糖尿病预测项目

赵思祺[1]，丘碧阳[1,2]

（1. 深圳技术大学，深圳 518000）

（2. 深圳技术大学，深圳 518000）

摘 要：本预测项目旨在通过年龄、体重指数、舒张压等参数，来判断一个人是否患有糖尿病。使用 Python 对拿到的各个医学数据与糖尿病的关系进行可视化分析。使用机器学习工具进行推断性分析。对数据进行标准化，使用逻辑回归算法模型和基于信息熵的决策树模型对测试集进行预测，最后使用混淆模型和准确率来对结果进行评估。最后得出结论：血浆葡萄糖浓度、三头肌皮褶厚度、血清胰岛素浓度、体重指数是与糖尿病相关性较强的参数。

关键词：糖尿病；逻辑回归模型；决策树模型

分类号：TP391

1 项目主题介绍

1.1 研究背景及意义

1.1.1 糖尿病背景介绍

糖尿病是一种以血葡萄糖浓度慢性增高为特征的代谢性疾病，其中主要包括 1 型糖尿病和 2 型糖尿病，该病症主要临床症状为多饮、多尿、多食以及体重下降，病情严重者可能会导致肾衰竭。据统计，我国糖尿病患者约 1.3 亿，潜在糖尿病患者约 5000 万，现如今越来越多的年轻人对甜品美食难以抗拒，对运动锻炼却“拒之门外”，使得年轻人患糖尿病的概率比以往增长了数倍，糖尿病正在快速年轻化。

1.1.2 研究意义

本文根据研究探究糖尿病与某些因素的潜在关系，并根据这些因素特征来对皮马印第安女性进行糖尿病预测，从而建立一套可以进行自测并提前预警的预测模型，来对有高风险患糖尿病的人群基于科学建议。

1.2 数据集介绍

1.2.1 数据集信息介绍

该数据集源自 kaggle 网站，统计于国家糖尿病、消化和肾脏疾病研究所，该数据集共有八个特征（分别为怀孕次数、血压、皮肤厚度、胰岛素、BMI、糖尿病遗传函数、年龄，如下表所示）和一个对应标签（糖尿病标签，0 为无糖尿病，1 为有糖尿病）。数据所统计的均为年龄为 21 岁以上的皮马印第安人血统的女性患者。

表 1 数据集特征介绍

数据特征、标签名	类型	描述
Pregnancies	Integer	怀孕次数
Glucose	Integer	口服葡萄糖耐量实验中血浆葡萄糖浓度为 2 小时
BloodPressure	Integer	舒张压(mmHg)
Skin Thickness	Integer	肱三头肌皮褶厚度(mm)
Insulin	Integer	2 小时血清胰岛素(μU/ml)
BMI	Integer	体重指数(kg/(m)^2)

Diabetes PedigreeFunction	Integer	糖尿病谱系功能
Age	Integer	年龄(岁)
Outcome	Integer	类变量(患糖尿病为 1,不患糖尿病为 0)

2 数据预处理

2.1 数据清洗

2.1.1 缺省数据分析

在刚收集到的数据集中会出现部分数据混乱，数据缺失或者数据意义不符合事实，这些数据会影响后续实验的准确性，所以首先进行缺省值的替换。在本项目里，我们将各个特征的均值算出，在将均值替换缺省值，从而达到提高数据集研究价值。

