分析流程

数据源

data1单因素.sav

算法配置

算法：随机森林分类

分析结果

随机森林分类基于准确率、召回率、精确率、F1指标对模型进行评价，请看详细结论。

分析步骤

1. 通过训练集数据来建立随机森林分类模型。
2. 通过建立的随机森林来计算特征重要性。
3. 将建立的随机森林分类模型应用到训练、测试数据，得到模型的分类评估结果。
4. 由于随机森林中具有随机性，每次运算的结果不一样，若保存本次训练模型，后续可以直接上传数据代入到本次训练模型进行计算分类。
5. 注：随机森林无法像传统模型一样得到确定的方程，通常通过测试数据分类效果来对模型进行评价。

详细结论

输出结果1：模型参数

复制

参数名	参数值
训练用时	0.633s
数据切分	0.8
数据洗牌	是
交叉验证	5
节点分裂评价准则	gini
决策树数量	100
有放回采样	true
袋外数据测试	false
划分时考虑的最大特征比例	auto
内部节点分裂的最小样本数	2
叶子节点的最小样本数	1
叶子节点中样本的最小权重	0
树的最大深度	10

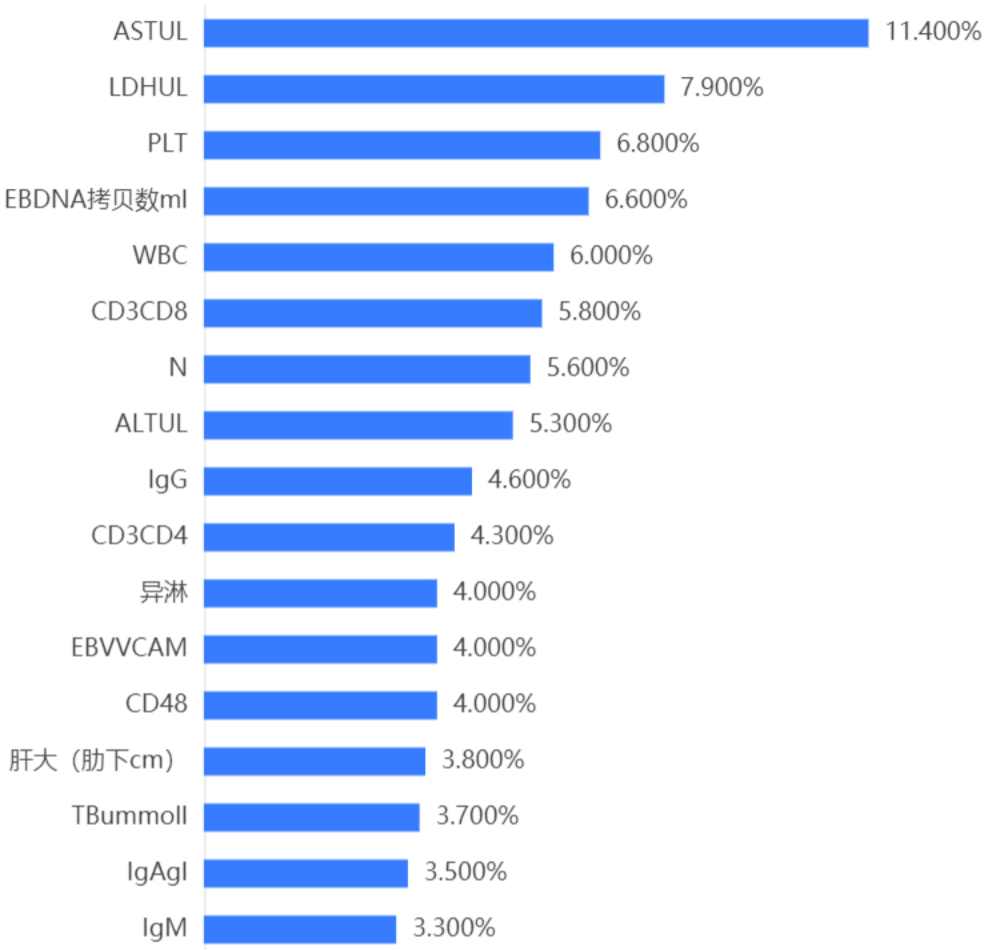
叶子节点的最大数量	50
节点划分不纯度的阀值	0

图表说明：

上表展示了模型各项参数配置以及模型训练时长。

输出结果2：特征重要性

柱形图

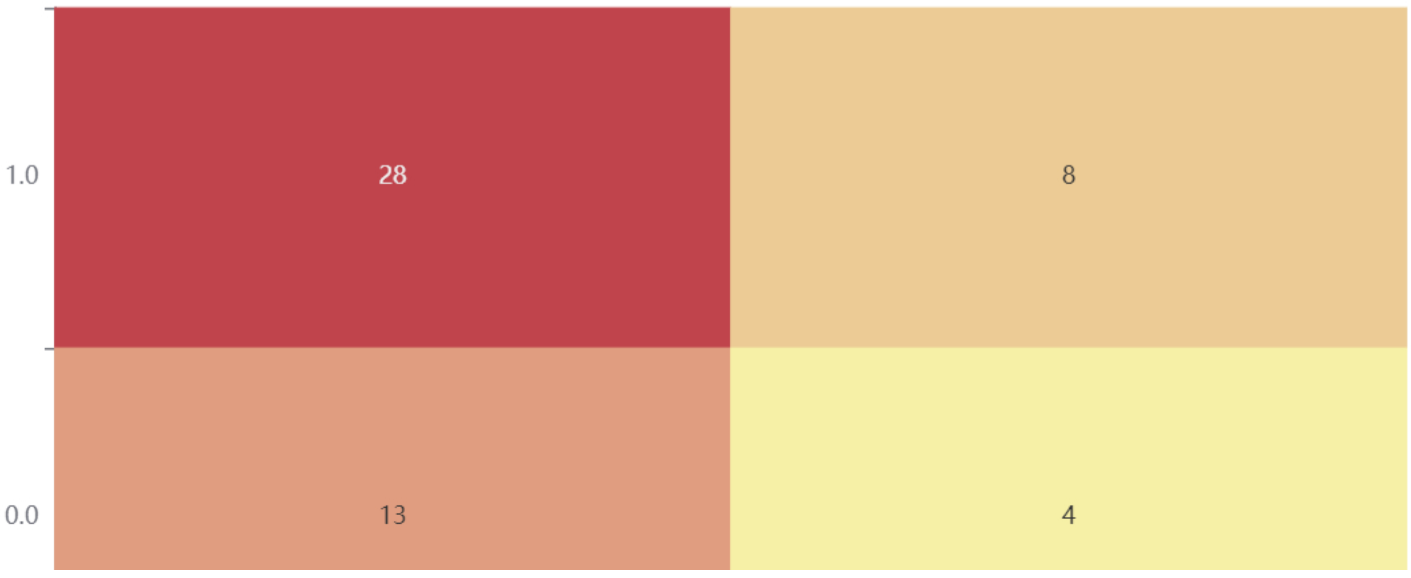


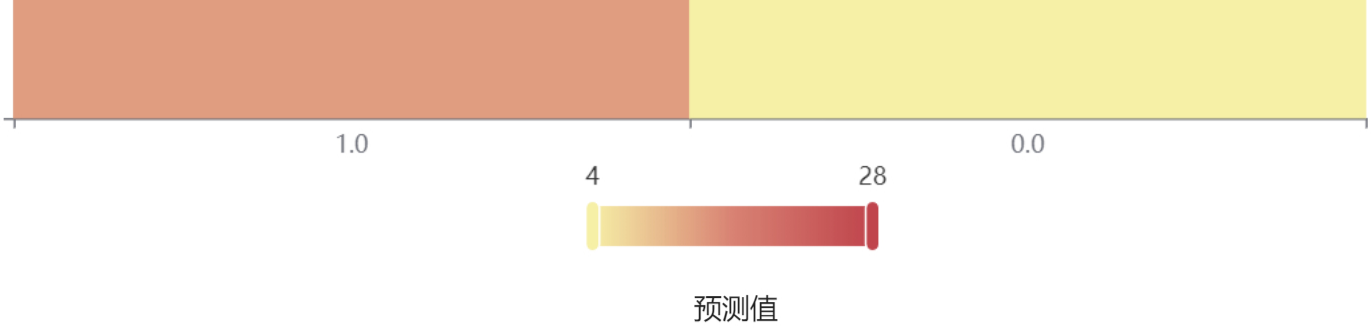
图表说明：

上柱形图或表格展示了各特征（自变量）的重要性比例。

输出结果3：混淆矩阵热力图

测试数据





图表说明：

上表以热力图的形式展示了混淆矩阵。

输出结果4：模型评估结果

复制

	准确率	召回率	精确率	F1
训练集	1	1	1	1
交叉验证集	0.673	0.673	0.695	0.665
测试集	0.604	0.604	0.571	0.582

