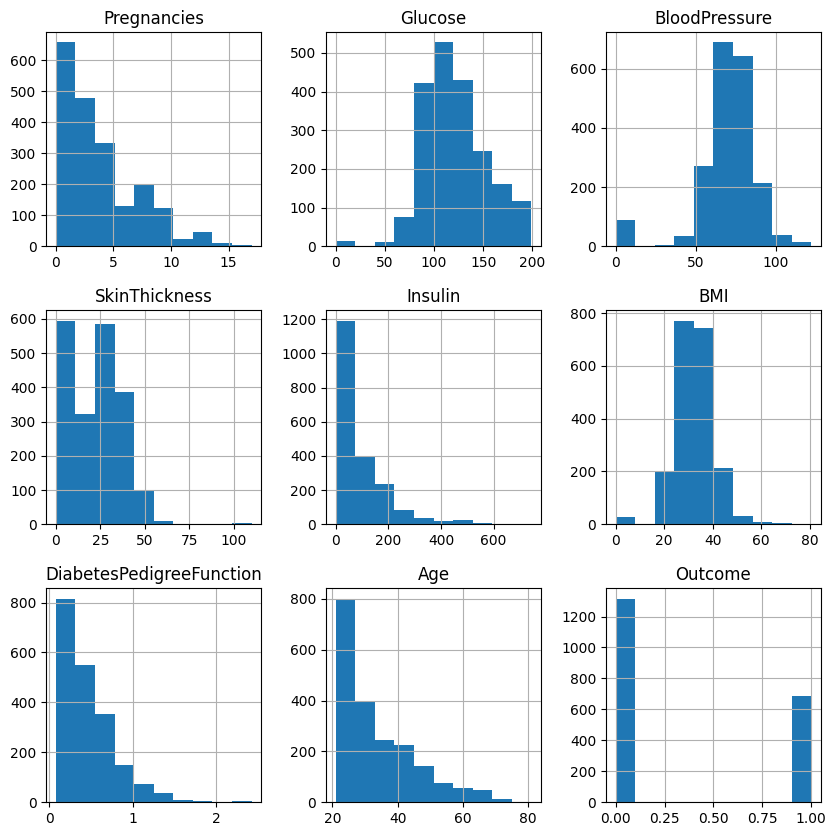


图3- 1直方图

* 1. **相关热力图**

通过绘制相关热力图，你可以了解数据集中各个特征之间的相关性。这对于判断特征之间是否存在线性关系以及它们与目标变量之间的关系非常有帮助。

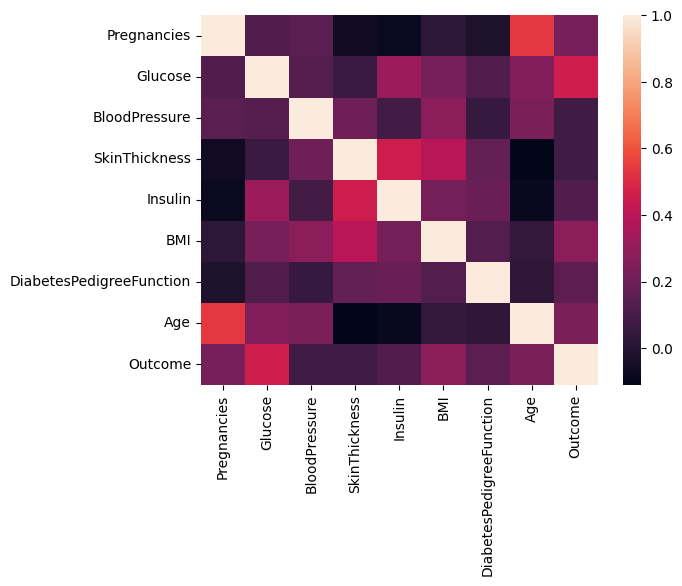
****

图3- 2相关热力图

从相关热力图中我们可以得到相关信息：

1.我们可以看到皮肤厚度、胰岛素、怀孕次数和年龄之间相互独立。

2.年龄和怀孕次数呈负相关。

* 1. **类别分布可视化**

类别分布可视化的主要作用是帮助我们了解数据集中类别变量的分布情况。对于二元分类问题，这通常是目标变量的分布，而对于多类别分类问题，可以是任何类别型变量的分布。

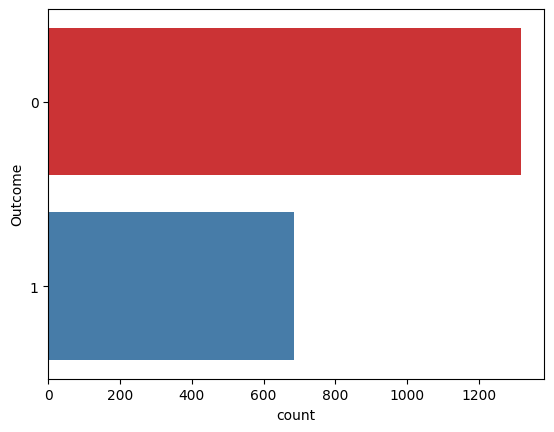
****

图3- 3计数条形图

0 表示没有糖尿病（no diabetes）。

1 表示患有糖尿病的患者（patient with diabetes）。

从图中我们可以看出患有糖尿病与没有糖尿病比例接近1：2。

1. 特征工程
   1. **绘制特征之间的关系图**

**sns.set(style="ticks")**

**sns.pairplot(df, hue="Outcome")**

sns.set(style="ticks"): 这是 Seaborn 库的一个设置函数，它设置了绘图的外观风格。在这里，使用 "ticks" 风格。

sns.pairplot(): 这是 Seaborn 库中用于绘制成对关系图的函数。它会绘制数据框中每两个数值型变量之间的散点图，并在对角线上显示每个单独变量的直方图。

hue="Outcome": 这表示通过 "Outcome" 变量的不同取值对散点图进行颜色标记。在这里，"Outcome" 是分类变量，代表是否患有糖尿病。因此，图中的散点会根据患糖尿病与否使用不同的颜色表示。

这个成对关系图的主要目的是通过可视化来探索数据集中各个数值型变量之间的关系，并看看这些关系是否随着糖尿病患病与否的不同而有所变化。每个点代表数据集中的一个样本，不同颜色的点表示不同的糖尿病患病情况。

