数据挖掘导论lab1报告

3220106424 张天逸

数据集来源

我选择的数据集是Polycystic Ovary Syndrome PCOS（https://www.kaggle.com/datasets/lucass0s0/polycystic-ovary-syndrome-pcos/data）

这是一个 3 000 例的合成 PCOS 临床指标数据集，提供年龄、BMI、雄激素水平、月经状况和卵泡计数等 5 个核心特征，用于探索或建模 PCOS 诊断。

维度	说明
数据类型	*合成* 临床数据，基于 2003 年国际共识的 **Rotterdam 诊断标准** 随机生成
样本量	3 000 条记录，代表 18–44 岁育龄女性
目标列	`PCOS_Diagnosis`　0 = 未诊断 PCOS　1 = 符合 Rotterdam 至少 2 项标准
字段总览	• `Age`（整数，18–44） • `BMI`（体质指数，8.5–44.7） • `Menstrual_Irregularity`（二元，月经是否不规律） • `Testosterone_Level(ng/dL)`（总睾酮，20.5–136.4） • `Antral_Follicle_Count`（窦卵泡计数，3–39） • `PCOS_Diagnosis`（二元，诊断结果）

部署

依赖：pandas numpy scikit-learn matplotlib scipy

如果存在python版本冲突，建议创建虚拟环境：

cd lab1 python -m venv .venv # 或 conda create -n pcos python=3.10 source .venv/bin/activate # Windows: .venv\Scripts\activate

预处理

预处理部分的代码可分为以下这些部分：

1. 去重 df = pd.read_csv(input_csv).drop_duplicates()读入 CSV 后立即删除完全重复的样本，避免同一人数据多次出现造成统计偏差。
2. 缺失值填补 df = df.fillna(df.median(numeric_only=True)) 对所有数值列用列中位数填补 NaN，保持分布不受极端值拉动。
3. 异常值截断（Winsorize） python
   for col in numeric_cols:
   q1,q3 = df[col].quantile([.25,.75])
   iqr = q3-q1

df[col] = df[col].clip(q1-1.5*iqr, q3+1.5*iqr) 对 4 个连续特征（Age、BMI、睾酮、AFC）按 1.5×IQR 规则截断上下尾，使极端值不再过度影响均值与模型梯度，同时长尾特征性仍保留。

4. 分位数离散化 python
   discretizer = KBinsDiscretizer(n_bins=n_bins, encode="ordinal", strategy="quantile")
   df[["BMI_bin","Testosterone_bin"]] = discretizer.fit_transform(df[["BMI","Testosterone_Level(ng/dL)"]]) 将 BMI 和 Testosterone 分成 n_bins (默认 4) 个等频箱，生成离散特征 *_bin。便于后续做卡方检验、信息增益或可视化。
5. 标准化 python
   scaler = StandardScaler()
   df[numeric_cols] = scaler.fit_transform(df[numeric_cols]) 让 Age、BMI、睾酮、AFC 均转化为 均值 0 / 方差 1，消除量纲差异，使距离度量与梯度下降更稳定。
6. 可选 PCA 降维 python
   if with_pca:
   pca = PCA(n_components=2, random_state=42)
   df[["PC1","PC2"]] = pca.fit_transform(df[numeric_cols]) 如 with_pca=True，使用 主成分分析 将 4 个数值特征投影到 2 维，便于可视化或降维建模；PC1、PC2 解释累计方差 > 95 %。
7. 保存结果 df.to_csv(output_csv, index=False) 输出为 new_pcos_processed.csv，供后续模型或可视化脚本直接读取。

可视化

可视化部分的代码可以分为以下这些部分：

1. 循环遍历每一数值列 for col in numeric: 逐列绘图，保持自动化
2. 新建画布 plt.figure() 避免多图叠加（每轮清空）
3. Boxplot plt.boxplot(df[col].dropna()) plt.title(f"Boxplot of {col}") plt.savefig(... "_box.png") 展示中位数、四分位距 (IQR) 与极端值；保存为 figs/<列名>_box.png
4. Histogram plt.hist(df[col].dropna(), bins=30) plt.title("Histogram ...") plt.savefig(... "_hist.png") 30 个柱子查看分布形态（偏态、峰度、多峰）；保存 _hist.png
5. QQ Plot stats.probplot(df[col].dropna(), dist="norm", plot=plt) plt.title("QQ Plot ...") plt.savefig(... "_qq.png") 把样本分位与正态分位比对，检验正态性；保存 _qq.png
6. 关闭画布 plt.close() 释放内存，确保下一轮不会把前面内容叠进来

分析结果

可视化图形揭示的主要特征

1.1 年龄 (Age)

- **直方图**近似钟形但轻度左偏——年轻样本略多。
- **QQ 图**几乎贴线 → 年龄分布与正态差异不大。
  → 数据集聚焦于 18–35 岁核心生育年龄段，年龄本身对 PCOS 诊断区分度有限。

1.2 BMI

- **盒须图**：上须显著长——高 BMI 离群值（> 30）集中在 PCOS 人群。
- **直方图**：右偏+长尾，即便 Winsorize 后仍明显。

- **QQ 图**：上端点远离直线说明右尾重。
  → 肥胖是 PCOS 共病的重要信号，清洗后仍保留了诊断信息。

1.3 睾酮 (Testosterone_Level)

- **所有三图**都显示浓重右左偏 → 高雄激素血症是 Rotterdam 三大诊断标准之一；数据合成过程很好地模拟了该临床特征。

1.4 窦卵泡计数 (AFC)

- **峰度高**（直方图尖且窄）、**QQ 图**上端明显上翘
  → 反映 PCOS 卵巢"多囊"影像学特征；与睾酮、月经失调共同组成高相关组合（corr ≈ 0.86 ~ 0.78）。

1.5 PCA 主成分 (PC1, PC2)

- 处理后 PC1、PC2 近似正态 → 说明标准化+PCA 消除了长尾，便于聚类或判别可视化。
- **散点图（若另外绘制）**展示健康 vs PCOS 在 PC1–PC2 平面已有粗分隔，可用于降维演示。

"有 vs 无" 预处理对可视化的影响

视角	原始数据	预处理后	差异
异常值	睾酮 > 250 ng/dL、BMI > 45、AFC > 50 等极端点淹没画面	Winsorize 截断至 1.5×IQR，盒须图更可读	极端样本权重下降，图形不再"散弹枪"
尺度	不同列量纲差异大 → PCA 前两维几乎只由 AFC 支配	StandardScaler 归一化后 PCA 方差更均衡	PC1/PC2 解释>95 % 总方差且近正态
直方图形态	极长右尾吞掉大部分区分度	尾巴被截但偏态仍在	易看出 BMI / Testosterone 幂律分布
QQ 图	上端远离红线到"折腰"状态	线性显著改善（PC1, PC2 近于直线）	统计检验更可信；仍可感知右偏
离散化列	无	生成 BMI_bin & Testosterone_bin（4 等频）	便于做分组箱线图 / 卡方检验

预处理 **减少了可视化的"噪音"**（极值、尺度不一），让箱线图和 QQ 图的形态更加易读，同时又 **保留了医学上有价值的右偏长尾**，在实践中能提高模型稳定性和解释性。