2.2 理解数据

2.2.1 可视化并分析数据

根据已知的数据集和各个特征，可以运用 python 来完成数据的可视化，进而观察特征与糖尿病是否有直接关系。

首先剖析数据集，运用柱状图绘图，由图 1 可直观得出数据集中患病人数远多于不患病人数，患病比大概为 1.866（500:268）

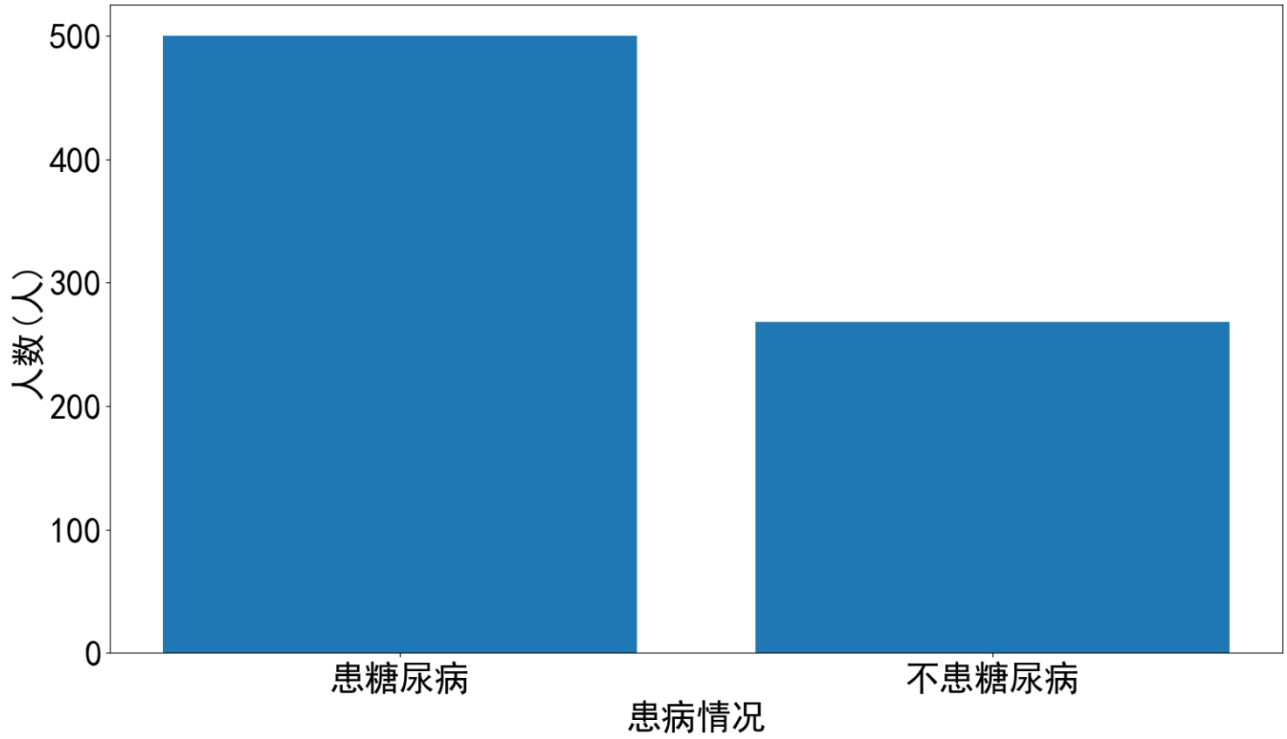

图 1 数据集中患病情况

接着用柱状图可以统计数据集中，分别在两种情况下（患糖尿病与不换糖尿病），六个特征（怀孕次数、葡萄糖浓度、舒张压、皮褶厚度、体重指数、年龄）的统计人数，由图 2 可以发现：

1、 女性在怀孕超过 6 次以后患糖尿病的概率更大；
2、 糖尿病患者的葡萄糖含量、血压、皮肤厚度、BMI 四个特征数值明显比不患病的人高，且糖尿病患者的的皮肤厚度数值大多集中在 36 左右，而不会患病的人皮肤厚度大多集中在 34；
3、 年龄在 40-60 岁之间患糖尿病的比率较高