通过这样的图表，可以观察到变量之间的散点分布、可能的趋势以及是否存在群集。这对于后续的特征选择、建模和解释模型中的结果都是有益的。



图4- 1 特征之间的关系图

通过特征之间关系图，我们可以大概了解到数据中是存在部分异常数据即图中有明显利群点，说明数据中存在离群值，我们需要对离群值进行处理。

**4.2 异常值检测和处理**

4.2.1 异常值检测和处理

可以通过绘制并分析变量箱型图来计算并删除离群值。

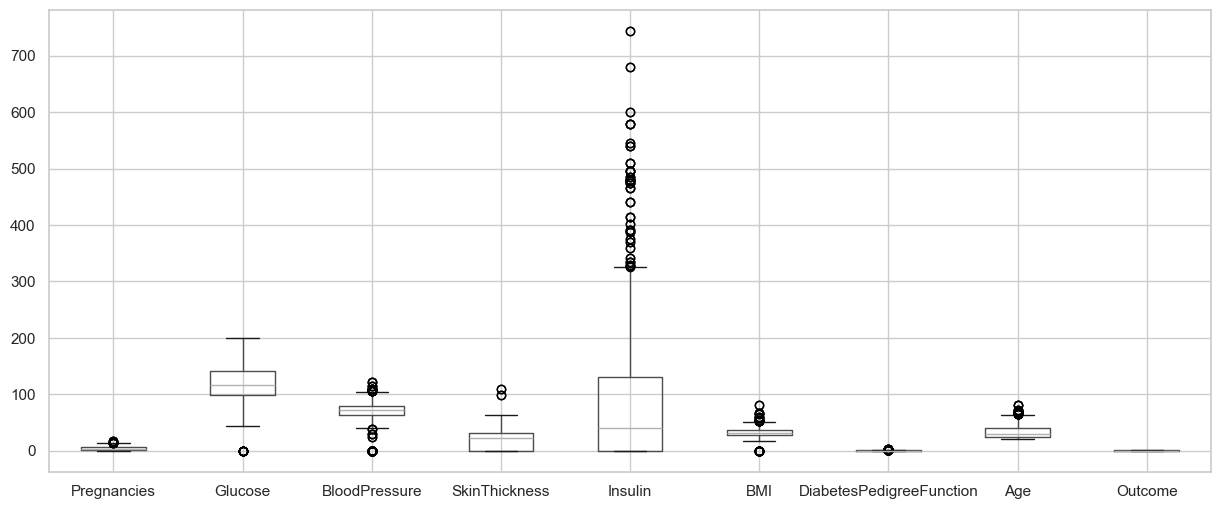


图4- 2 箱型图

查看Insulin,BloodPressure,DiabetesPedigreeFunction变量箱型图

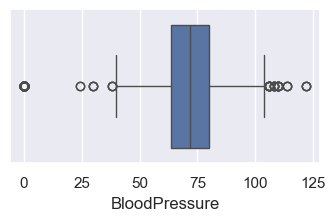


图4- 3 BloodPressure

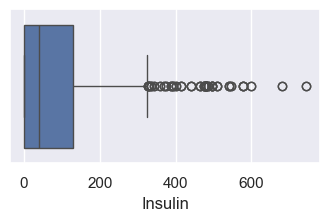


图4- 4 Insulin

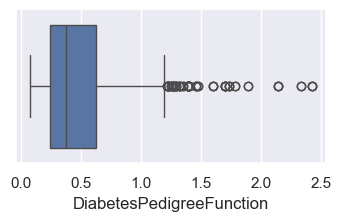
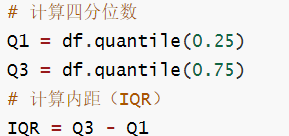


