

# 血液疾病辅助诊断方法研究

2019 年 6 月 25 日

## 血液样本数据指标项

表: 血常规指标项 (正负样本字段)

指标项	统一字段名	指标项	统一字段名
平均血红蛋白含量	MCH	中性粒细胞比率	NEUT_P
平均血红蛋白浓度	MCHC	中性粒细胞计数	NEUT
红细胞平均体积	MCV	血小板压积	PCT
平均血小板体积	MPV	<u>红细胞分布宽度</u>	<u>RDW</u>
嗜碱性粒细胞计数	BAS	淋巴细胞计数	LY
嗜碱性粒细胞比率	BAS_P	淋巴细胞比率	LY_P
嗜酸性粒细胞计数	EOS	单核细胞计数	MONO
嗜酸性粒细胞比率	EOS_P	单核细胞比率	MONO_P
血红蛋白	HB	红细胞计数	RBC
血小板分布宽度	PDW	红细胞压积	HCT
血小板计数	PLT	性别	Sex
白细胞计数	WBC	年龄	Age
诊断结果	Result		

# 正样本数据

表: 疾病分类汇总数据

疾病分类	英文对照	总数
淋巴瘤	Lymphoma	5202
多发性骨髓瘤 [卡勒病] (M97320/3)		2183
急性髓样白血病		2051
血小板减少性紫癜		1155
急性淋巴细胞白血病		1154
过敏性紫癜 [亨诺克 (- 舍恩莱因) 紫癜]		1105
再生障碍性贫血 NOS		458
噬血细胞综合症		374
骨髓增生异常综合征		229
合计		13911

## 负样本数据

合计: 728,599

# 实验流程

## 1 样本比例

- 1 : 1
- 1 : 2
- 1 : 5
- 1 : 10

## 2 训练模型

- 逻辑回归
- 随机森林
- Light GBM
- XGBoost
- 神经网络

## 3 记录结果参数与过程

# Imbalanced Class Distribution(aka. Skewed Class)

- Challenges with Standard Machine Learning Techniques
- Resampling Techniques
- Algorithmic Ensemble Techniques

# Challenges with Standard Machine Learning Techniques