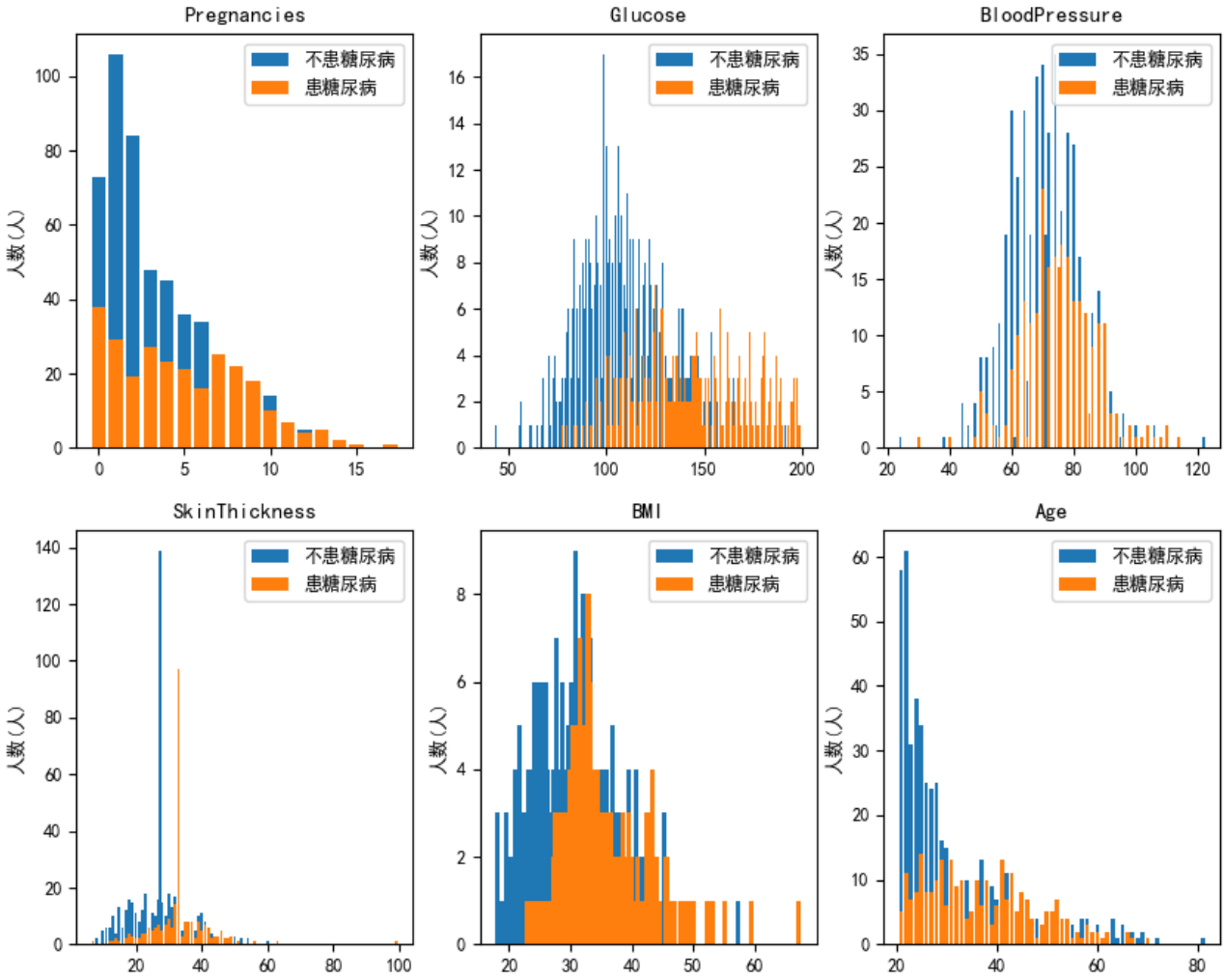

图 2：六个特征分别在两个情况下的统计人数

以均值作为纵坐标，特征类别作为横坐标，分别画出患病与不患病的折线，根据图 3 初步预测与患病最有直接关联的特征因素为胰岛素含量。

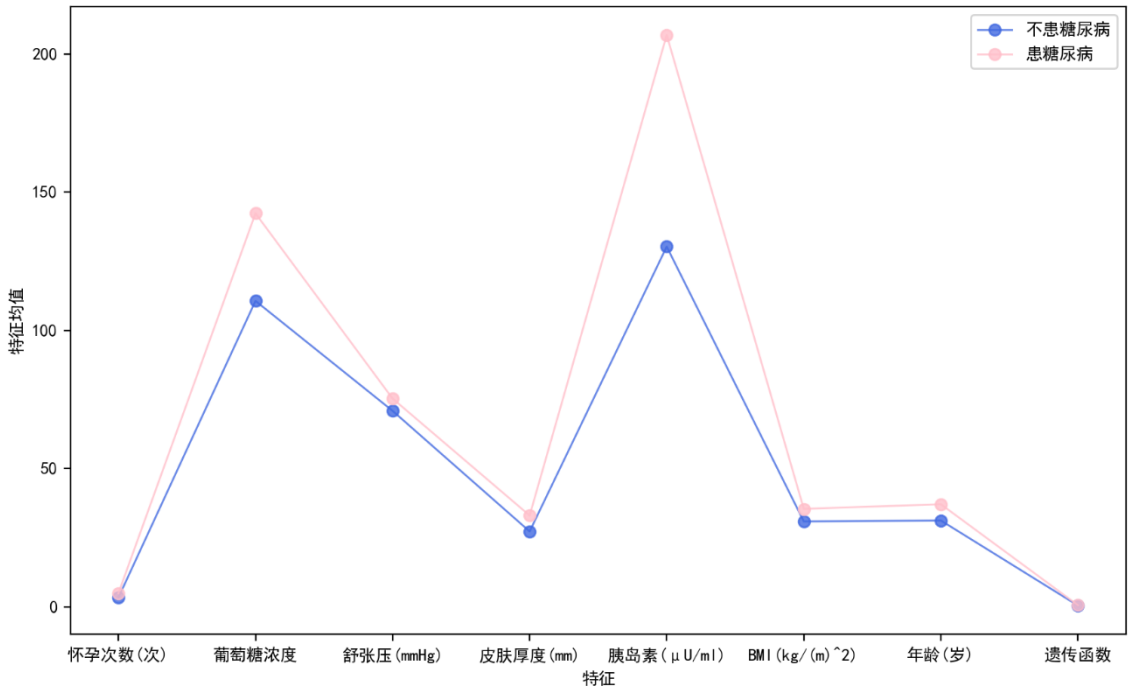

图 3：各特征分别在两个情况下（患病与没患病）的均值