图4- 5 DiabedigreeFunction

根据箱型图，我们计算四分位数和内距



---Q1---

Pregnancies 1.000

Glucose 99.000

BloodPressure 63.500

SkinThickness 0.000

Insulin 0.000

BMI 27.375

DiabetesPedigreeFunction 0.244

Age 24.000

Outcome 0.000

Name: 0.25, dtype: float64

---Q3---

Pregnancies 6.000

Glucose 141.000

BloodPressure 80.000

SkinThickness 32.000

Insulin 130.000

BMI 36.800

DiabetesPedigreeFunction 0.624

Age 40.000

Outcome 1.000

Name: 0.75, dtype: float64

---IQR---

Pregnancies 5.000

Glucose 42.000

BloodPressure 16.500

SkinThickness 32.000

Insulin 130.000

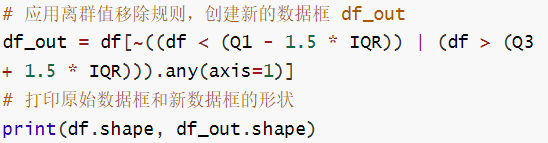
BMI 9.425

DiabetesPedigreeFunction 0.380

Age 16.000

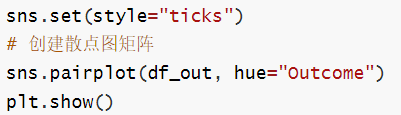
Outcome 1.000

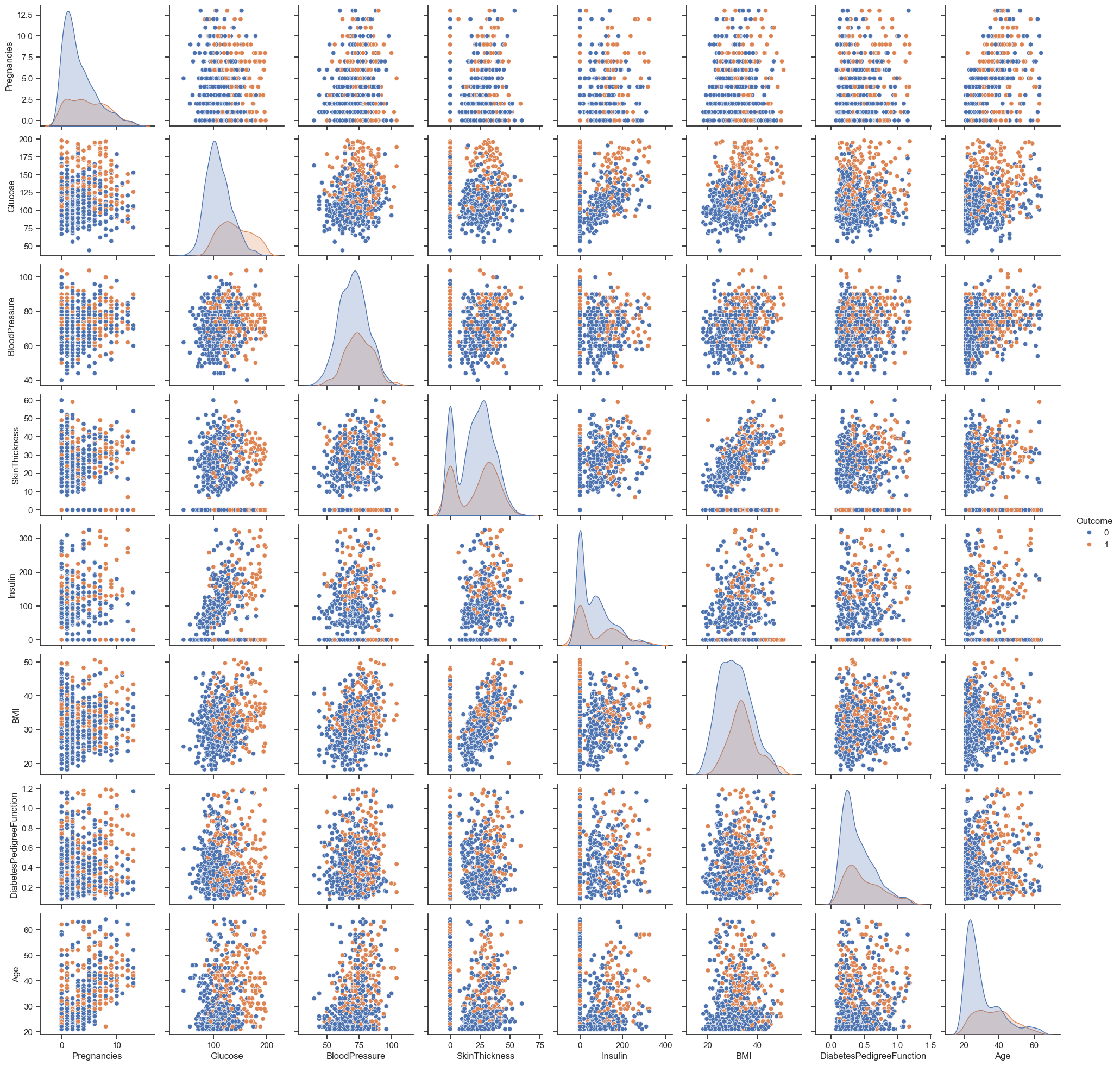
