大数据人工智能使用说明书

# 使用环境要求

1. 安装 python 3.6
2. 安装 pandas，sklearn 第三方库。

# 数据集说明

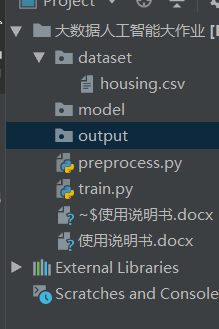
使用的是有关楼价预测的数据集。有 9 个特征，预测街区楼市中位数的数据集。

链接为：<https://www.kaggle.com/harrywang/housing?rvi=1>

# 使用说明

## 使用 pycharm 打开项目【大数据人工智能大作业】

文件结构如下图：



dataset/ 储存原始的数据集和预处理后的训练集与测试集

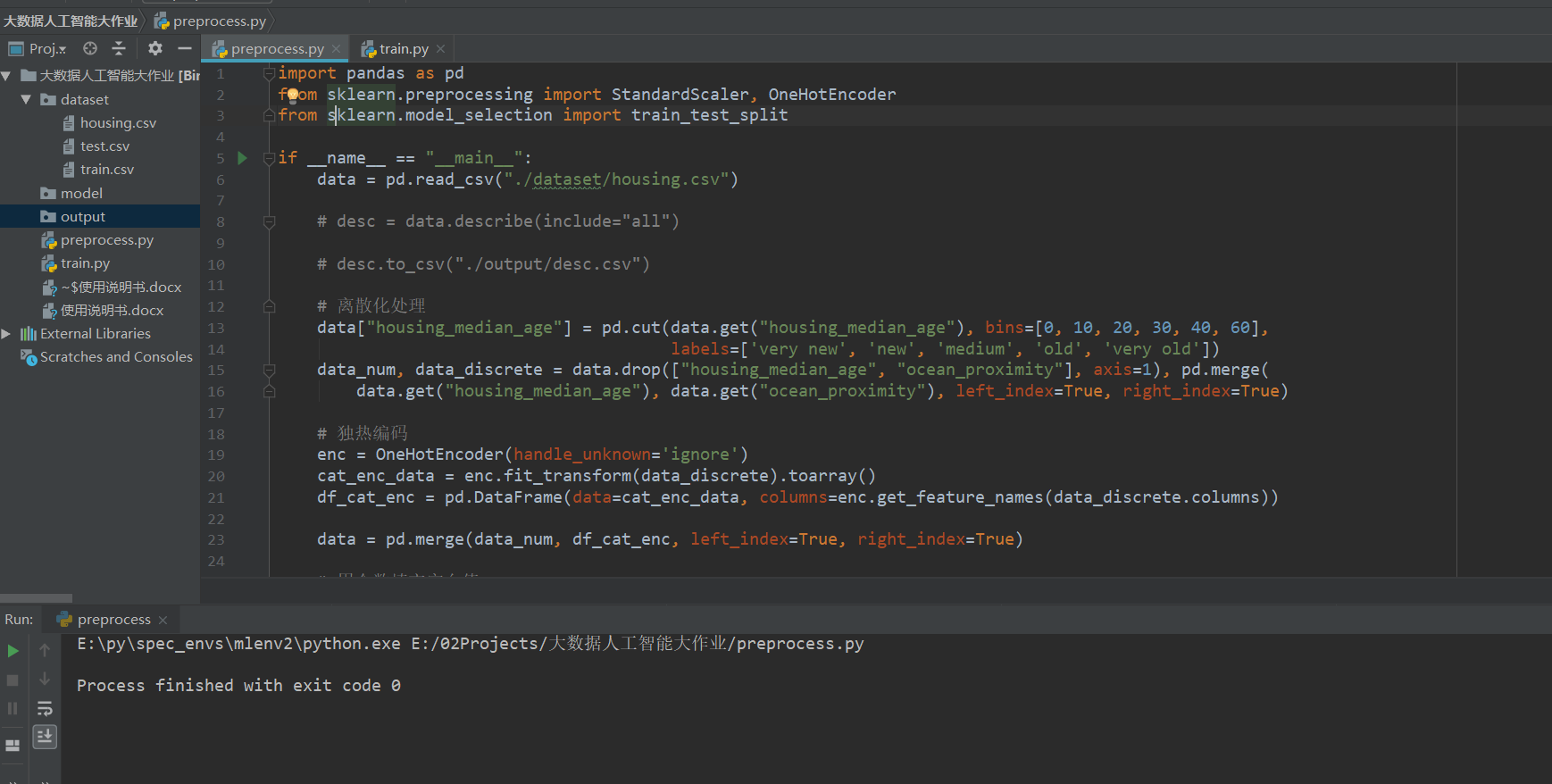
model/ 储存训练好的模型

