1 决策树分类

* 1. 分类模型C4.5决策树的基本流程  
      首先，C4.5是决策树算法的一种。决策树算法作为一种分类算法，目标就是将具有p维特征的n个样本分到c个类别中去。相当于做一个投影，c=f(n)，将样本经过一种变换赋予一种类别标签。决策树为了达到这一目的，可以把分类的过程表示成一棵树，每次通过选择一个特征pi来进行分叉。那么怎样选择分叉的特征呢？每一次分叉选择哪个特征对样本进行划分可以最快最准确的对样本分类呢？不同的决策树算法有着不同的特征选择方案。如下：ID3用信息增益、C4.5用信息增益率、CART用gini系数

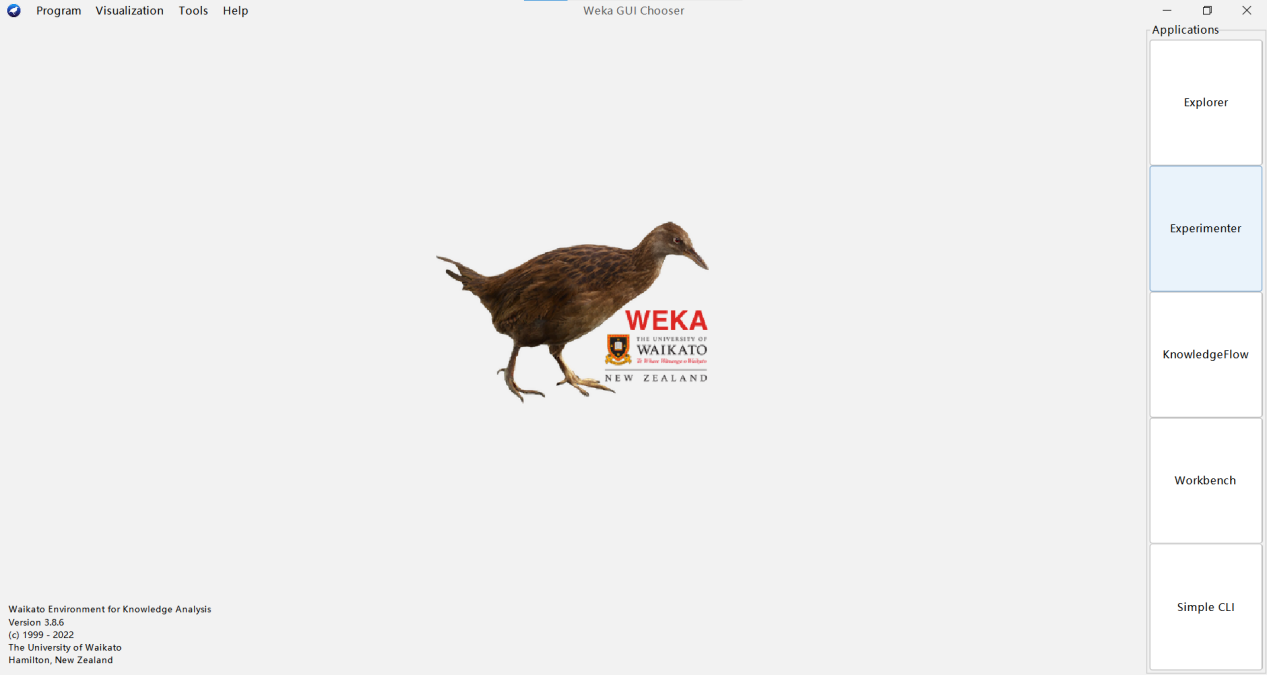
C4.5决策树是一种经典的分类模型，其基本流程如下：  
 （1）收集数据：收集包含目标变量和特征变量的数据集。

（2）数据预处理：对数据进行处理，如去除重复数据、缺失值  
 （3）数据准备（数据处理+特征选择）：选择对分类有决定性作用的特征，提高分类的准确性和效率。C4.5采用信息增益比作为特征选择的标准，选择信息增益比最大的特征作为划分标准。  
 （4）构建决策树：基于选定的特征进行决策树的构建，构建过程中采用递归方法，在每个节点处选定最优特征，生成相应的子节点。  
 （5）决策树剪枝：对生成的决策树进行剪枝，以避免过拟合。  
 （6）模型评估：使用测试数据集来评估模型的性能和准确性。