除此之外，还可以画出标准热力图，观察各个特征之间的情况。

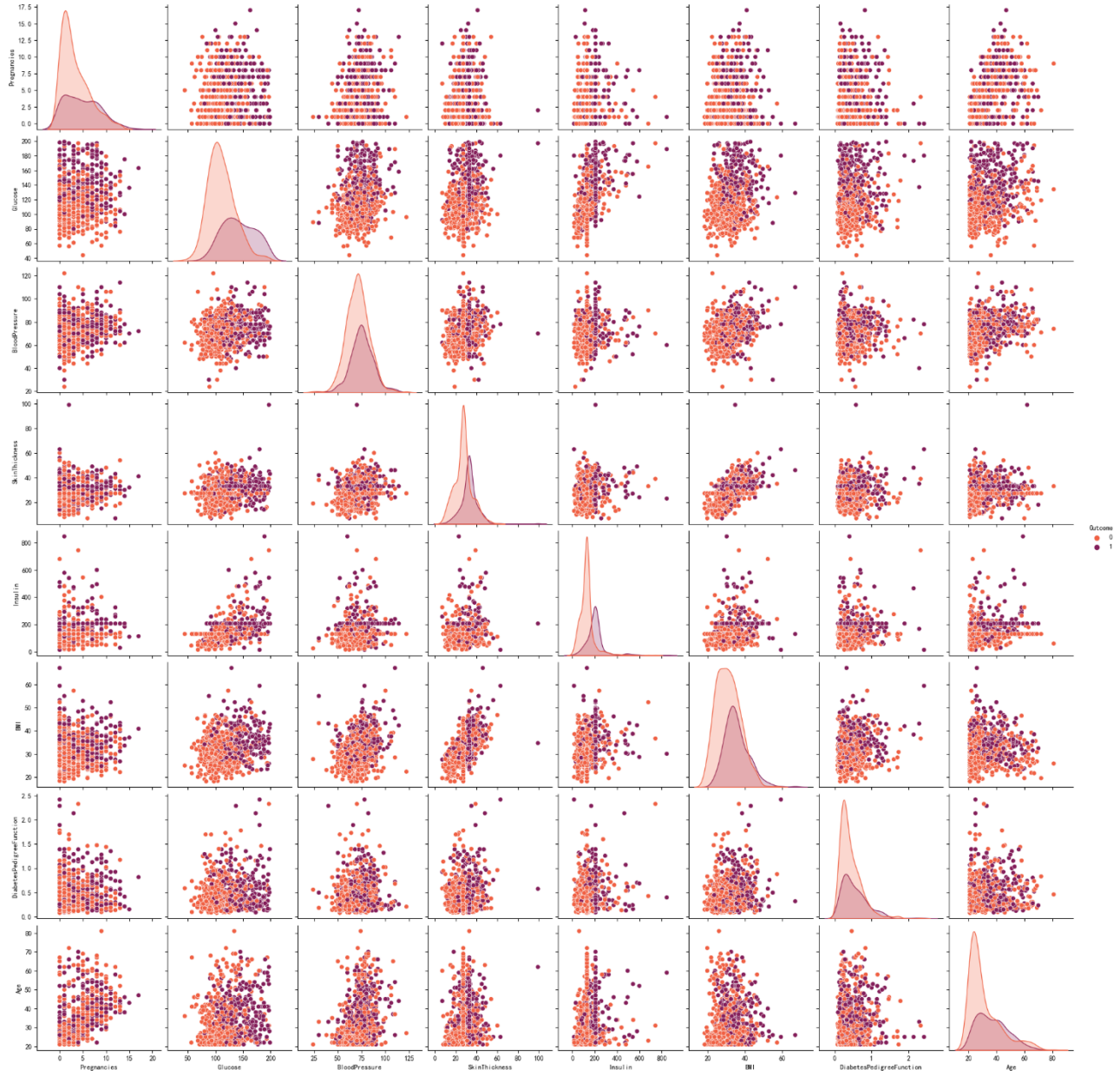

图 4：标准热力图

2.2.2 提出假设

根据以上分析，可以做出三个预测：

1、 女性在怀孕超过 6 次后，患糖尿病概率更大

2、 年龄在 40-60 岁的女性更易患糖尿病

3、 平均葡萄糖浓度、平均舒张压、平均皮褶厚度、平均血清胰岛素、平均体重指数更高的女性更容易患糖尿病

2.3 特征工程

2.3.1 特征选取

为减少后续模型训练时间，并且能够增强模型的泛化能力，需要进行特征选取，采用相关系数法，可以利用热力图来观察各个特征之间的特征相关性，通过图 5 可得，相关性较强的为葡萄糖浓度、胰岛素含量、体重指数，相关系数分别为 0.5，0.41，0.32。