图表说明：

上表中展示了训练集和测试集的分类评价指标，通过量化指标来衡量随机森林对训练、测试数据的分类效果。

- 准确率：预测正确样本占总样本的比例，准确率越大越好。
- 召回率：实际为正样本的结果中，预测为正样本的比例，召回率越大越好。
- 精确率：预测出来为正样本的结果中，实际为正样本的比例，精确率越大越好。
- F1：精确率和召回率的调和平均，精确率和召回率是互相影响的，虽然两者都高是一种期望的理想情况，然而实际中常常是精确率高、召回率就低，或者召回率低、但精确率高。若需要兼顾两者，那么就可以用F1指标。
- oob_score：对于分类问题，oob_score是袋外数据的准确率。若在建立树过程中选择有放回抽样时，大约1/3的记录没有被抽取。没有被抽取的自然形成一个对照数据集，可用于模型的验证。所以随机森林不需要另外预留部分数据做交叉验证，其本身的算法类似交叉验证，而且袋外误差是对预测误差的无偏估计（当算法参数选择了“袋外测试数据”后，才会通过oob_score来检验模型的泛化能力）。

输出结果5：预测结果

测试集

下载

预测结果Y	住院天数是否大于7天	预测测试结果概率_0.0	预测测试结果概率_1.0	WBC	N	PLT	异淋	ALTUL	ASTUL	TBummolI	LDHUL	EBVVCAM	EBDN拷贝数ml
1.0	1.0	0.3504134389603116	0.6495865610396884	10.93	31.2	243	15	42	59	5	432	160	155000

0.0	0.0	0.65	0.35	19.51	14.8	361	11	14	23	7.5	253	71.8	18500
1.0	1.0	0.1424215238138658	0.8575784761861343	27.67	13.9	192	11	371	452	19	673	160	791000
1.0	1.0	0.09310316106206651	0.9068968389379335	22.11	12.7	211	7	301	304	11.2	751	160	23050
0.0	1.0	0.5079598334155296	0.4920401665844703	7.59	25.5	192	11	24.4	49.7	4.6	474.2	98.1	23050
1.0	1.0	0.2320981671320367	0.7679018328679632	10.12	16	235	11	189	149	6.4	513.6	160	23050
0.0	1.0	0.6239035563592527	0.37609644364074746	7.71	37.4	346	28	23	64	4.5	428	87.8	36300
1.0	0.0	0.3450750360750361	0.6549249639249638	14.57	34.3	213	16	15	35	5	460	109	22400
1.0	1.0	0.1652870378536762	0.8347129621463237	19.2	18.1	234	15	25	58	5.1	623	160	64400
1.0	1.0	0.2949061771561771	0.7050938228438228	19.62	25.3	289	14	27	46	5.7	584	160	39200
1.0	1.0	0.41128321678321683	0.5887167832167832	13.55	30.5	159	14	40	30	4.5	397	160	23050
1.0	0.0	0.49349650349650354	0.5065034965034965	17.49	20.4	160	13	16	36	5.5	513.6	64.4	33500
1.0	1.0	0.08983596990706609	0.9101640300929339	22.57	18.6	219	13	104	101	8.8	638	160	127000
1.0	0.0	0.3294731943923581	0.6705268056076418	28.41	19.8	330	17	159	124	5.8	635	160	23050
1.0	0.0	0.39627523742052057	0.6037247625794794	24.53	7.6	157	17	66.8	65.8	4.3	635.9	98.4	65500

图表说明：

上表格为预览结果，只显示部分数据，全部数据请点击下载按钮导出。

上表展示了随机森林模型对测试数据的分类结果，分类结果值是拥有最大预测概率的分类组别。

输出结果6：模型预测与应用

请选择文件所在路径

模型预测

数据是否包括实际因变量值Y

图表说明：

- 系统会自动保存模型，需要注意的是：在机器学习中的随机森林算法保存的模型是非常复杂的，不是类似于线性回归那样可以用一个公式保存，系统以二进制文件方式进行序列化保存。
- 由于随机森林具有随机性，每次训练的模型可能不一致，若保存本次训练模型，后续可以直接上传数据代入到本次训练模型进行计算预测。
- 若删除本分析报告将会直接删除模型的缓存。

参考文献

[1] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11)[Online Application Software]. Retrieved from https://www.spsspro.com.

[2] 周志华. 机器学习[M]. 清华大学出版社, 2016.