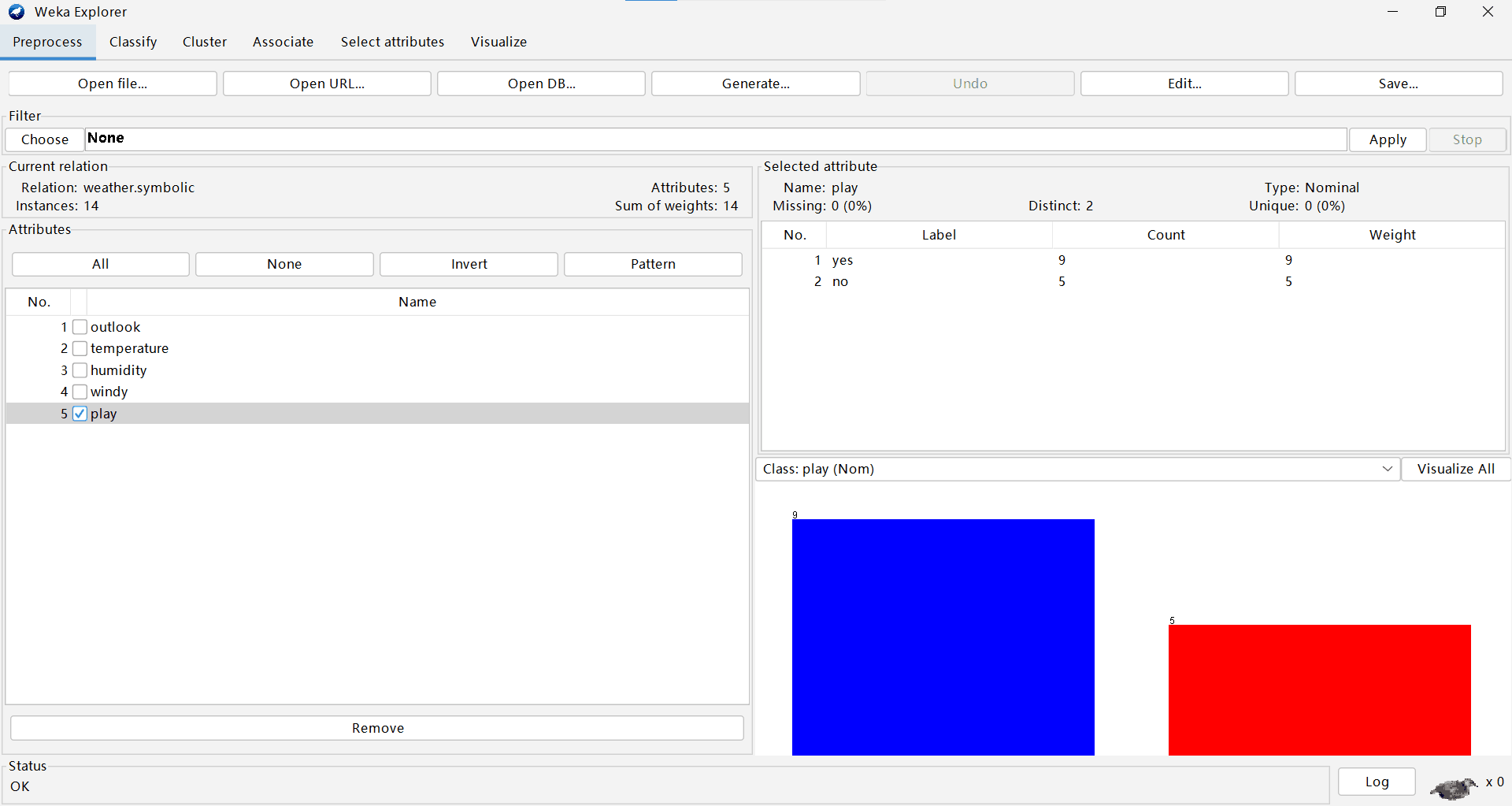
（7）模型应用：使用训练好的模型对新数据进行分类。

1.2 使用内嵌数据集与内嵌C4.5分类算法运行数据

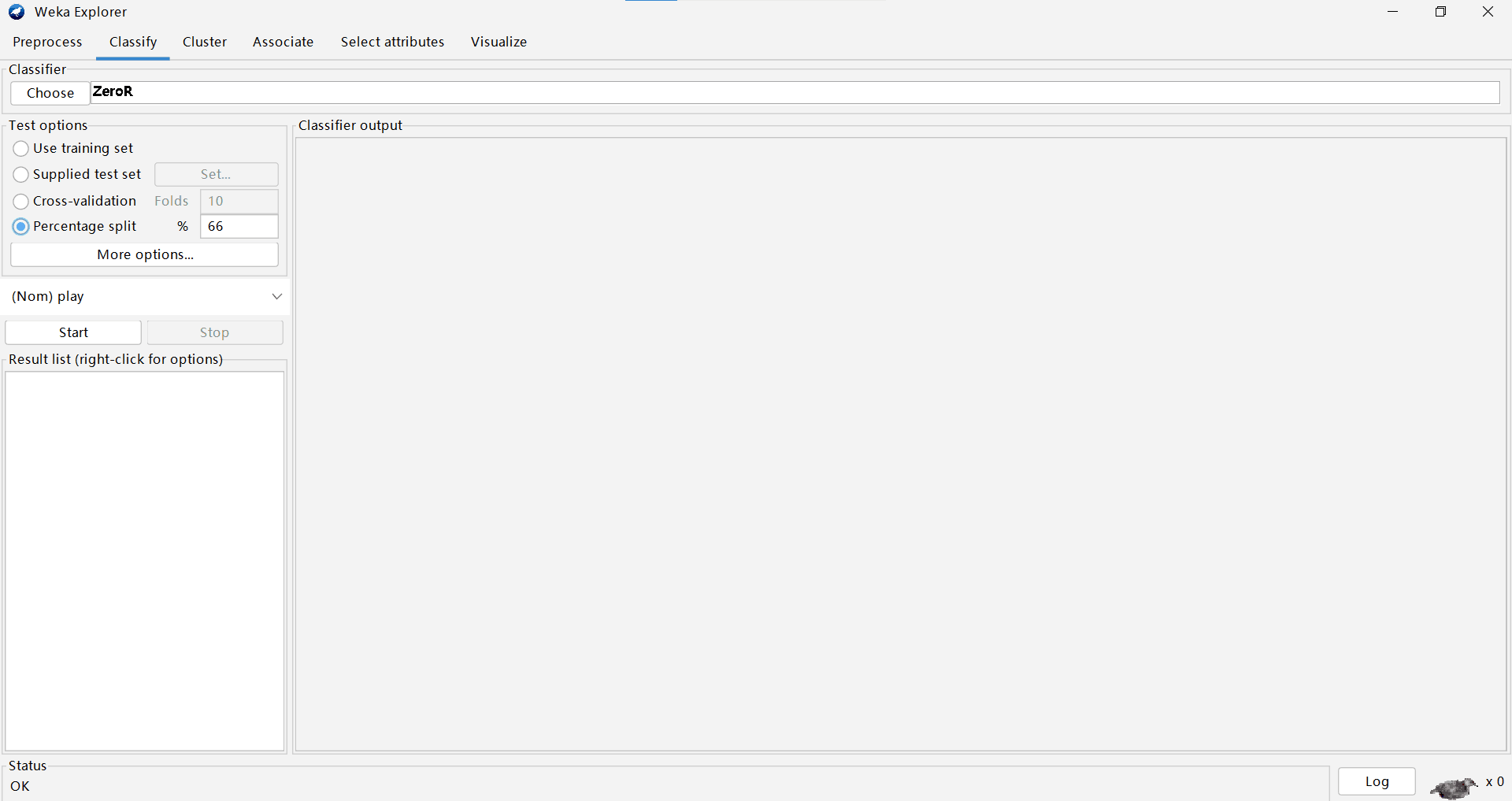
打开软件界面。



导入内嵌数据weather.nominal.arff。

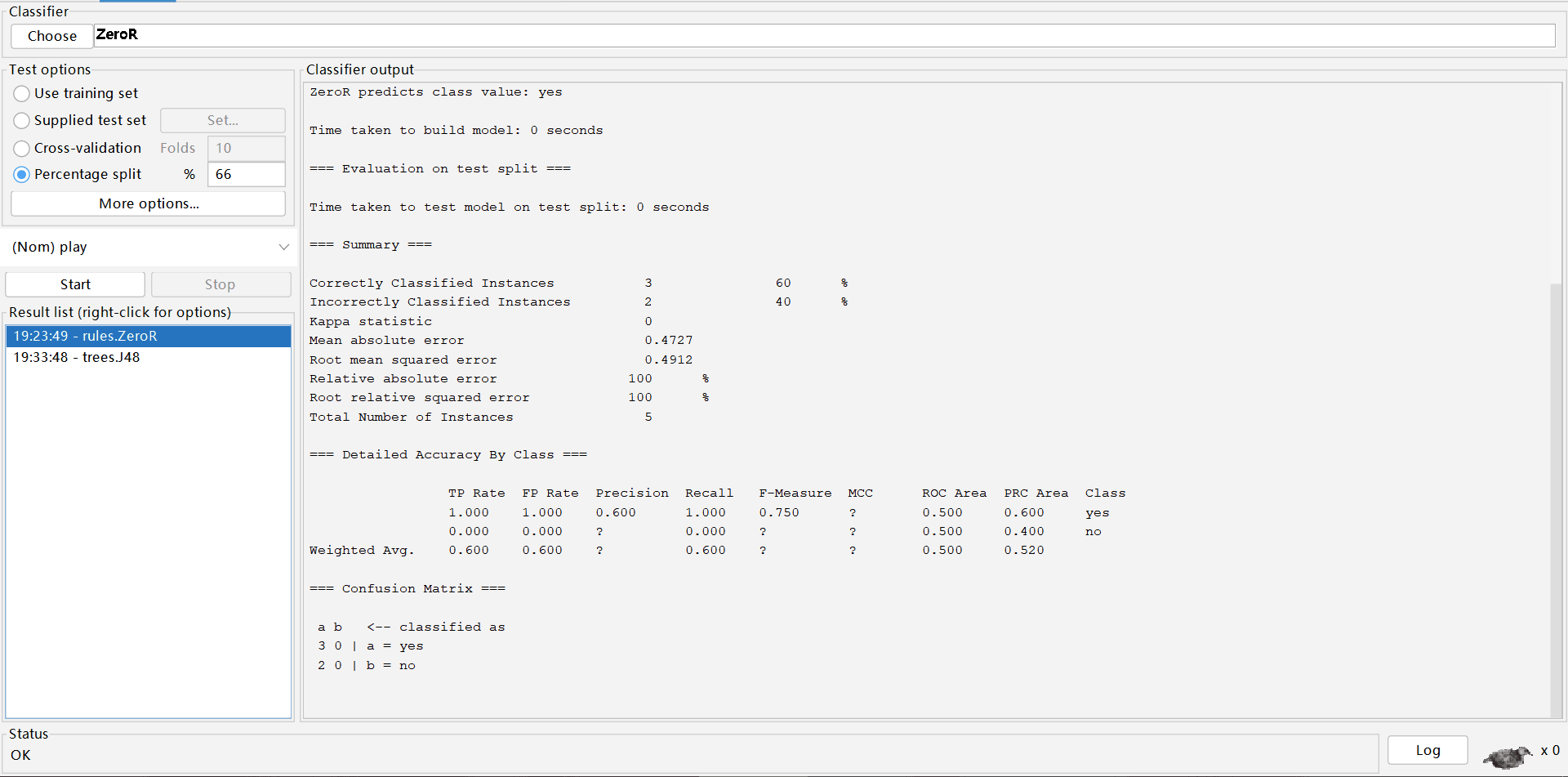


选择ZeroR分类器，并按照比例分割。分割的比例为66%。



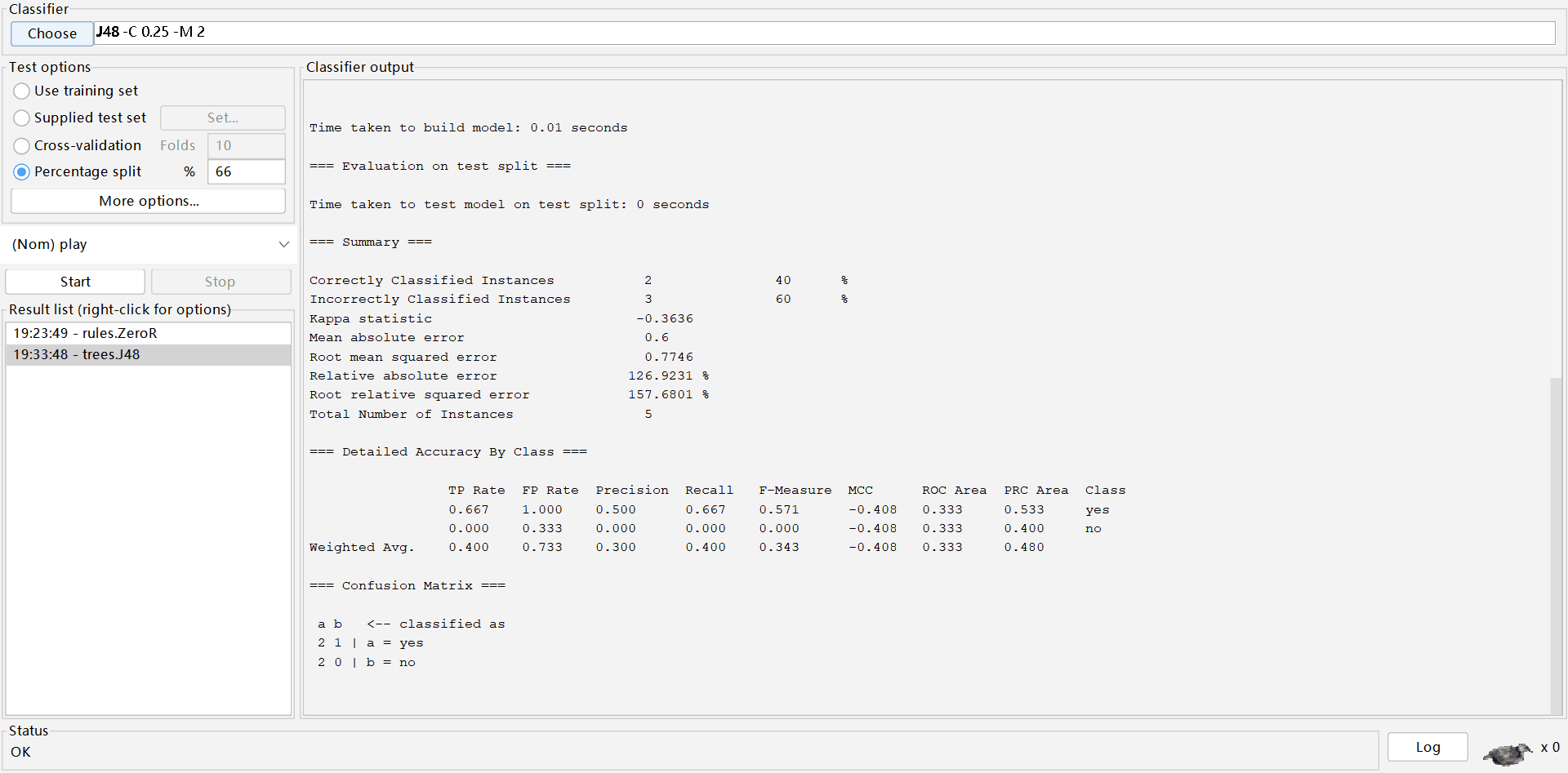