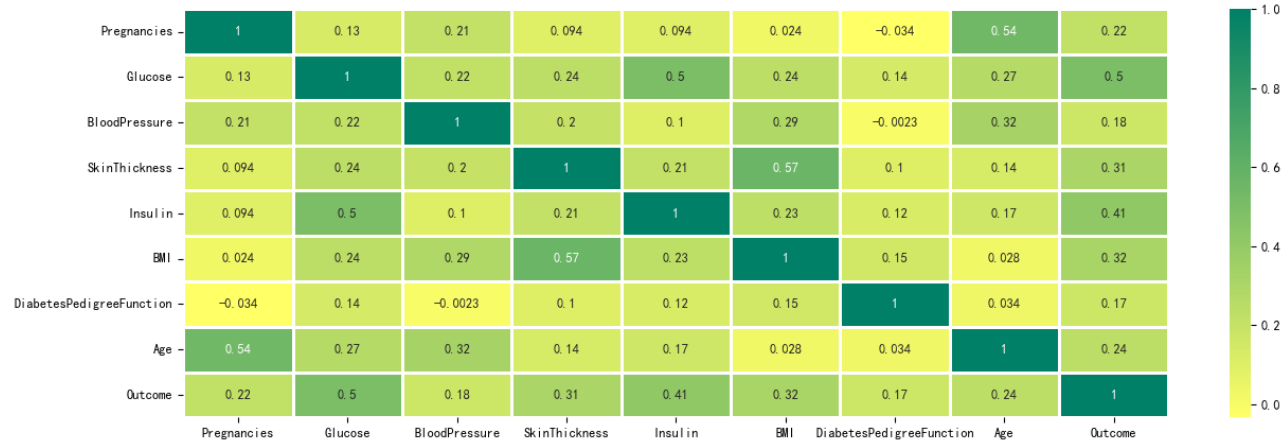

图 5：相关系数热力图

2.4 数据标准化

2.4.1 数据标准化

为了后续让数据有可比性，需要进行数据标准化，使用 StandardScaler 模型进行标准化，标准化后数据如图 6。

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	0.639947	0.864020	-0.035389	0.645088	0.561272	0.167806	0.468492	1.425995
1	-0.844885	-1.205478	-0.531332	-0.027701	-0.300842	-0.850452	-0.365061	-0.190672
2	1.233880	2.013741	-0.696647	0.420825	0.561272	-1.330487	0.604397	-0.105584
3	-0.844885	-1.074081	-0.531332	-0.700491	-0.709475	-0.632253	-0.920763	-1.041549
4	-1.141852	0.502679	-2.680419	0.645088	0.123830	1.549727	5.484909	-0.020496

图 6：标准化后数据

3 模型构建

3.1 分类算法

分类算法是十分普遍的一种技术。它简单，易于理解。可以通过原始数据集中的特征属性来找到潜在的分类规则，从而对新的数据集快速的进行有效的分类。

在糖尿病预测的原始数据集分析中，进行的主要是二分类的预测问题。将是否患有糖尿病这一自变量作为分类属性，其他自变量作为条件属性，将各项数据划分到不同的子集中。还可以通过特征的选取来发现致病的主要因素。

3.2 数据集分类

将原始数据集以 4:1 的比例分成训练集 train 和测试集 test 训练集用于模型的训练，测试集用于结果的测试。最终训练集包含 614 条数据，测试集包含 154 条数据。

3.3 逻辑回归算法模型

3.3.1 模型简介

逻辑回归算法是经典的分类方法，实际上是在用线性回归模型的预测结果去逼近真实标记的对数几率。它是一种广义的线性回归分析模型，属于机器学习中的监督学习。其推导过程与计算方式类似于回归的过程，但实际上主要是用来解决二分类问题（也可以解决多分类问题）。分类的主要思想是：根据现有数据

对分类边界线建立回归公式，以此进行分类。

3.3.2 利用逻辑回归算法训练模型

在 Python 中，调用 LogisticRegression()和 fit()函数训练模型，最后查看各个参数的权重。得出来的权重： [[0.28520852 0.8902513 -0.11502468 0.30043912 0.54487822 0.40286415 0.25176905 0.19159546]]。

3.3.3 混淆模型

混淆矩阵也称误差矩阵，是表示精度评价的一种标准格式，用 n 行 n 列的矩阵形式来表示。在本文中，混淆模型的结果为，154 名皮马印第安女性中正确识别为未患糖尿病的正常人数量为 95 名，正确识别为患糖尿病的糖尿病患者数量为 29 名，识别为未患糖尿病的糖尿病患者为 12 名，识别为糖尿病的正常人为 18 名。准确率为 80.52%。

3.4 决策树模型

3.4.1 模型简介

在多层次或多阶段的决策中，当一个阶段的决策完成之后，可能会产生新的不同自然状态，每种自然状态又会产生不同的策略选择。决策树是一种能帮助决策者进行序列决策分析的有效工具，其方法是将问题中有关策略、自然状态、概率及收益值等通过线条和图形用类似于树状的形式表示出来。

