基于糖尿病数据集的决策树分类

简介

diabetes.csv 最初来自国家糖尿病和消化和肾脏疾病研究所。该数据集的目标是根据数据集中包括的某些诊断测量来进行诊断预测，预测患者是否患有糖尿病。这些实例的选择受到若干约束，特别是这里的所有患者都是至少 21 岁的皮马印第安传统的女性。从 diabetes.csv 中您可以找到几个变量，其中一些是独立的（几个医学预测变量），并且只有一个目标依赖变量（结果）。

指示

1. 使用 Copilot 聊天来在您的项目中创建一个新笔记本。使用命令 /newnotebook 并将其命名为 "Diabetes Tree Classifier"。
2. 使用 Copilot 和 Copilot 聊天来制定练习并支持您的学习。

练习

该项目的目标是基于 Scikit-learn 和 Python 构建一个决策树分类器。该分类器应能够基于数据集中包括的某些诊断测量来预测患者是否患有糖尿病。

1. 导入所需的库来构建决策树分类器

2. 加载数据集

3. 探索性数据分析

3.1. 显示数据框的前 5 行。
3.2. 显示数据框中的行数和列数。
3.3. 显示每列的数据类型。
3.4. 显示每列中的缺失值数量。
3.5. 显示每列中的唯一值数量。

4. 特征选择

4.1. 将数据拆分为特征和目标变量。
4.2. 将数据拆分为训练集和测试集。

5. 构建决策树分类器

5.1. 实例化 DecisionTreeClassifier 类。
5.2. 将模型拟合到训练数据。
5.3. 预测测试数据的标签。
5.4. 评估模型。

6. 可视化决策树

7. 结论