1.3 数据预处理过程、算法处理结果

使用ZeroR分类器，并按照比例分割的结果。

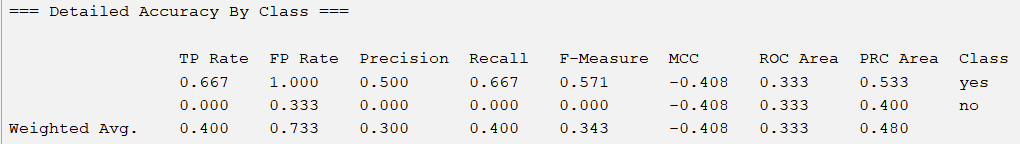


1.4 可视化决策树结果以及分析当前数据集分类统计指标

打开之前生成的.arff文件，点击面板中的Clissify 、 Choose 、Trees 、J48 、Start。右边会输出结果，包括树的结构，正确率等。也可以右击左下角面板的Result List -> Visualize Tree，这样可以更直观地理解决策树的结构。如下是运行后的结果。



分析当前数据集分类统计指标。这是Weka中的分类器评估指标之一，下面解释每个指标的含义：  
 （1）TP Rate：True Positive Rate，即真正例率，表示正例中被分类器正确预测为正例的比例。  
 （2）FP Rate：False Positive Rate，即假正例率，表示负例中被分类器错误预测为正例的比例。  
 （3）Precision：精确率，表示分类器预测为正例的实例中，真正例的比例。  
 （4）Recall：召回率，表示正例中被分类器预测为正例的比例。  
 （5）F-Measure：F值，综合考虑精确率和召回率，是精确率和召回率的调和平均数，F-Measure越大，表示分类器性能越好。  
 （6）MCC：Matthews Correlation Coefficient，是一种常用的分类器性能评估指标，值在-1到1之间，值越大表示分类器性能越好。  
 （7）ROC Area：ROC曲线下的面积，ROC Area越大，表示分类器性能越好。  
 （8）PRC Area：PRC曲线下的面积，PRC Area越大，表示分类器性能越好。  
 （9）Class：表示数据集中的类别，上述指标是对每个类别单独计算的。



决策树如下。

