数据挖掘报告

胡正浩 5110369037

一、数据描述

我选用的数据集是 EEG EYE STATE，共 14980 条记录。每条记录由 15 个属性来描述，前 14 个属性分别为
AF3，F7，F3，FC5，T7，P7，O1，O2，P8，T8，FC6，F4，F8，AF4，表示大脑各个部分的脑电波的数值，都为连续型的数值变量；最后一个属性表示分类的类别，由 0 和 1 表示，代表眼睛是处于开还是闭的状态。

二、分类

分类过程分为数据预处理阶段，数据分析阶段，和评估阶段。数据预处理阶段将原始数据经过恰当的处理以方便后续数据分析的进行。数据分析阶段包括模型选择和模型的计算机实现两个部分。评估阶段对分类模型的复杂度和分类的准确度。

（一）、预处理阶段

在预处理阶段，为了方便后续对于数据集的操作，我先将所有数据读入一个 14980 行 15 列的二位数组中。将数据全部读入内存只需占用 1.7MB 的内存空间，但能使得数据的操作和运行效率大大提升，因此这么做是合理的。

同时，为了避免 overfitting，以及减少 outlier 出现的概率，在预处理阶段，我用随机读取的方法，仅选取了部分的数据加入到二维数组中。于此同时，为了避免由于训练数据过少造成的 underfitting，因此选取 60%的数据作为训练数据。为了使每次抽取的数据被抽到的概率相同，采用了有放回的抽样策略，即每次抽取出一条记录后不将该记录从数据集中删去。

（二）、数据分析阶段

1、模型选择

在这次数据的分类中我选择的是决策树分类模型。主要考虑因素如下：

（1）由于数据均为连续数据，因此不适用基于规则的分类模型

（2）计算各个属性间的相关系数，发现很多属性之间都有较大的关联度，因此贝叶斯分类器也不适用。

（3）对所有属性进行简单离散化，发现两个分类在所有属性上都有很高的重合度，因此也不适用最近邻分类器。

（4）考虑到大作业时间有限，以及其他课业负担的因素，对于 SVM 和人工神经网络等分类器也不太熟悉，因此选择了决策树分类模型。

（5）用 Weka 进行分类测试时，发现基于决策树的分类具有最高的准确度。

2、算法实现

模型生成的算法是基于贪心算法和递归函数实现的。生成决策树时，每次都选取 Gini 系数最小的属性（同时也选取该属性中 Gini 系数最小的划分点）作为一个分类节点，将分类后的数据集作为输入再一次计算下一节点，进行递归计算。在算法实现中主要面临一下几个问题：连续数据离散化，节点选择，先/后剪枝实现，泛化误差估计。

（1）连续数据离散化采用 Gini 系数的方法，先将一个属性的所有值从小到大进行排序，计算以每两个相邻数值的平均值来划分区间的 Gini 系数，选取 Gini 系数最小的

划分点作为最优划分。

（2）节点选择同样采用计算 Gini 系数的方法，通过计算除所有 14 个属性最优划分点的 Gini 系数，选取其中 Gini 系数最小的属性的划分点作为决策树的一个节点。并用此节点来对数据集进行分类，将分好的数据集作为输入，递归计算出所有节点。可通过修改 MIN_TH_CLASS 变量来改变决策树何时停止生长。若数据集的数据个数小于设定值，则停止分类。

（3）为了减少树的大小，我同时采用了先剪枝和后剪枝两种剪枝技术来精简决策树。先剪枝是在生成决策树时设置一个正确率的阈值，若通过某一节点分类的正确率达到该阈值，则决策树停止生长。决策树的阈值通过公共变量 TH_CRATE 进行设置。后剪枝是在决策树生成之后，重新扫描整个决策树。实现时，我采用的是 REP 技术，分别计算出通过某一节点直接预测分类的错误率，记为 E（node），和通过以该节点为根的子树的错误率，记为 C（node），来判断是否进行剪枝。若 P（node）与 C（node）相当，或 P（node）< C（node），则将该子树替换成相应叶节点。