删除离群值



80多条记录被删除

* 1. **绘制散点矩阵并对比**





通过比对，明显看出离群值被删除。

**4.3 提取特征值和目标**

X = df\_out.drop(columns=['Outcome']) # 特征矩阵，删除 Outcome 列

y = df\_out['Outcome'] # 目标变量，仅包含 Outcome 列

**4.4拆分数据集和测试集**

**from sklearn.model\_selection import train\_test\_split**

**# 使用 train\_test\_split 函数分割数据**

**# test\_size=0.2 表示测试集占总数据的20%**

**train\_X, test\_X, train\_y, test\_y = train\_test\_split(X, y, test\_size=0.2)**

**print(train\_X.shape, test\_X.shape, train\_y.shape, test\_y.shape)**



图4- 6 输出结果

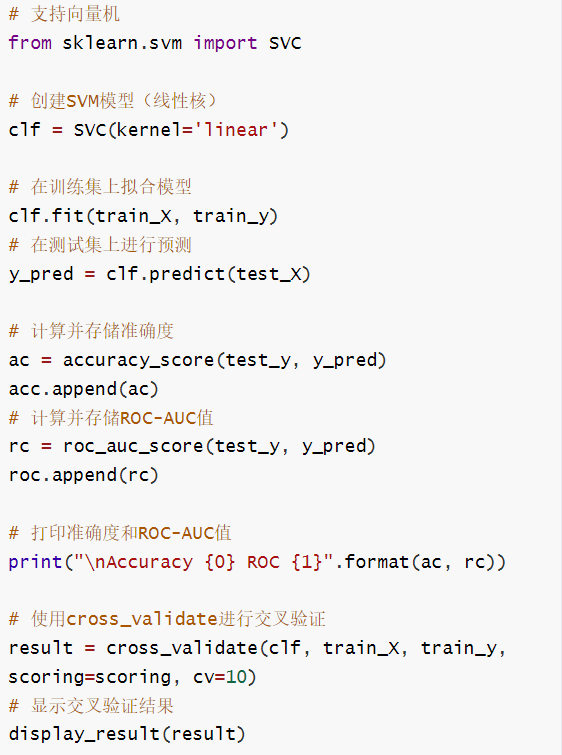
1. 建立模型与性能评估
   1. **定义性能评估函数**



图5- 1 相关代码

定义过性能评估函数后，进行初始化列表，用于存储准确度和ROC-AUC值，我们接下来就可以进行建模预测。

* 1. **支持向量机**



预测结果如下：

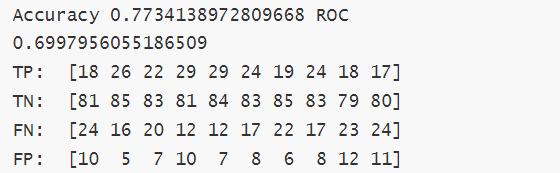
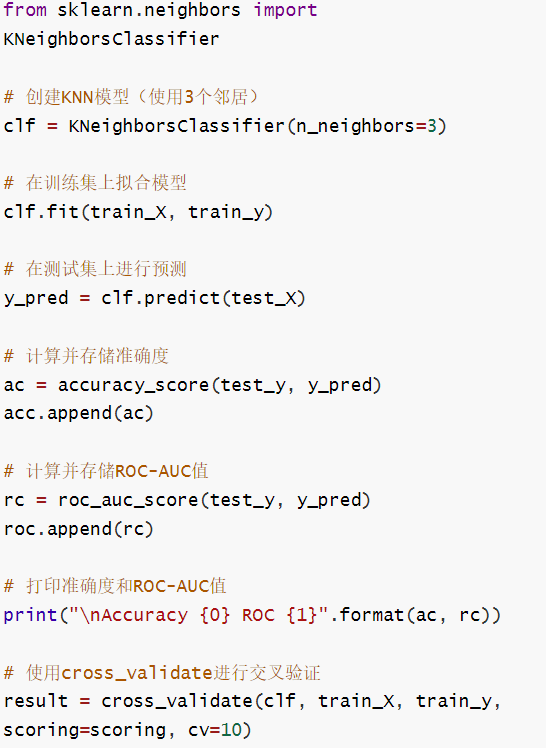


