Decision Tree Example

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.tree import DecisionTreeClassifier

from sklearn import preprocessing
from matplotlib import rc,font_manager

ticks_font = font_manager.FontProperties(family='Times New Roman', style='normal',
    size=12, weight='normal', stretch='normal')
ax=plt.gca()

## Loading Data ##
df=pd.read_csv('D:\Python\edx\Machine Learning\Classification\drug200.csv')
with open('Decision_Tree.txt','a') as f:
    print(df.head(),file=f)

# Preprocessing #
X=df[['Age','Sex','BP', 'Cholesterol','Na_to_K']].values
with open('Decision_Tree.txt','a') as f:
    print(X[0:5],file=f)

#Sklearn Decision trees dont handle categorical variable, so need to convert to numerical

le_sex=preprocessing.LabelEncoder()
le_sex.fit(['F','M'])
X[:,1]=le_sex.transform(X[:,1])

le_BP=preprocessing.LabelEncoder()
le_BP.fit(['LOW','NORMAL','HIGH'])
X[:,2]=le_BP.transform(X[:,2])

le_Chol=preprocessing.LabelEncoder()
le_Chol.fit(['NORMAL','HIGH'])
X[:,3]=le_Chol.transform(X[:,3])

y=df['Drug']
with open('Decision_Tree.txt','a') as f:
    print(X[0:5],file=f)
    print(y[0:5],file=f)
```

```
# Train Test Split #
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3,random_state=3)
with open('Decision_Tree.txt','a') as f:
    print('Train Set: ', X_train.shape,y_train.shape,file=f)
    print('Test Set: ', X_test.shape,y_test.shape,file=f)

#Modeling#
#Create DecisionTreeClassifier instance and then specify criterion as entrpy to find inf

dtree=DecisionTreeClassifier(criterion='entropy',max_depth=4)
dtree.fit(X_train,y_train)

#Prediction#
ptree=dtree.predict(X_test)
with open('Decision_Tree.txt','a') as f:
    print('Prediction Set: ', ptree[0:5],file=f)
    print('Y Test Set: ', y_test[0:5],file=f)

#Evaluation#
## Checking accuracy with sklearn accuracy metric

from sklearn import metrics
with open('Decision_Tree.txt','a') as f:
    print('Accuracy for Decision Tree Model (with sklearn) is : ', metrics.accuracy_scor

##Calculating accuracy without sklearn library
acc=np.mean(y_test==ptree)
with open('Decision_Tree.txt','a') as f:
    print('Accuracy for Decision Tree Model (without sklearn) is : ', acc,file=f)


## Note: Don't use sklearn.externals.six, need to use 0.20.3, pip install --upgrade scik
```

Solution:

```
   Age Sex      BP Cholesterol  Na_to_K   Drug
0   23   F    HIGH        HIGH   25.355  drugY
1   47   M     LOW        HIGH   13.093  drugC
2   47   M     LOW        HIGH   10.114  drugC
3   28   F  NORMAL        HIGH    7.798  drugX
4   61   F     LOW        HIGH   18.043  drugY
[[23 'F' 'HIGH' 'HIGH' 25.355]
 [47 'M' 'LOW' 'HIGH' 13.093]
 [47 'M' 'LOW' 'HIGH' 10.113999999999999]
```

```
 [28 'F' 'NORMAL' 'HIGH' 7.797999999999999]
 [61 'F' 'LOW' 'HIGH' 18.043]]
[[23 0 0 0 25.355]
 [47 1 1 0 13.093]
 [47 1 1 0 10.113999999999999]
 [28 0 2 0 7.797999999999999]
 [61 0 1 0 18.043]]
0    drugY
1    drugC
2    drugC
3    drugX
4    drugY
Name: Drug, dtype: object
Train Set:  (140, 5) (140,)
Test Set:  (60, 5) (60,)
Prediction Set:  ['drugY' 'drugX' 'drugX' 'drugX' 'drugX']
Y Test Set:  40     drugY
51     drugX
139    drugX
197    drugX
170    drugX
Name: Drug, dtype: object
Accuracy for Decision Tree Model (with sklearn) is :  0.9833333333333333
Accuracy for Decision Tree Model (without sklearn) is :  0.9833333333333333
```