

数据挖掘-决策树

概述

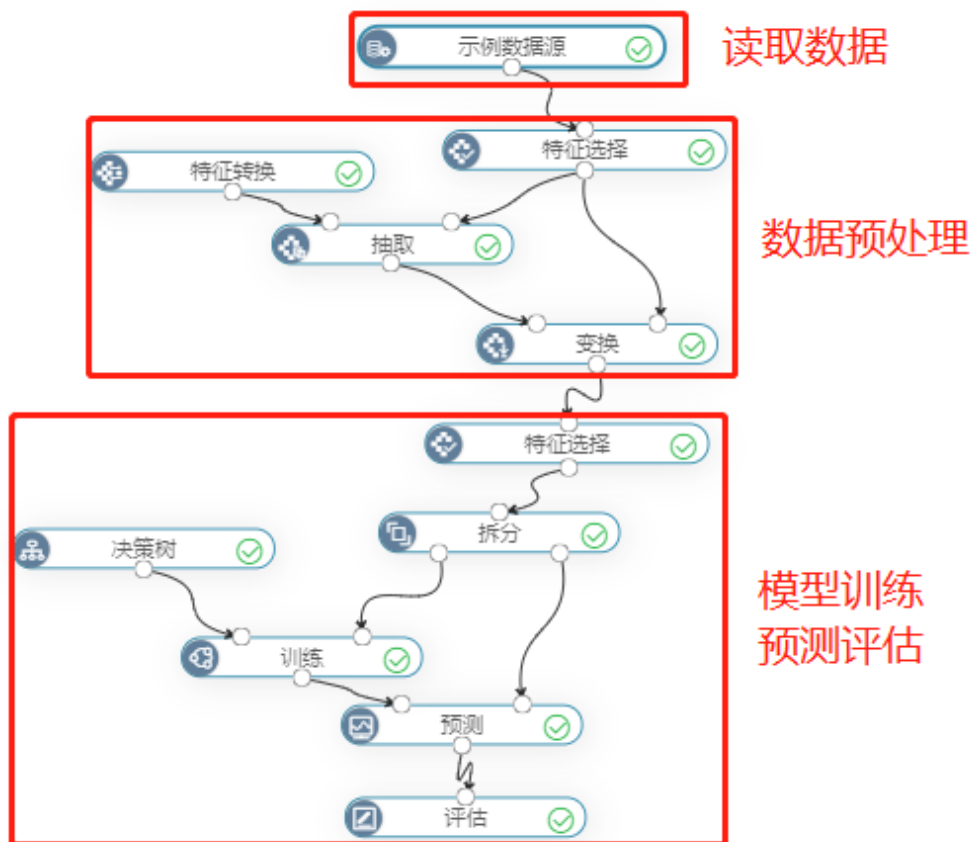
决策树是一种常用的分类算法，它是一种树形结构，其中每个内部节点表示一个属性上的测试，每个分支代表一个测试输出，每个叶节点代表一种类别。根节点到每个叶子节点均形成一条分类的路径规则。而对新的样本进行测试时，只需要从根节点开始，在每个分支节点进行测试，沿着相应的分支递归地进入子树再测试，一直到达叶子节点，该叶子节点所代表的类别即是当前测试样本的预测类别。

优势：可直接查看决策树分析的可视化效果，如下图：



示例

使用“糖尿病预测”案例数据，预测是否有糖尿病。



特征转换是为了将各变量中的类别型变量转换成数值型变量，类别型无法进入模型，转换后方便算法模型学习；

决策树的参数如下：

参数名称	值	说明
分裂特征的数量	取值范围：>=2的整数； 默认值：32。	对连续类型特征进行离散时的分箱数； 该值越大，模型会计算更多连续型特征分裂点且会找到更好的分裂点，但同时也会增加模型的计算量；
树的深度	取值范围：[1, 30]的整数；默认值为4。	当模型达到该深度时停止分裂； 树的深度越大，模型训练的准确度更高，但同时也会增加模型的计算量且会导致过拟合；
计算信息增益的方式	gini	分裂标准，Entropy表示熵值，Gini表示基尼指数；
	entropy	