常见科室及医疗疾病诊断预测

印第安人糖尿病数据集：

属性名称	描述
妊娠	妊娠次数
葡萄糖	口服糖耐量试验 2 小时的血糖浓度
血压	舒张压（mm Hg）
皮肤厚度	三头肌皮褶厚度（mm）
胰岛素	2 小时血清胰岛素（uU/ml）
BMI	体重指数
糖尿病谱系功能	糖尿病谱系函数
年龄	年龄（年）
诊断	0 或 1（0：无糖尿病，1：糖尿病）

可视化数据集直方图：

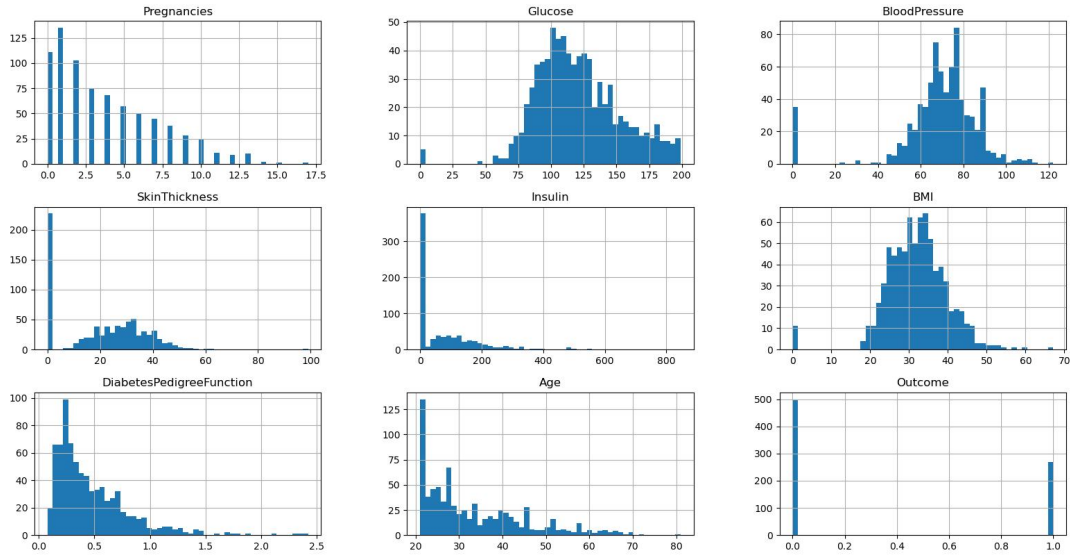


图 1 各个属性的相关数据显示

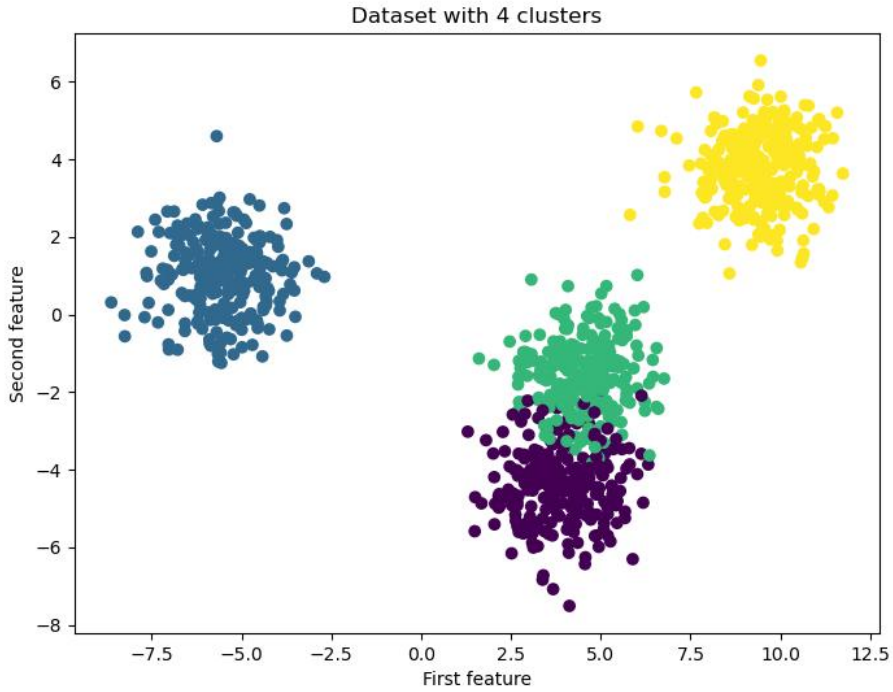


图 2 数据分类

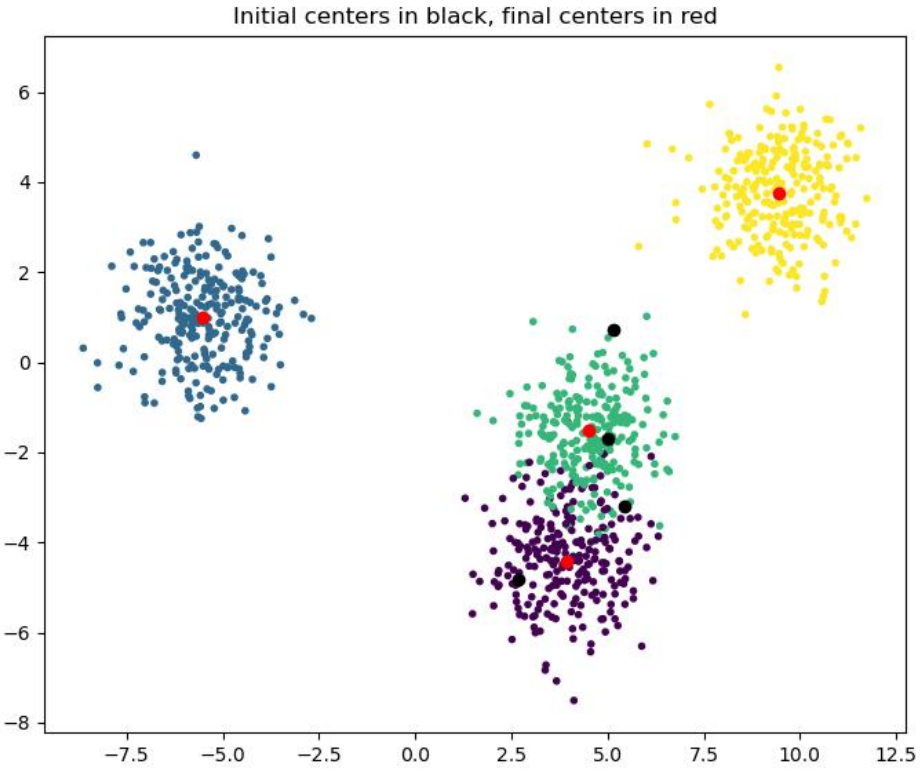


图 3 Kmeans 聚类方法

另一数据来源：通过网页爬取和人工筛选的方式获得寻医问药网、好大夫经典问诊网页、39 健康网数据集以及 Chinese medical dialogue data 开源数据网获得的中

文医疗对话数据集，其中好大夫网数据集更真实可靠。（好大夫拒绝使用 python 爬虫，数据造假，这里使用软件采集网页的各科病例的症状数据一栏）
开源数据网站：GitHub - Toyhom/Chinese-medical-dialogue-data: Chinese medical dialogue data 中文医疗对话数据集
好大夫网站：https://www.haodf.com/bingcheng/list-xinxueguanneike.html
39 健康网症状收集：39 健康网_优质医疗保健信息与在线健康服务平台

(1)开源数据集的截取:

label	title	ask	answer
心血管科	高血压患	我有高血	高血压病人可以口服党参的。党参有降血脂，降血压的作用，可以彻底消除血
内分泌科	糖尿病还	糖尿病有	2型糖尿病的隔代遗传概率为父母患糖尿病，临产的发生率为40%，比一般人患
消化科	哪家医院	烧心，打	建议你用奥美拉唑同时，加用吗丁啉或莫沙必利或援生力维，另外还可以加用
呼吸科	感冒咳嗽	妹妹昨天	感冒初期如果是有咳发怕风怕冷等许多症多考虑是咽喉炎症上呼吸道感风寒咳
血液科	求问特发	到底怎么	你好，特发性血小板减少性紫癜一般是原因搞不清楚影响致使自身免疫功能障

