# 实验报告

姓名：张嘉铭

学号：1829210092

## 实验名称：Colab 3: 聚类

### 任务概述：

使用spark机器学习工具包中聚类相关的算法，对sklearn自带的乳腺癌数据集进行聚类分析，比较聚类结果与真实标签之间的差异。

### 使用数据

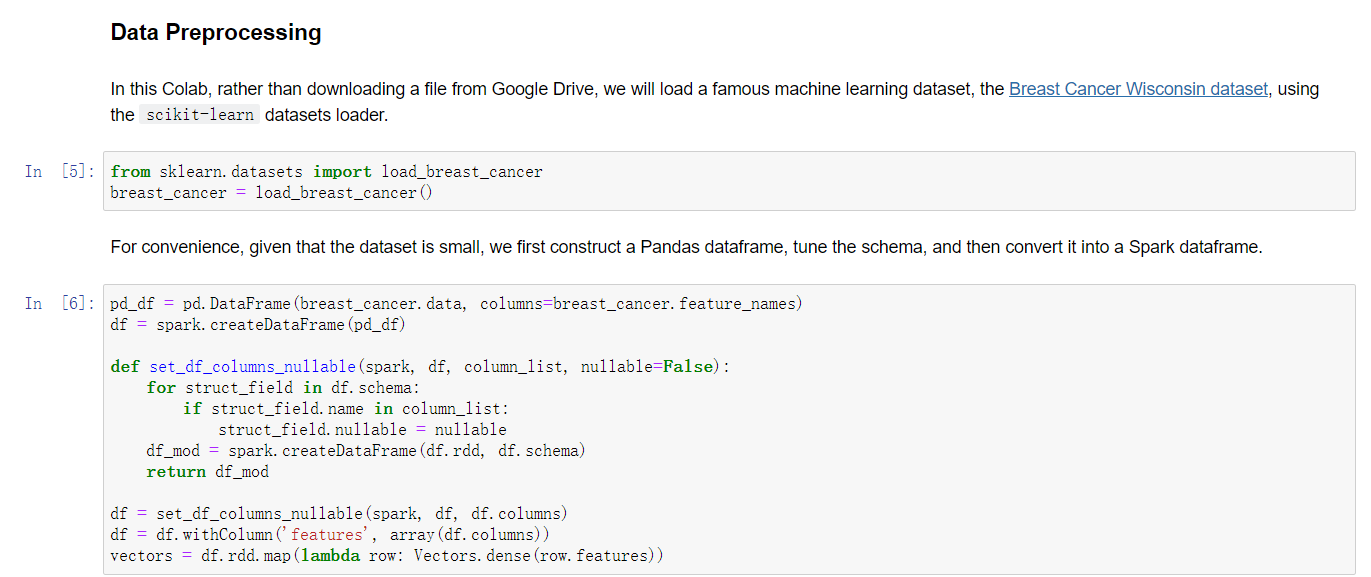
sklearn.datasets.load\_breast\_cancer: 乳腺癌患者数据集