图5- 2 预测结果

* 1. **KNN模型**



预测结果如下：

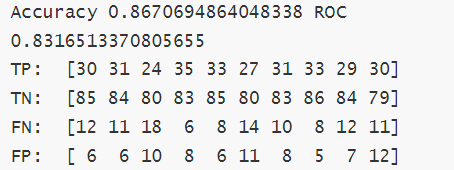
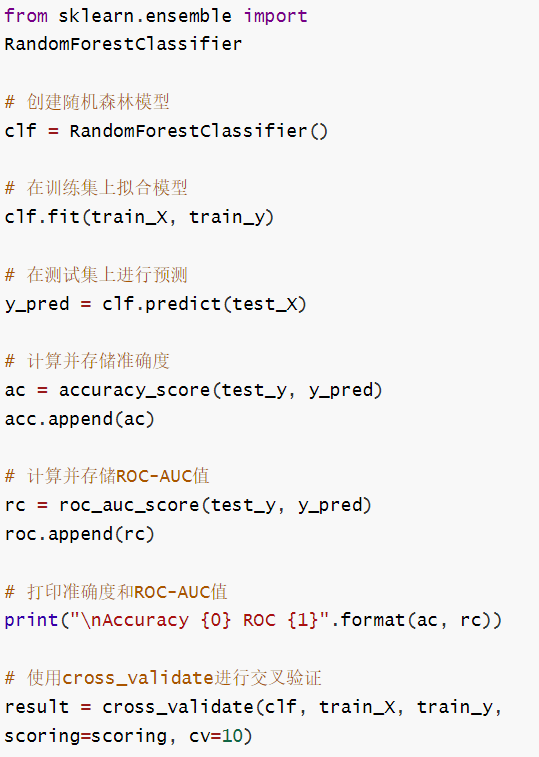


图5- 3 预测结果

* 1. **随机森林模型**



预测结果如下：

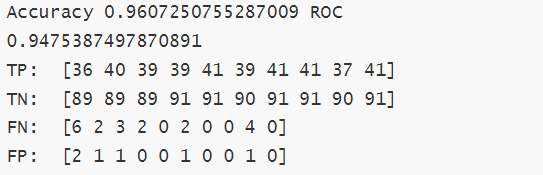
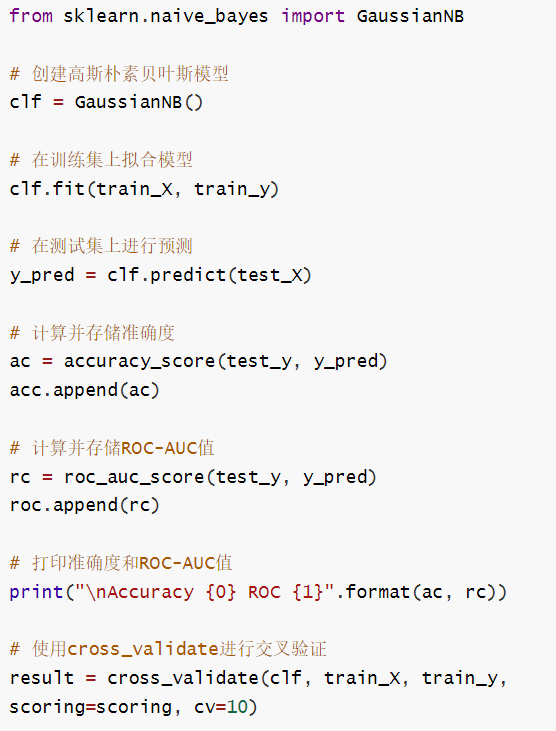


