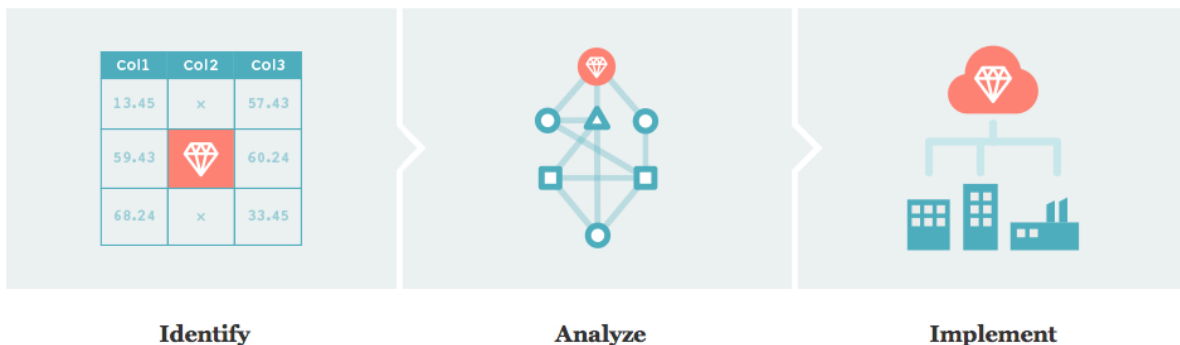


首发于
数据之路

写文章

登录



scikit-learn实战

Jack jacksu · 3 个月前

上一篇《[使用python机器学习（五）-scikit-learn](#)》简单介绍了scikit-learn的基本知识，此文主要通过一个公开数据集，使用scikit-learn进行实战，其中会使用到numpy、pandas、matplotlib等，可以参考前面的文章。

数据加载

首先，数据要被加载到内存中，才能对其操作。Scikit-Learn库在它的实现中使用了NumPy数组，所以我们将用Numpy来加载*.csv文件。让我们从[UCI Machine Learning Repository](#)下载其中印度人糖尿病的数据集。该数据集共有九列，分别为：

- 怀孕次数
- 口服葡萄糖耐量试验中2小时中血浆葡萄糖浓度
- 舒张压（mm Hg）
- 三头肌皮褶厚度（mm）
- 2小时血清胰岛素（ $\mu\text{U}/\text{ml}$ ）
- 体重指数（ $\text{kg}/(\text{身高}(\text{m}))^2$ ）
- 糖尿病谱系功能
- 年龄
- 是否得糖尿病label

