“DataGuard”数据安全报告

一、数据集分析

表格名称：医保_个人基本信息.xlsx，记录条数：53，属性个数：64。

- 准标识符名称：性别、民族、学历、邮政编码
- 隐私属性名称：无
- 效用属性名称：无

处理后属性个数：64，总体非缺失值占比：43.87%。

序号	属性名	属性类别	非缺失值占比	处理方式
1	统筹区号	其他	100.0	保留
2	联系人	其他	0.0	删除（行业规范）
3	姓名	其他	100.0	删除（行业规范）
4	户口所在地详细地址	其他	0.0	删除（全空列）
5	手机号码	其他	71.7	保留
6	居住证所在地地址	其他	0.0	删除（全空列）
7	常住地详细地址	其他	100.0	保留
8	家庭主要联系人(联系电话)	其他	3.8	保留
9	家庭地址	其他	0.0	删除（全空列）
10	家庭主要联系人(姓名)	其他	1.9	保留
11	身份证编码	其他	100.0	保留
12	异地行政区划（作为就医地时，为参...	其他	98.1	保留
13	居民人员行政区划	其他	96.2	保留
14	个人编号	其他	100.0	保留
15	性别	准标识符	100.0	K匿名
16	民族	准标识符	100.0	K匿名
17	出生日期	其他	100.0	删除（行业规范）
18	参加工作日期	其他	66.0	保留
19	户口性质	其他	100.0	保留
20	学历	准标识符	94.3	K匿名
21	用工形式	其他	86.8	保留
22	证件类型	其他	100.0	保留
23	离退休状态	其他	100.0	保留
24	国家/地区代码	其他	100.0	保留
25	个人管理码	其他	100.0	保留
26	资金来源标识(财政拨款/单位缴纳...	其他	0.0	删除（全空列）
27	医疗后延缴费标志	其他	0.0	删除（全空列）

28	医疗后延缴费终止时间	其他	0.0	删除（全空列）
29	医疗后延缴费开始时间	其他	0.0	删除（全空列）
30	养老后延缴费标志	其他	0.0	删除（全空列）
31	有效标志	其他	100.0	保留
32	放弃医保待遇	其他	0.0	删除（全空列）
33	精准扶贫人员标识	其他	0.0	删除（全空列）
34	内卡号	其他	90.6	保留
35	离退休日期	其他	41.5	保留
36	医疗人员类别	其他	100.0	保留
37	常住地行政区划代码	其他	98.1	保留
38	个人唯一识别码	其他	0.0	删除（全空列）
39	免缴类型	其他	1.9	保留
40	医疗异地行政区划	其他	1.9	保留
41	养老隶属关系	其他	1.9	保留
42	医保异地参保标识	其他	1.9	保留
43	死亡日期	其他	9.4	保留
44	邮政编码	准标识符	17.0	K匿名
45	灵活就业养老缴费档次	其他	3.8	保留
46	灵活就业缴费档次	其他	3.8	保留
47	是否属于金坛茅山老区(1-是,0...	其他	1.9	保留
48	家庭主要联系人(与参保人关系)	其他	1.9	保留
49	城乡医疗人员类别	其他	30.2	保留
50	武进个人编号	其他	37.7	保留
51	溧阳个人编号	其他	0.0	删除（全空列）
52	补偿金月数_武进	其他	11.3	保留
53	医疗退休类型	其他	30.2	保留
54	参保地性质(1户籍所在地参保,2...	其他	0.0	删除（全空列）
55	武进居民养老退休时间	其他	11.3	保留
56	医疗离退休日期	其他	32.1	保留
57	城乡养老缴费档次	其他	5.7	保留
58	本地或外地	其他	94.3	保留
59	通讯地址	其他	0.0	删除（全空列）
60	人员登记状态	其他	100.0	保留
61	城乡养老人员类别	其他	5.7	保留
62	原医疗系统个人编号	其他	30.2	保留
63	原企业养老系统个人编号	其他	26.4	保留
64	是否启动省卡	其他	98.1	保留

二、效果分析

匿名方法：K匿名

序号	参数值	效用指标			风险指标						权衡指标
		GIL	DM	CAVG	Ra	Rb	Rc	r_low	r_high	uni	tradeoff
1	原始	0	1	1	0.42	1.0	0.42	0.19	0.28	0.28	-
2	2.0	0.33	22.25	4.42	0.32	0.5	0.23	0.19	0.19	0.0	0.165
3	3.0	0.38	13.42	3.53	0.06	0.33	0.13	0.42	0.06	0.0	0.125
4	4.0	0.4	12.32	2.65	0.0	0.2	0.11	0.25	0.09	0.0	0.08
5	5.0	0.4	12.32	2.12	0.0	0.2	0.11	0.25	0.09	0.0	0.08
6	6.0	0.44	15.53	2.21	0.0	0.14	0.08	0.42	0.13	0.0	0.062
7	7.0	0.44	15.53	1.89	0.0	0.14	0.08	0.42	0.13	0.0	0.062
8	8.0	0.52	18.96	2.21	0.0	0.09	0.06	0.42	0.21	0.0	0.047

推荐取值：参数值 = 8.0。

1、效用指标：

- GIL（generalized information loss，泛化信息损失）通过量化已泛化的域值的比例，捕获泛化特定属性时产生的损失。GIL越小，表示效用损失越少。
- DM（discernibility metric，分辨力指标）通过给每个记录分配一个惩罚值，来衡量一个记录与其他记录的不可区分程度，惩罚值等于它所属的等价类的大小。DM越小，表示效用损失越少。
- CAVG（average equivalence class size metric，平均等价类指标）测量等价组的创建是否接近最佳情况。CAVG越小，表示效用损失越少。
- dist（distance，距离）衡量原始数据集和新数据集的平均欧式距离。dist越小，表示效用损失越少。
- acc（downstream prediction task accuracy，下游预测任务准确率）基于隐私处理后的数据集训练下游任务的机器学习模型，计算下游任务分类准确率（此指标的计算需指定效用属性）。acc越高，表示效用损失越少。

2、风险指标：

- Ra（lowest prosecutor risk，最低检察官风险）刻画重识别概率大于0.2的数据记录占总体的比例，Ra越小，表示风险越低。

- Rb（highest prosecutor risk，最高检察官风险）刻画数据集中最大的重识别概率，Rb越小，表示风险越低。
- Rc（average prosecutor risk ，平均检察官风险）刻画平均重识别概率，Rc越小，表示风险越低。
- low（records affected by lowest risk，最低风险影响的记录数）是达到最低风险的记录占总体的比例，在Rc较低的情况下low越大，表示风险越低。
- high（records affected by highest risk，最高风险影响的记录数）是达到最高风险的记录占总体的比例，在Rc较低的情况下high越小，表示风险越低。
- uni（sample uniques，唯一样本）是等价组大小为1的记录数量占总体的比例，uni越小，表示风险越低。
- priv（private attribute inference attack accuracy，隐私属性推断攻击准确率）模拟攻击者的角色，基于隐私处理后的数据集训练隐私属性预测的机器学习模型，计算隐私属性推断的准确率。priv越低，表示风险越低。

3、权衡指标：

- tradeoff（trade-off between utility and risk，风险与效用权衡指标）定义为：（1）对于数据匿名算法：效用指标GIL和风险指标Rb的乘积；（2）对于指定效用属性的其他算法：隐私指标priv乘以效用指标acc下降的值；（3）对于未指定效用属性的其他算法：隐私指标priv乘以效用指标dist。
- tradeoff越小，表示风险越低，同时效用损失越小。我们基于tradeoff指标自动为用户推荐合适的参数取值。