图5- 4 预测结果

* 1. **高斯朴素贝叶斯模型**



预测结果：

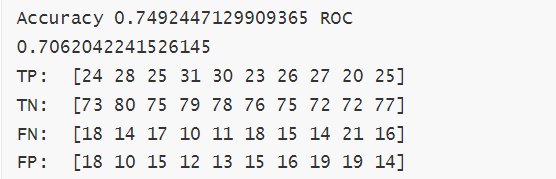


图5- 5 预测结果

* 1. **梯度提升分类器**



预测结果：

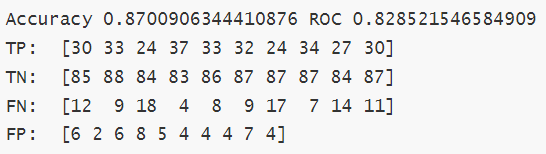
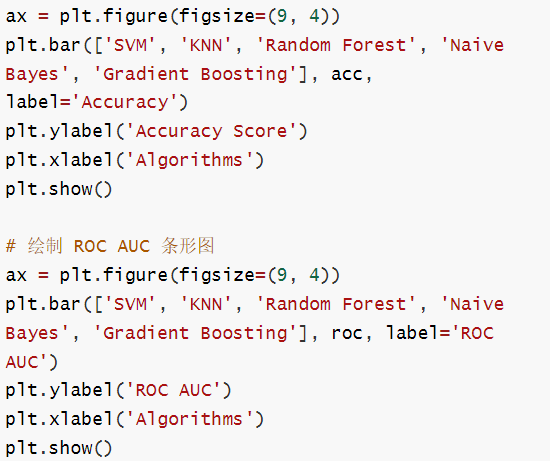