3.4.2 利用决策树训练模型

在 Python 中，调用 DecisionTreeClassifier()和 fit()函数训练模型。最后得出准确率为 90.26%。

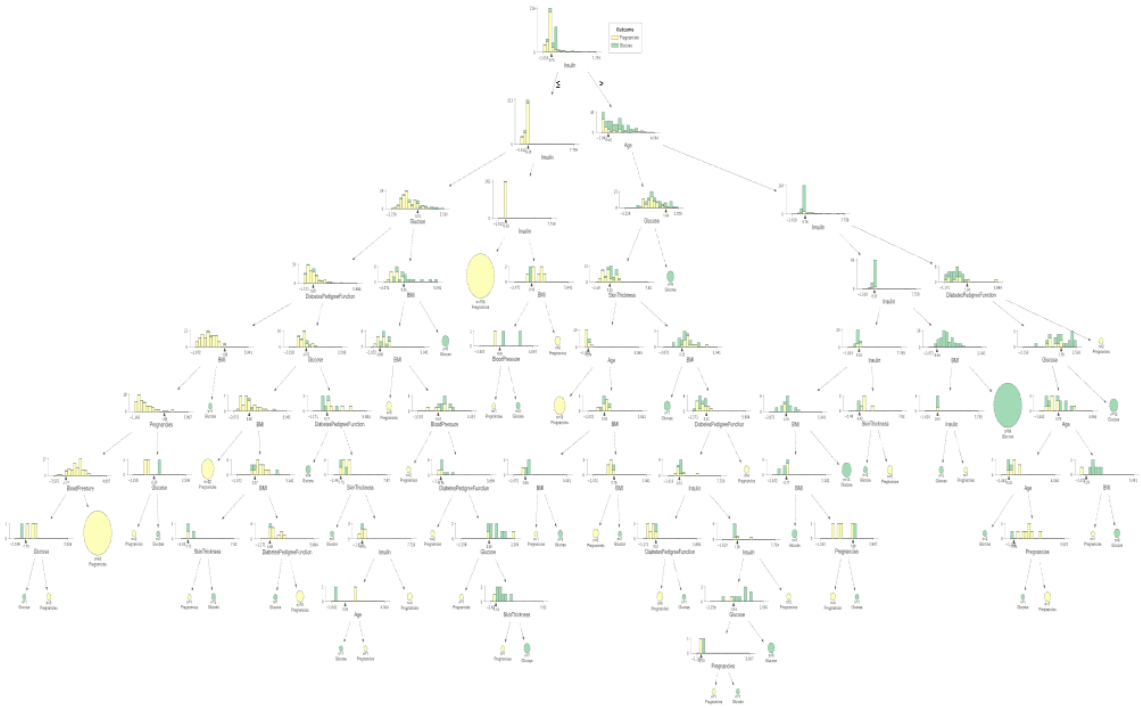

图 7：决策树可视化

4 结论

本文针对年龄大于等于 21 岁的皮马印第安女性是否患有糖尿病训练出合理的预测模型。利用了多种数据预处理技术对数据进行清洗。对清洗后的数据进行初步查看分析。将原始数据集分成训练集和测试集来进行后续模型训练。本次预测实验中使用了逻辑回归模型和决策树模型来训练数据，最终结果逻辑回归模型准确率为 80.52%，决策树模型准确率为 90.26%。决策树模型训练的效果优于逻辑回归模型。

最后的出结论：768 人中，有 268 人患病，500 人不患病，患病率为 34.90%。平均葡萄糖浓度、平均

舒张压、平均皮褶厚度、平均血清胰岛素、平均体重指数数值较高更容易患糖尿病。在 40-60 岁之间或者怀孕次数超过 6 次的女性更容易患糖尿病。

本次糖尿病预测项目，使用的数据集具有一定的局限性。若要更准确的提高预测的准确率，可以扩大数据集的范围，进一步扩大各年龄段和各个人种以及性别的条件，对各个特征进行进一步分析。

Diabetes prediction project

Zhao Siqi [1], Qiu Biyang [1, 2]

(1. *Shenzhen Technology University , Shenzhen,518000)*

(2. *Shenzhen Technology University , Shenzhen,518000)*

Abstract: The project aims to determine whether a person has diabetes by age, body mass index, diastolic blood pressure and other parameters. Python was used for visual analysis of the relationship between various medical data and diabetes. Use machine learning tools for inferential analysis. The data was standardized, the logistic regression algorithm model and the decision tree model based on information entropy were used to predict the test set, and finally the confusion model and accuracy were used to evaluate the results. It is concluded that plasma glucose concentration, triceps skin fold thickness, serum insulin concentration and body mass index are strongly correlated with diabetes mellitus.

Keywords: Diabetes; Logistic regression model; Decision tree model