# 数据分析处理

## 一、概括分析

包括id和label一共107列，包括表头一共6297行

其中有缺失数据的一共75列，没有缺失数据的一共32列

对于有缺失数据的列，都只缺少了724-752个左右

## 二、label分析

分析所有label的分布情况，可以看出0类占比绝大多数

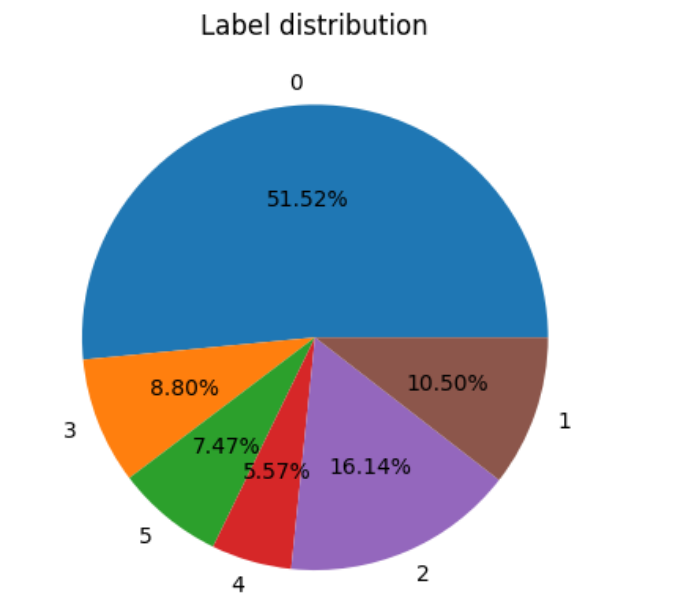


图1 标签分布情况

## 三、特征分析

从feature0-feature104一共105个特征

对于每一列分别进行数据分析，主要包括：

分布情况、最大最小值、方差大小、与label的关系(相关系数)、

特殊的几个特征：

**完全重复数值:**

feature17、feature70、feature86全部都是0建议直接删除

**大量重复数值：**

**这些特征可以选择直接丢弃或者全部用它重复的那个值填充**

**feature4** 一共6296个数据5400个0.88，732个缺失数据，只有164个有效数据，这164个数据里面只有124个0，40个1，所以也建议会直接删除

**feature23** 一共6296个数据5379个1126.4，734个缺失数据，只有183个有效数据，这183个数据里面只有132个0，51个1，所以也建议会直接删除

**feature30** 一共6296个数据5387个11，734个缺失数据，只有175个有效数据，这175个数据里面只有134个0，41个1，所以也建议会直接删除

**feature38** 一共6296个数据5403个1126.4，728个缺失数据，只有165个有效数据，这165个数据里面只有129个0，36个1，所以也建议会直接删除

**feature41** 一共6296个数据5403个1126.4，728个缺失数据，只有165个有效数据，这165个数据里面只有129个0，36个1，所以也建议会直接删除

feature2含有大量重复数值，重复的数值为：288358.4，重复了5379次

feature4含有大量重复数值，重复的数值为：0.88，重复了5400次

feature23含有大量重复数值，重复的数值为：1126.4，重复了5379次

feature30含有大量重复数值，重复的数值为：11.0，重复了5387次

feature36含有大量重复数值，重复的数值为：4505.6，重复了5373次

feature38含有大量重复数值，重复的数值为：1126.4，重复了5403次

feature41含有大量重复数值，重复的数值为：1980.0，重复了5369次

feature48含有大量重复数值，重复的数值为：288358.4，重复了5368次

feature58含有大量重复数值，重复的数值为：18022.4，重复了5386次

feature64含有大量重复数值，重复的数值为：72954675.2，重复了5397次

feature79含有大量重复数值，重复的数值为：44.0，重复了5397次

另外feature2缺失数据为：740个，有效数据只有：177个

另外feature4缺失数据为：732个，有效数据只有：164个

另外feature23缺失数据为：734个，有效数据只有：183个

另外feature30缺失数据为：734个，有效数据只有：175个

另外feature36缺失数据为：744个，有效数据只有：179个

另外feature38缺失数据为：728个，有效数据只有：165个

另外feature41缺失数据为：752个，有效数据只有：175个

另外feature48缺失数据为：749个，有效数据只有：179个

另外feature58缺失数据为：737个，有效数据只有：173个

另外feature64缺失数据为：733个，有效数据只有：166个

另外feature79缺失数据为：735个，有效数据只有：164个