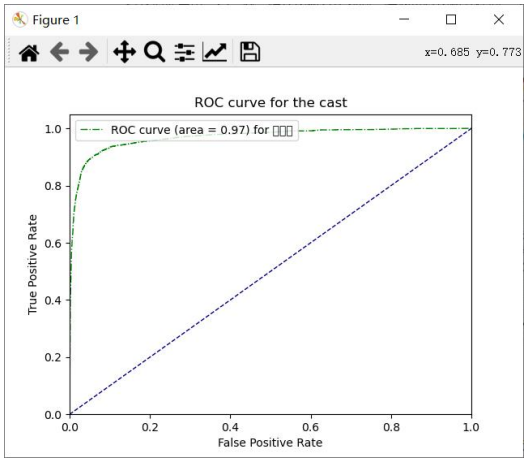
(2)好大夫问诊截取：
0-4 对应神经内科、呼吸科、消化内科、内分泌科、心血管科

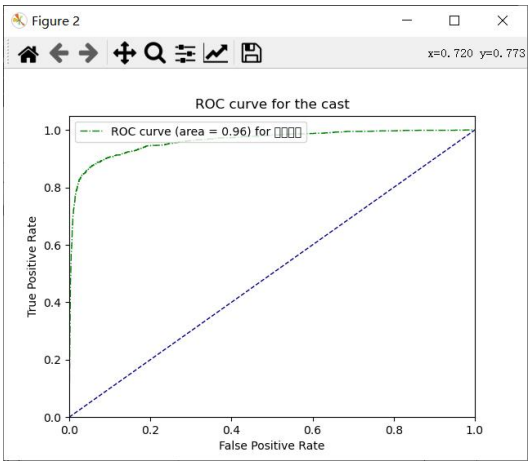
label	question
心血管内	就是家族高血压史，容易焦虑，有过惊恐发作的痕迹，110 150 甚至更高，就是手心出汗，
消化内科	明天下午1点半请您做胃肠镜，请问怎么吃药？紧急，今天晚上就要吃磷酸钠盐口服液吗，非
神经内科	患者于2018年前高血压脑出血后经保守治疗后遗留左侧偏身麻木无力，伴左侧全肢体疼痛，2
内分泌科	甲状腺肿大，去医院医生开了测血和彩超，现在结果出来了自己看不懂（2023-03-28填写）
呼吸科	复方新诺明和米诺环素治疗奴卡菌感染一个月。但明显感觉到头晕。不知道哪种药物引起的。

(3)具体科室预测指标：

具体科室指标%	精确率 P	召回率 R	F1
呼吸科	82.49	87.22	84.79
消化内科	88.63	90.10	89.36
内分泌科	84.36	84.68	84.52
心血管科	86.80	83.38	85.06
肾内科	88.15	79.59	83.65