### **实**验过程及结果

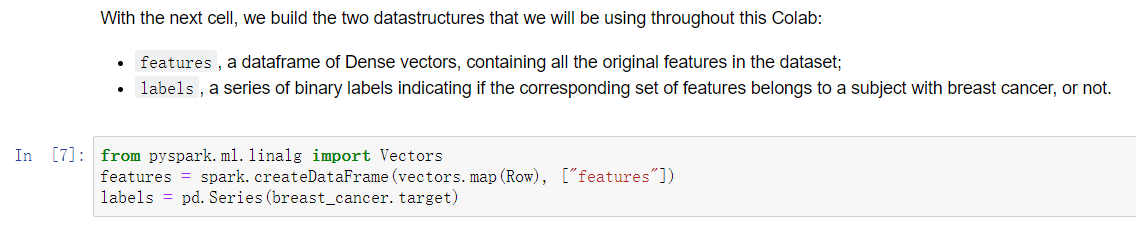
1. 环境配置



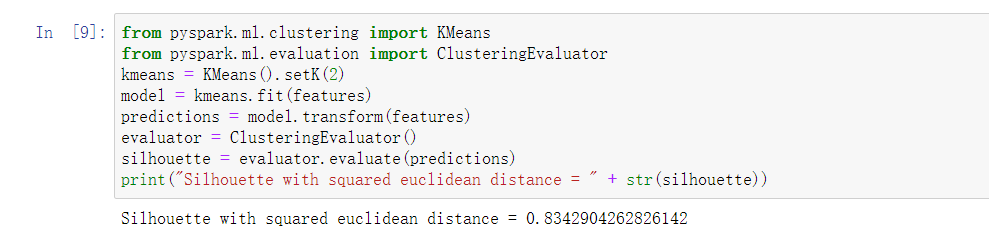
1. 加载数据集并将其转成Spark dataframe格式

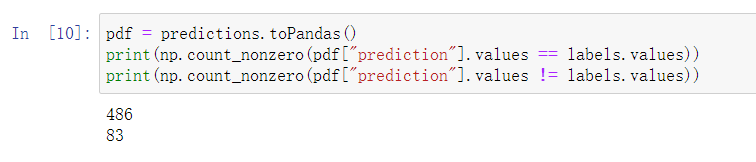


1. 得到向量集(features)与其是否对应乳腺癌患者(labels)

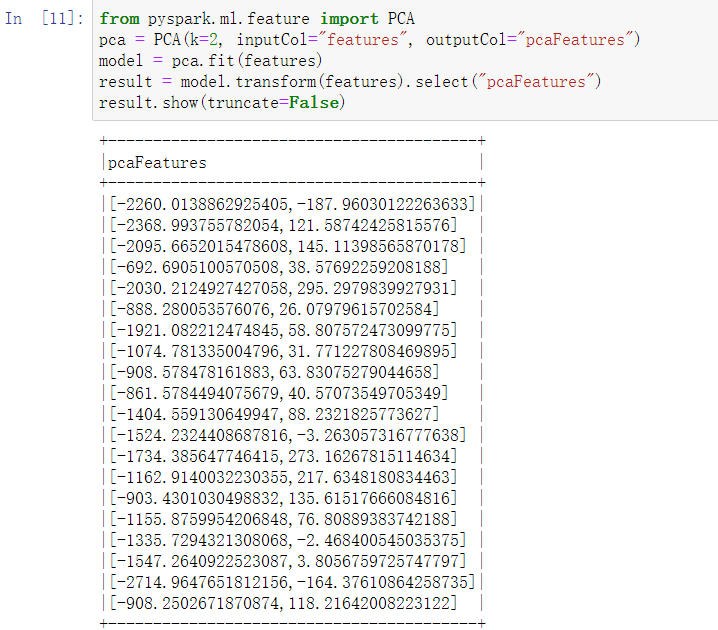


1. 用keams算法聚成2簇，并计算得分、与答案比较(486个对上)。

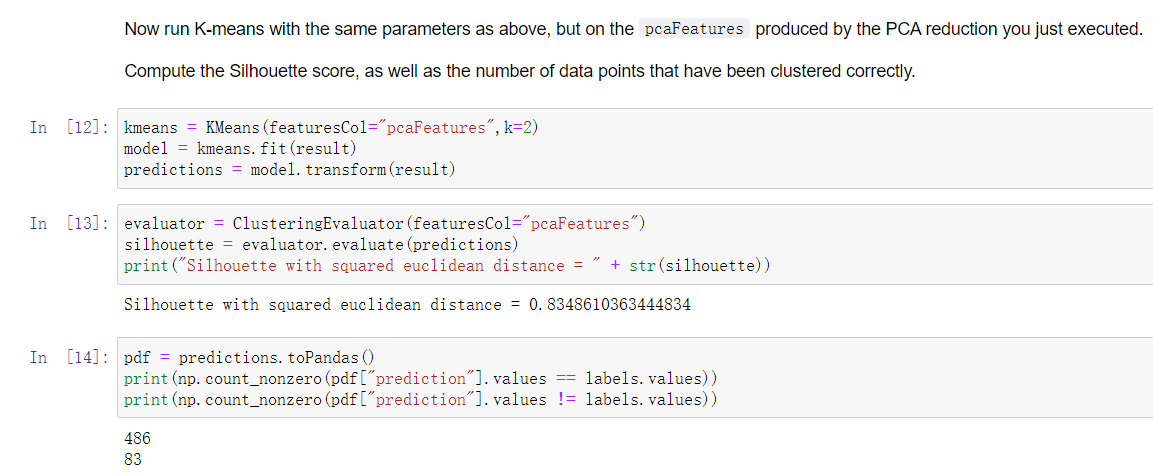




1. 使用PCA将原30维向量集转为2维向量集。



1. 再次使用keams算法将上面的二维向量集聚成两簇，计算得分、与答案比较(486个对上)。



### 实验中的问题及解决方法

无。