（4）在后剪枝过程中，为了引入决策树复杂度带来的代价问题，我增加了泛化误差的考虑。即，当 P（node）<= C（node）+ G（node）时进行剪枝操作。其中 G（node）是泛化误差，为泛化误差系数乘以子树节点个数的值。泛化误差系数可通过 GERR 变量进行设置，默认为 0.5。

（三）、评估阶段

在评估阶段我采用 Holdout 的方法，以 2/3 作为训练数据，1/3 作为测试数据，由于还要考虑到 outlier 和 overfitting 的因素，因此留下 10%的数据不做使用。因此，随机取 30%的数据作为测试数据来评估模型。Houldout 大小可通过 HOLDOUT 变量进行修改。

评估内容有三个部分，混淆矩阵，准确率，决策树复杂度。

通过使用已生成的决策树和测试数据集，计算出混淆矩阵。再通过混淆矩阵可以计算出准确率。同时，由于决策树是二叉树，因此决策树的复杂度可由节点数量来表示。

四、实验测试

index:7; <= 4051.280000

---> index:2; <= 3988.720000

---> ---> index:2; <= 3972.565000

---> ---> index:2; > 3972.565000

---> ---> ---> index:2; <= 3974.870000

---> ---> ---> index:2; > 3974.870000

---> ---> ---> ---> index:2; <= 3975.900000

---> ---> ---> ---> index:2; > 3975.900000

---> ---> ---> ---> ---> index:2; <= 3986.150000

---> ---> ---> ---> ---> index:2; > 3986.150000

---> index:2; > 3988.720000

---> ---> index:2; <= 4042.305000

---> ---> ---> index:2; <= 3998.460000

---> ---> ---> ---> index:2; <= 3992.820000

---> ---> ---> ---> index:2; > 3992.820000

---> ---> ---> index:2; > 3998.460000

---> ---> ---> ---> index:2; <= 4030.770000

---> ---> ---> ---> ---> index:2; <= 4008.720000

---> ---> ---> ---> ---> ---> index:2; <= 4004.620000

---> ---> ---> ---> ---> ---> ---> index:2; <= 4003.590000

---> ---> ---> ---> ---> ---> ---> index:2; > 4003.590000

---> ---> ---> ---> ---> ---> index:2; > 4004.620000

---> ---> ---> ---> ---> index:2; > 4008.720000

---> ---> ---> ---> ---> ---> index:2; <= 4010.260000

---> ---> ---> ---> ---> ---> index:2; > 4010.260000

---> ---> ---> ---> ---> ---> ---> index:2; <= 4028.210000

---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4021.540000

---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4021.540000

---> ---> ---> ---> ---> ---> ---> index:2; > 4028.210000

---> ---> ---> ---> index:2; > 4030.770000

---> ---> index:2; > 4042.305000

index:7; > 4051.280000

---> index:2; <= 3984.620000

---> ---> index:2; <= 3953.850000

---> ---> index:2; > 3953.850000

---> ---> ---> index:2; <= 3983.590000

---> ---> ---> ---> index:2; <= 3978.970000

---> ---> ---> ---> ---> index:2; <= 3977.950000

---> ---> ---> ---> ---> ---> index:2; <= 3961.030000

---> ---> ---> ---> ---> ---> index:2; > 3961.030000

---> ---> ---> ---> ---> ---> ---> index:2; <= 3963.080000

---> ---> ---> ---> ---> ---> ---> index:2; > 3963.080000

---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3969.740000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3967.690000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3967.690000

---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3969.740000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3973.330000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3972.820000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3970.260000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3970.260000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3971.280000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3971.280000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3972.820000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3973.330000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3974.870000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3974.870000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3977.440000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3975.900000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3975.900000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3977.440000

