皮马印第安人糖尿病预测模型

第三组：贺轶凡、李洁琼、施呈韵、王秋成、肖彤、张小娅

选题背景与概述

美国糖尿病及消化疾病研究所曾针对21岁以上的皮马印第安女性，在初检中获取与糖尿病相关的8项特征数据，并在5年内对初检未患病样本进行复诊，诊断是否后续患糖尿病：

数据来源：UCI

样本分类：糖尿病患者*（268个样本），非糖尿病患者（500个样本）

样本特征：

1. 怀孕次数
2. 葡萄糖耐量
3. 舒张压
4. 三头肌皮褶厚度
5. 血胰岛素浓度
6. 体重指数
7. 遗传指数
8. 年龄

模型选择：决策树（5层），精确度：79.5%

模型成果：根据样本特征判定其是否会在未来患糖尿病




*该研究对糖尿病确诊依据为葡萄糖耐量测试，口服葡萄糖后2小时血糖浓度大于200mg/dl

原始数据样例

样本编号	怀孕次数	葡萄糖耐量 (mg/dl)	舒张压 (mm Hg)	三头肌皮褶厚度 (mm)	血胰岛素浓度 (mu U/ml)	体重指数	遗传指数	年龄	是否糖尿病
1	3	126	88	41	235	39.3	0.704	27	0
2	11	143	94	33	146	36.6	0.254	51	1
3	10	125	70	26	115	31.1	0.205	41	1
4	1	97	66	15	140	23.2	0.487	22	0
5	13	145	82	19	110	22.2	0.245	57	0
6	3	158	76	36	245	31.6	0.851	28	1
7	3	88	58	11	54	24.8	0.267	22	0
8	4	103	60	33	192	24	0.966	33	0
9	4	111	72	47	207	37.1	1.39	56	1
10	3	180	64	25	70	34	0.271	26	0
11	9	171	110	24	240	45.4	0.721	54	1
12	1	103	80	11	82	19.4	0.491	22	0
13	1	101	50	15	36	24.2	0.526	26	0
14	5	88	66	21	23	24.4	0.342	30	0
15	8	176	90	34	300	33.7	0.467	58	1

特征相关性

特征相关系数	怀孕次数	葡萄糖耐量	舒张压	三头肌皮褶厚度	血胰岛素浓度	体重指数	遗传指数	年龄
怀孕次数	1.00	0.13	0.14	-0.08	-0.07	0.02	-0.03	0.54
葡萄糖耐量	0.13	1.00	0.15	0.06	0.33	0.22	0.14	0.26
舒张压	0.14	0.15	1.00	0.21	0.09	0.28	0.04	0.24
三头肌皮褶厚度	-0.08	0.06	0.21	1.00	0.44	0.39	0.18	-0.11
血胰岛素浓度	-0.07	0.33	0.09	0.44	1.00	0.20	0.19	-0.04
体重指数	0.02	0.22	0.28	0.39	0.20	1.00	0.14	0.04
遗传指数	-0.03	0.14	0.04	0.18	0.19	0.14	1.00	0.03
年龄	0.54	0.26	0.24	-0.11	-0.04	0.04	0.03	1.00

代码实现

分离特征与标签：

```
def tokenize(item):
    vector = Vectors.dense(float(item[0]),float(item[1]),float(item[2]),float(item[3]),
             float(item[4]),float(item[5]),float(item[6]),float(item[7]))
    if item[8] == '0':
        label = 0.0
    else:
        label = 1.0

    item = LabeledPoint(label,vector)
    return item
```

训练样本（622个）与验证样本（146个）分离：

```
training, testing = diabete_points.randomSplit([0.8, 0.2], 11)
```

决策树模型：最大深度5层

```
model = DecisionTree.trainClassifier(training, numClasses = 2, maxDepth = 5, maxBins = 32, categoricalFeaturesInfo={})
```

决策树模型

```
If (feature 1 <= 155.0)
   If (feature 5 <= 26.2)
    If (feature 1 <= 124.0)
     If (feature 6 <= 0.661)
      Predict: 0.0
     Else (feature 6 > 0.661)
      If (feature 6 <= 0.695)
       Predict: 0.0
      Else (feature 6 > 0.695)
       Predict: 0.0
    Else (feature 1 > 124.0)
     If (feature 2 <= 56.0)
      Predict: 1.0
     Else (feature 2 > 56.0)
      If (feature 7 <= 39.0)
       Predict: 0.0
.....
```