output/ 储存最终输出的预测结果

**2. 执行 preprocessing.py，进行数据预处理**

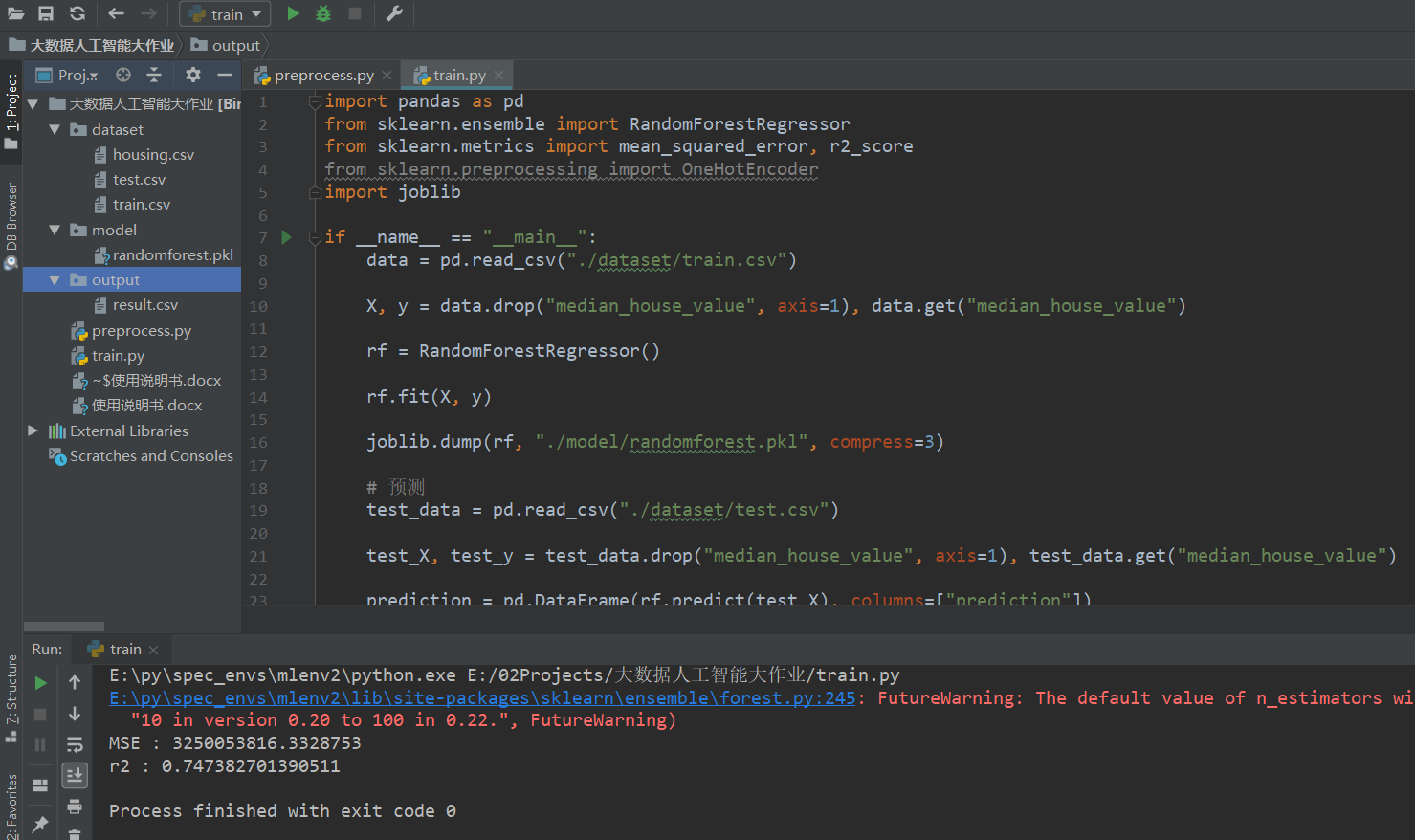
执行后，会在 dataset/ 下生成 【train.csv】 与 【test.csv】文件。

该文件中的特征都已经经过了 one-hot 编码，缺失值使用众数填充，以及归一化处理。



**3. 执行 train.py，进行模型训练与回归预测**

执行后，会在 model/ 下生成 【randomforest.pkl】文件。以及在 output/ 下生成【result.csv】文件。并输出测试集上预测结果与真实值的均方差和 R2 系数。



result.csv 文件截图：左侧是真实值，右侧是回归值