---> ---> ---> ---> ---> index:2; > 3977.950000

---> ---> ---> ---> index:2; > 3978.970000

---> ---> ---> ---> ---> index:2; <= 3979.490000

---> ---> ---> ---> ---> index:2; > 3979.490000

---> ---> ---> ---> ---> ---> index:2; <= 3982.560000

---> ---> ---> ---> ---> ---> ---> index:2; <= 3982.050000

---> ---> ---> ---> ---> ---> ---> index:2; > 3982.050000

---> ---> ---> ---> ---> ---> index:2; > 3982.560000

---> ---> ---> index:2; > 3983.590000

---> index:2; > 3984.620000

---> ---> index:2; <= 4056.410000

---> ---> ---> index:2; <= 4018.460000

---> ---> ---> ---> index:2; <= 3985.640000

---> ---> ---> ---> index:2; > 3985.640000

---> ---> ---> ---> ---> index:2; <= 3997.950000

---> ---> ---> ---> ---> ---> index:2; <= 3991.790000

---> ---> ---> ---> ---> ---> ---> index:2; <= 3988.210000

---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3987.690000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3986.150000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3986.150000

---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3987.690000

---> ---> ---> ---> ---> ---> ---> index:2; > 3988.210000

---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3989.230000

---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3989.230000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3989.740000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3989.740000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3990.260000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3990.260000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3990.770000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3990.770000

---> ---> ---> ---> ---> ---> index:2; > 3991.790000

---> ---> ---> ---> ---> ---> ---> index:2; <= 3995.900000

---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3993.330000

---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3993.330000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3994.360000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3994.360000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3995.380000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3995.380000

---> ---> ---> ---> ---> ---> ---> index:2; > 3995.900000

---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3996.920000

---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3996.920000

---> ---> ---> ---> ---> index:2; > 3997.950000

---> ---> ---> ---> ---> ---> index:2; <= 4011.790000

---> ---> ---> ---> ---> ---> ---> index:2; <= 4010.770000

---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4010.260000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4002.560000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4002.050000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4001.540000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4000.000000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3998.460000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3998.460000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 3998.970000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 3998.970000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4000.000000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4000.510000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4000.510000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4001.540000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4002.050000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4002.560000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4006.150000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4005.130000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4004.620000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4004.100000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4003.080000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4003.080000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4003.590000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4003.590000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4004.100000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4004.620000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4005.130000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4006.150000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4007.690000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4007.180000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4007.180000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4007.690000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4009.230000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4008.720000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4008.720000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4009.230000

---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4010.260000

---> ---> ---> ---> ---> ---> ---> index:2; > 4010.770000

---> ---> ---> ---> ---> ---> index:2; > 4011.790000

---> ---> ---> ---> ---> ---> ---> index:2; <= 4017.950000

---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4014.360000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4012.820000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4012.820000

---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4014.360000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4015.900000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4015.380000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4015.380000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4015.900000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4016.920000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4016.920000

---> ---> ---> ---> ---> ---> ---> index:2; > 4017.950000

---> ---> ---> index:2; > 4018.460000

---> ---> ---> ---> index:2; <= 4034.870000

---> ---> ---> ---> ---> index:2; <= 4034.360000

---> ---> ---> ---> ---> ---> index:2; <= 4020.000000

---> ---> ---> ---> ---> ---> index:2; > 4020.000000

---> ---> ---> ---> ---> ---> ---> index:2; <= 4025.640000

---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4024.620000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4024.100000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4023.590000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4023.080000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4022.560000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4021.030000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4021.030000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4022.050000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4022.050000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4022.560000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4023.080000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4023.590000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4024.100000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4024.620000

---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4025.640000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4027.690000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4027.690000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4033.330000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4031.790000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4029.230000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4029.230000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4031.790000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4033.330000