图5- 6 预测结果

* 1. **绘制各个模型准确条形图及ROC AUC条形图**



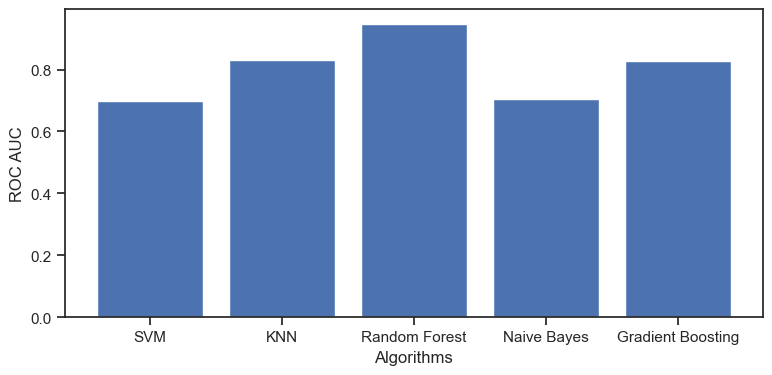
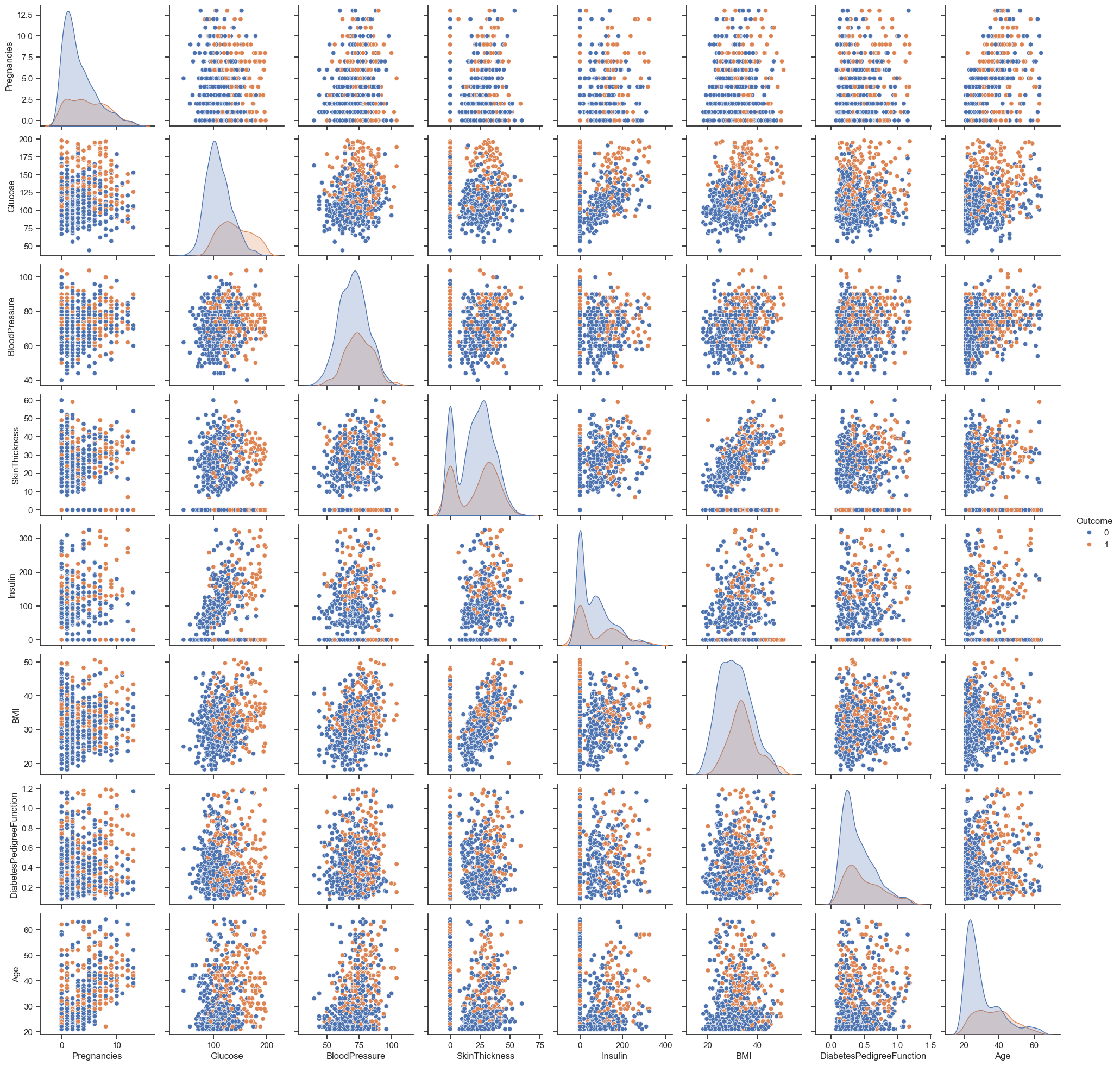


图5- 7ROC AOC条形图

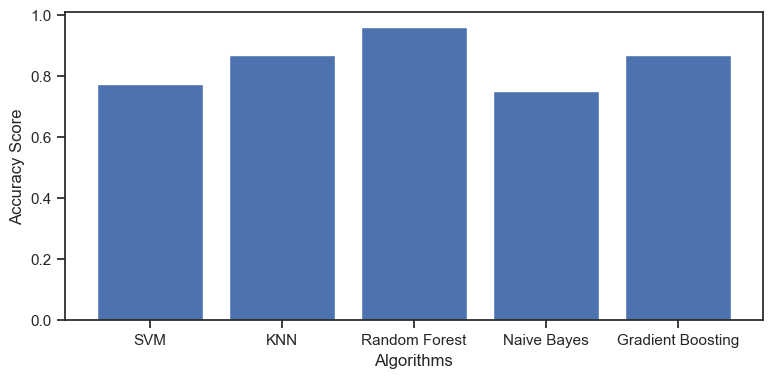


图5- 8 准确度条形图

随机森林的准确度最高为98%，ROC\_AUC曲线为97%

在我们的模型中，30%的样本是糖尿病患者，70%的样本是非糖尿病患者