绘制 ROC 曲线：
①标准数据集以 0 呼吸科和 1 消化内科为示例：





②好大夫截取数据集以呼吸科和内分泌科为例：；

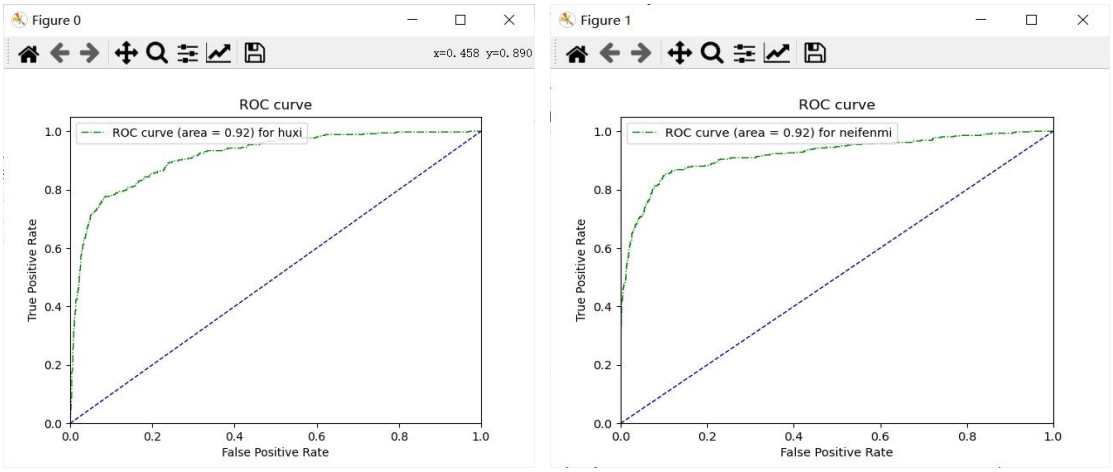


图 4 测试各个科室的预测 ROC 曲线绘制

部分 python 运行截图：
未添加 TD-IDF

```
Epoch 12/15
1448/1448 [==============================] - 106s 73ms/step - loss: 0.0928 - accuracy: 0.9679
Epoch 13/15
1448/1448 [==============================] - 107s 74ms/step - loss: 0.0847 - accuracy: 0.9699
Epoch 14/15
1448/1448 [==============================] - 108s 75ms/step - loss: 0.0781 - accuracy: 0.9719
Epoch 15/15
1448/1448 [==============================] - 110s 76ms/step - loss: 0.0750 - accuracy: 0.9727
161/161 - 1s - loss: 0.9425 - accuracy: 0.8538 - 1s/epoch - 6ms/step
322/322 [==============================] - 1s 3ms/step
```

添加人工词典库：

```
Epoch 12/15
1287/1287 [==============================] - 97s 76ms/step - loss: 0.0907 - accuracy: 0.9681
Epoch 13/15
1287/1287 [==============================] - 98s 76ms/step - loss: 0.0842 - accuracy: 0.9705
Epoch 14/15
1287/1287 [==============================] - 97s 75ms/step - loss: 0.0812 - accuracy: 0.9716
Epoch 15/15
1287/1287 [==============================] - 95s 74ms/step - loss: 0.0730 - accuracy: 0.9736
322/322 - 2s - loss: 1.0137 - accuracy: 0.8428 - 2s/epoch - 5ms/step
644/644 [==============================] - 2s 3ms/step
```

添加 TF-IDF 筛选词频：

```
Epoch 12/15
1448/1448 [==============================] - 120s 83ms/step - loss: 0.0824 - accuracy: 0.9701
Epoch 13/15
1448/1448 [==============================] - 121s 84ms/step - loss: 0.0762 - accuracy: 0.9726
Epoch 14/15
1448/1448 [==============================] - 119s 82ms/step - loss: 0.0724 - accuracy: 0.9735
Epoch 15/15
1448/1448 [==============================] - 115s 79ms/step - loss: 0.0683 - accuracy: 0.9748
161/161 - 1s - loss: 0.9726 - accuracy: 0.8635 - 1s/epoch - 7ms/step
322/322 [==============================] - 1s 3ms/step
```

使用 skip-gram 模型：

```
Epoch 9/15
1448/1448 [==============================] - 116s 80ms/step - loss: 0.1016 - accuracy: 0.9642
Epoch 10/15
1448/1448 [==============================] - 115s 80ms/step - loss: 0.0917 - accuracy: 0.9676
Epoch 11/15
1448/1448 [==============================] - 117s 81ms/step - loss: 0.0848 - accuracy: 0.9698
Epoch 12/15
1448/1448 [==============================] - 118s 81ms/step - loss: 0.0801 - accuracy: 0.9713
Epoch 13/15
1448/1448 [==============================] - 117s 81ms/step - loss: 0.0748 - accuracy: 0.9727
Epoch 14/15
1448/1448 [==============================] - 118s 81ms/step - loss: 0.0703 - accuracy: 0.9744
Epoch 15/15
1448/1448 [==============================] - 117s 81ms/step - loss: 0.0654 - accuracy: 0.9756
161/161 - 1s - loss: 0.9800 - accuracy: 0.8678 - 1s/epoch - 7ms/step
322/322 [==============================] - 1s 3ms/step
```