feature 0: 怀孕次数

feature 1: 葡萄糖耐量测试

feature 2: 舒张压

feature 3: 三头肌皮褶厚度

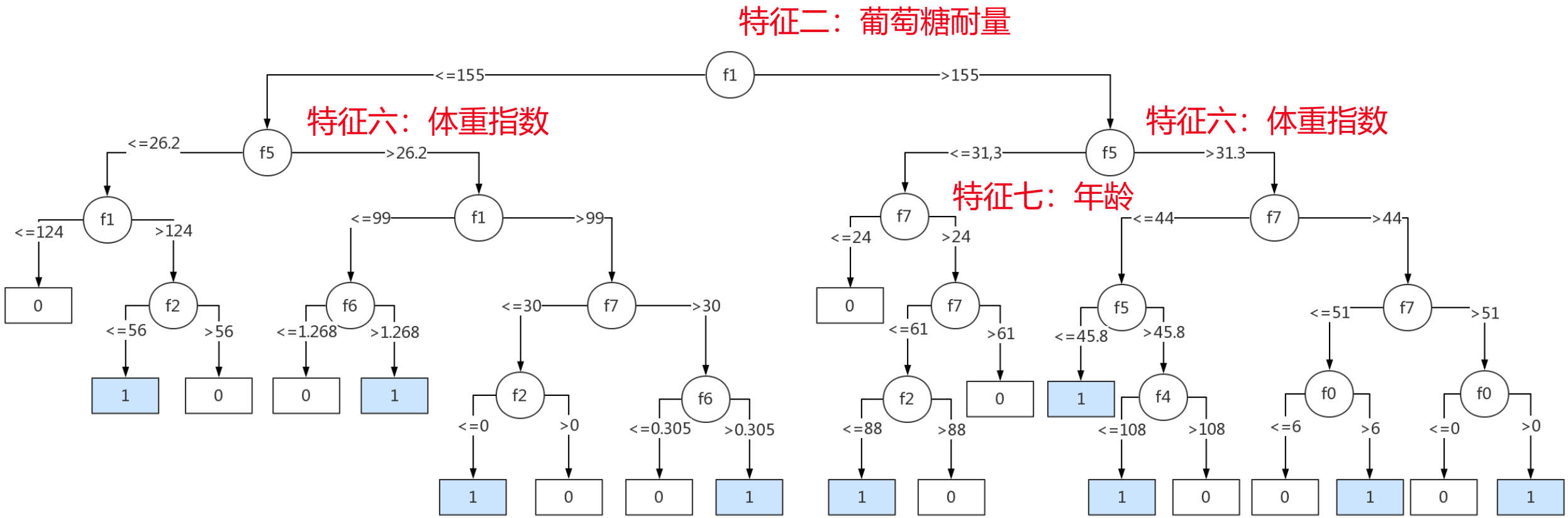
feature 4: 血胰岛素

feature 5: 体重指数

feature 6: 遗传指数

feature 7: 年龄

决策树模型


特征二：葡萄糖耐量
特征六：体重指数
特征六：体重指数
特征七：年龄
f1
<=155
>155
f5
<=26.2
>26.2
f5
<=31,3
>31.3
f1
<=124
>124
f1
<=99
>99
f7
<=24
>24
f7
<=44
>44
0
f2
f6
<=30
f7
>30
0
f7
<=61
>61
f5
<=45.8
>45.8
f7
<=51
>51
<=56
>56
<=1.268
>1.268
f2
<=0
>0
f6
<=0.305
>0.305
f2
<=88
>88
1
f4
<=108
>108
f0
<=6
>6
f0
<=0
>0
1
0
0
1
1
0
0
1
1
0
1
0
1
0
0
1
1
0
1
0
0
1

研究结果

验证样本（146）

模型精准率：**79.5%**

混淆矩阵

29	24	+
6	87	−
+	−	

模型判定　　实际结果

- 决策树模型对验证数据判定精准率达到79.5%，可在增加样本量之后可进一步提升模型精准率；
- 模型所需特征均为可量化客观特征，且容易获取，模型执行成本低
- 基于混淆矩阵结果，模型的误诊率较低，仅6例被模型判定将会患糖尿病的样本实际未患病，因此对于被该模型判定患糖尿病的新例很有可能在未来患糖尿病，需要开始进行干预；
- 基于决策树的分层结果，较上层的特征，除了葡萄糖耐量，体重指数与年龄也是影响力较大的特征；
- 样本来自21岁以上皮马印第安女性，因此模型应用于其他人群，如中国女性，需重新输入样本数据，生成决策树，作为预判患糖尿病的有效模型，对潜在病患进行及时干预。

谢谢！

皮马印第安人糖尿病预测模型

第三组：贺轶凡、李洁琼、施呈韵、王秋成、肖彤、张小娅