---> ---> ---> ---> ---> index:2; > 4034.360000

---> ---> ---> ---> index:2; > 4034.870000

---> ---> ---> ---> ---> index:2; <= 4036.920000

---> ---> ---> ---> ---> index:2; > 4036.920000

---> ---> ---> ---> ---> ---> index:2; <= 4039.490000

---> ---> ---> ---> ---> ---> index:2; > 4039.490000

---> ---> ---> ---> ---> ---> ---> index:2; <= 4044.100000

---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4043.590000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4041.540000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4041.540000

---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4043.590000

---> ---> ---> ---> ---> ---> ---> index:2; > 4044.100000

---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4047.180000

---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4047.180000

---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4055.900000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4055.380000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4053.850000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4052.820000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4051.280000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; <= 4050.260000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4050.260000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4051.280000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4052.820000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4053.850000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4055.380000

---> ---> ---> ---> ---> ---> ---> ---> ---> ---> ---> index:2; > 4055.900000

---> ---> index:2; > 4056.410000

---> ---> ---> index:7; <= 4083.590000

---> ---> ---> ---> index:7; <= 4054.870000

---> ---> ---> ---> index:7; > 4054.870000

---> ---> ---> ---> ---> index:7; <= 4060.000000

---> ---> ---> ---> ---> index:7; > 4060.000000

---> ---> ---> ---> ---> ---> index:7; <= 4075.380000

---> ---> ---> ---> ---> ---> index:7; > 4075.380000

---> ---> ---> index:7; > 4083.590000

---> ---> ---> ---> index:2; <= 4080.000000

---> ---> ---> ---> index:2; > 4080.000000

++++++++++++++++++++++++++++++++Evaluation++++++++++++++++++++++++++++++

Accurency: 0.610666666667
Tree Complexity: 222

五、程序说明

程序中有四个全局变量作为可修改的参数：

MIN_TH_CLASS：为可继续分类的最小数据数目，用于限制决策树的生长。默认值为 5。

TH_CRATE　　：为先剪枝时停止决策树生长的阈值。默认值为 0.95。

HOLDOUT　　：为用 Holdout 策略划分数据集时，训练数据的个数。默认值为 14980（总数据条数）× 0.6（60%）

GERR　　：为泛化误差参数。默认值为 0.5。

数据预处理阶段：randPick 函数会从原始数据中随即挑选出固定条数的数据，数据数目可通过调整 HOLDOUT 变量来设置，或把数值大小作为参数传入。initialize 函数将挑选好的数据组织成二维数组的形式，并以此作为返回值。

数据分析阶段：treeGen 函数通过接收数据集作为参数，生成决策树，返回值为决策树的根节点。treeGen 函数中还包含几个子函数，分别为：stopping_cond 函数，决定生成决策树的终止条件；nodePick 函数和 discret 函数，通过计算 Gini 系数来选择局部最优的分类节点。另外，pruning 函数提供后剪枝功能，传入参数为根节点和测试数据集。

评估阶段：confuseMtx 函数用于生成混淆矩阵；accurency 函数通过计算出的混淆矩阵计算出分类的准确率；complexity 函数计算决策树节点的个数。

其他函数：Tprint 函数的功能是打印决策树；classify 函数用于判断某条数据是否能被模型正确分类。

六、遇到问题与解决

在实际编程时，由于 python 有最大递归次数的限制，因此一旦树的深度过大，程序就会报错。解决方法有两个，一是适当放宽生成决策树的算法的终止条件，减少树的深度；另一个方法是在程序开头添加如下代码，调大最大递归次数。

```
# Main Function
# Set Maxmium Recursion Depiton to 1000000000
import sys
sys.setrecursionlimit(1000000000)
```

还遇上一些由于数据集中某个属性数据值的重复导致运算时会产生除零错误。解决方法是对所有可能出现除零出错的代码部分添加 try，execpt 语句，接收并对出错信息进行处理，避免程序崩溃。