接下来，我们将利用AC\_201…..中的数据来研究不同的factors 如（……）对于不同郡县的药物使用情况类型的影响。在这个过程中我们希望可以找到对于其影响最为显著的几个特征。并通过对其进行控制来改善每个郡县的药物使用状况，并且提出可能的决策建议。

首先通过观察我们发现数据中存在缺值，且存在较多的相互相关性很大的factors。为此为了优化建模结果以及便于筛选影响力最大的factors ，我们将对于数据进行预处理

I.Preprocessing:

……

II.Decision Tree:

对于以及通过数据清洗以及特征筛选的数据，我们选择采用数据挖掘分类模型中的决策树模型，他的目标为通过我们选取的特征尽量准确地预测郡县的药物使用分类（’Severe’, ‘Critical’, ‘Normal’, ‘Good’）。而我们使用这种分类器的原因是：决策树模型可以很好的排序出，在不同的情况之中，不同的factors（特征属性）对于分类的影响的优先级。模型的具体建立如下：

**II.1: Basic Priciple of Decision Tree:**

一棵决策树由一个根节点和若干内部节点以及叶节点构成。其中每个内部节点都具有若干子节点和父节点。而每一个叶节点代表了预测的结果，它只有父节点，而输出的是不同的分类结果。内部节点则代表了一次关于关于特定数据集的划分决策过程。既从父节点接受特定数据集并将数据集中的对象划分到自己的不同的子节点中。 根节点则包含所有的数据集。从根节点到叶节点的路径则代表了一个判定的序列。

而我们的目标是生成一个这样的决策树来表示我们的分类决策过程。其生成思路如下：

对于每一个节点，我们检视每一个从父节点接收到的对象。

1. 若全是同一类，则该节点成为叶节点，返回该类；
2. 若集合中的每一个对象的特征值都一样，则无法分类，则该节点成为叶节点，返回该集合中数量最多的类。
3. 在其他的情况下，我们将根据该集合的情况，选取一个最优的属性来划分集合，并生成对应的子节点。每一个子节点将接受该集合中，该特征属性取值相同的一组的子集合

我们可以看出，对于一棵决策树来说，最关键的步骤是对于内部节点进行最优划分特征属性的选取。既对于每一个对象，判定对于其分类结果影响力最大的一个特征属性。而这个划分属性的决策过程如下

**II.2 属性划分：**

首先，为了更好的做出划分决策，我们需要建立一些量来描述该节点的集合的数据分布

II.2.1.1 Entropy ( 信息熵 )

信息熵是用来确定随机变量的不确定性的量。假设X是一个有限取值的离散随机变量，而它的概率分布为：（行内公式）

其中pi 为Xi 取值出现的概率，在该数据量级下，我们近似的使用在数据集中出现的频率来估算。而随机变量的X的信息熵的定义如下：（公式二）

**西瓜书公式4.1**

信息熵值越大，X的不确定性越大。而当我们用该值来描述某一个数据集的时候，我们考虑此处X为药品使用分类， 而pi为不同的分类 （Good, Normal…）在该集合的频率。那么我们可以得到，当信息熵值越小，该集合中的分类情况越集中于某一个特定分类，我们称其“纯度”（purity）越高。当Ent（D）= 0 时候，该集合中只有一个分类。那么，我们的决策目标则是尽可能的使得子节点的纯度提高

II.2.1.2 信息增益（Information gain）:

假定某一个离散属性A 具有 V 个不同的取值 {A\_1, A\_2, …A\_V}. 当我们选取A为划分该节点的属性，它将产生V个子节点，我们记为D^(k), k \in {1,2,…,V}。之后，我们可以根据**公式2，**计算出每一个子节点的信息熵。考虑到不同的节点的规模不同，我们给其赋予权重|D^(k)|/|D|。于是我们得以计算出，该划分决策的“信息增益”Gain(D, A)

起具体计算公式如下：

**西瓜书公式4.2**

不难得出，当该划分的信息增益越大，分类结果的子节点的纯度越高。所以我们将选取带来信息增益最大的特征属性为最优划分属性，并以此为标准划分子节点。

**II.3 连续值处理：**

在以上考虑的情况下， 我们考虑的都是有限离散的分布的特征属性。但在我们所要处理的情况下，大部分的特征属性变量都是连续变化的。为此我们需要对于该模型进行一些改进

II.3.1 连续属性离散化：

为了使得连续数据可以划分出有限个取值，我们将对于连续属性进行离散化处理。这里我们选取的方法是二分法。假设存在节点样本集 D 和连续属性 B， 而在数据集中，B 有N 个取值 {B\_1,B\_2, …… B\_N }. 那么我们可以设置一个划分点t， 使得B\_k 大于t 的集合生成一个子集 D\_t^+，小于t的生成一个子集 D\_t^-。那么，将得到N-1 个可能划分结果，因为在两个相邻取值B\_i, B\_(i+1)选取的任一值划分结果都是一样的。因此我们仍然可以得到有限个划分结果。并比较其信息增益。