常见的分类方法：

（1）SVM 分类

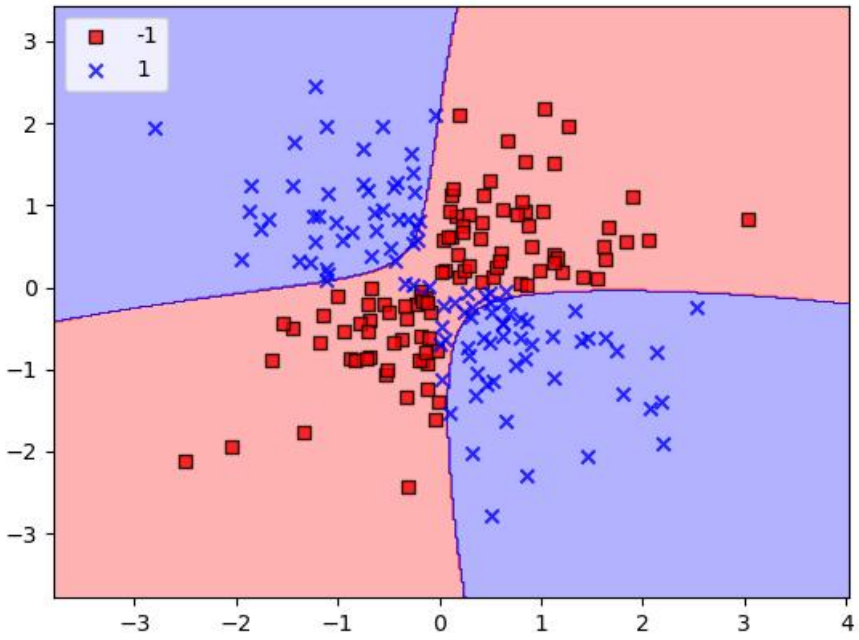


图 5 使用非线性 Kernel-SVM

（2）决策树分类

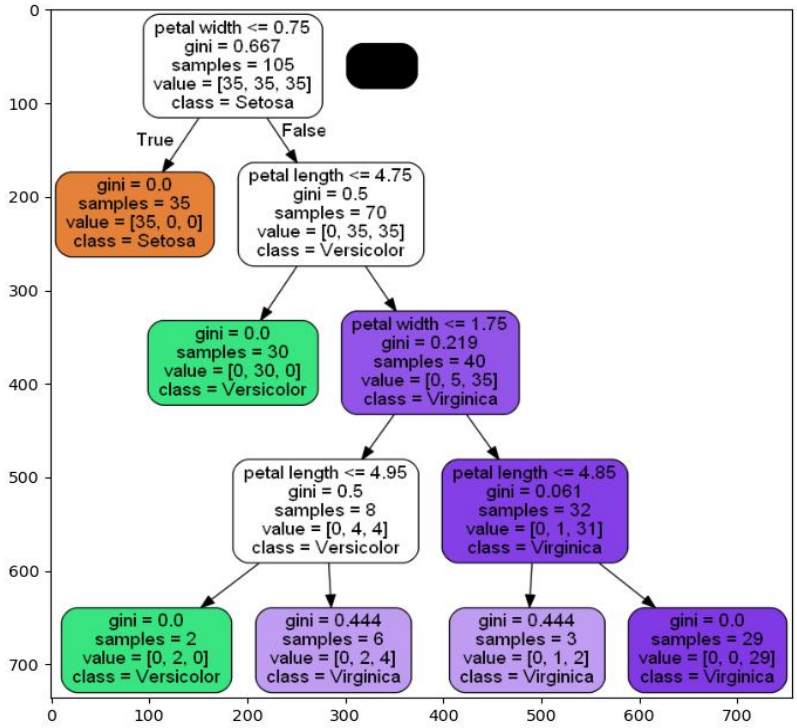


图 6 构建目标决策树

目前加密方案 CP-ABE-PRE 加密方案：

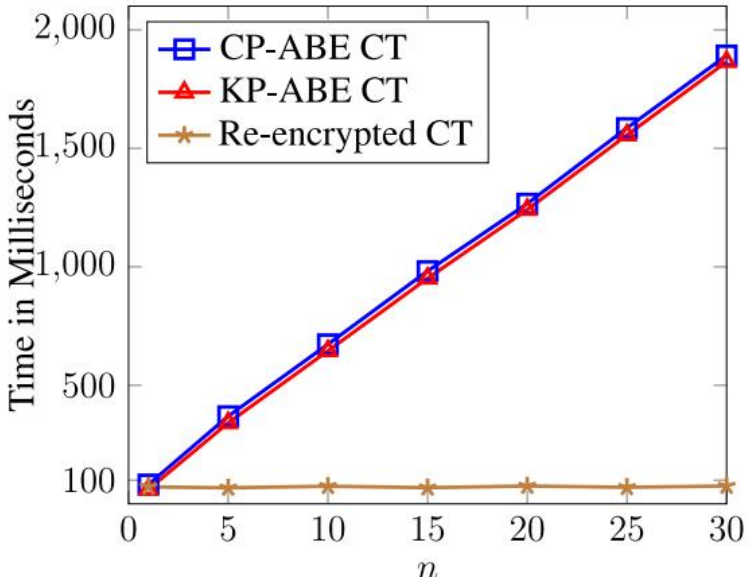


图 7 ABE 密文和重新加密密文的解密执行时间。