In [1]:

```
import pandas as pd
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

executed in 29.1s, finished 18:26:20 2022-01-05

In [2]:

```
sales_data=pd.read_csv("Company_Data.csv")
sales_data.head()
```

executed in 350ms, finished 18:26:33 2022-01-05

Out[2]:

	Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Age	Education	Urba
0	9.50	138	73	11	276	120	Bad	42	17	Y
1	11.22	111	48	16	260	83	Good	65	10	Y
2	10.06	113	35	10	269	80	Medium	59	12	Y
3	7.40	117	100	4	466	97	Medium	55	14	Y
4	4.15	141	64	3	340	128	Bad	38	13	Y

Initial investigation

In [3]:

```
sales_data.shape
```

Out[3]:

```
(400, 11)
```

In [4]:

```
sales_data.dtypes
```

Out[4]:

```
Sales          float64
CompPrice        int64
Income           int64
Advertising      int64
Population       int64
Price            int64
ShelveLoc       object
Age              int64
Education        int64
Urban           object
US              object
dtype: object
```

In [5]:

```
sales_data.isnull().sum()
```

Out[5]:

```
Sales          0
CompPrice      0
Income         0
Advertising    0
Population     0
Price          0
ShelveLoc      0
Age            0
Education      0
Urban          0
US             0
dtype: int64
```

In [6]:

```
sales_data.describe()
```

Out[6]:

	Sales	CompPrice	Income	Advertising	Population	Price	Age	E
count	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	4(
mean	7.496325	124.975000	68.657500	6.635000	264.840000	115.795000	53.322500	1
std	2.824115	15.334512	27.986037	6.650364	147.376436	23.676664	16.200297	
min	0.000000	77.000000	21.000000	0.000000	10.000000	24.000000	25.000000	1
25%	5.390000	115.000000	42.750000	0.000000	139.000000	100.000000	39.750000	1
50%	7.490000	125.000000	69.000000	5.000000	272.000000	117.000000	54.500000	1
75%	9.320000	135.000000	91.000000	12.000000	398.500000	131.000000	66.000000	1
max	16.270000	175.000000	120.000000	29.000000	509.000000	191.000000	80.000000	1

Number of features and records in the given data set is 11 and 400 respesctively

There is no null values in the data set

The categorical data can be converted into numeric data type by using encoder so that the model can learn the things more easily

Data preparation

In [7]:

```
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
```

In [8]:

```
sales_data['ShelveLoc'].unique(),sales_data['Urban'].unique(),sales_data['US'].unique()
```

Out[8]:

```
(array(['Bad', 'Good', 'Medium'], dtype=object),
 array(['Yes', 'No'], dtype=object),
 array(['Yes', 'No'], dtype=object))
```

In [9]:

```
sales_data['ShelveLoc']=le.fit_transform(sales_data['ShelveLoc'])
sales_data['Urban']=le.fit_transform(sales_data['Urban'])
sales_data['US']=le.fit_transform(sales_data['US'])
sales_data.dtypes
```

Out[9]:

```
Sales          float64
CompPrice        int64
Income           int64
Advertising      int64
Population       int64
Price            int64
ShelveLoc        int32
Age              int64
Education        int64
Urban            int32
US               int32
dtype: object
```

Converting sales to category of high,medium and low sales

In [10]:

```
sales_data.insert(11,'sales_category','')
sales_data
```

Out[10]:

	Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Age	Education	Urban	US	sal
0	9.50	138	73	11	276	120	0	42	17	1	1	
1	11.22	111	48	16	260	83	1	65	10	1	1	
2	10.06	113	35	10	269	80	2	59	12	1	1	
3	7.40	117	100	4	466	97	2	55	14	1	1	
4	4.15	141	64	3	340	128	0	38	13	1	0	
...	...	...	...	...	...	...	...	...	...	...	...	
395	12.57	138	108	17	203	128	1	33	14	1	1	
396	6.14	139	23	3	37	120	2	55	11	0	1	
397	7.41	162	26	12	368	159	2	40	18	1	1	
398	5.94	100	79	7	284	95	0	50	12	1	1	

In [11]:

```
for i in range(0,len(sales_data['Sales']),1):
    if sales_data['Sales'][i]>=11.0:
        sales_data["sales_category"][i]='high'
    elif sales_data['Sales'][i]<=6.0:
        sales_data['sales_category'][i]='Low'
    else:
        sales_data['sales_category'][i]='Medium'
```

In [12]:

```
sales_data['sales_category'].nunique()
```

Out[12]:

3

In [13]:

```
sales_data.head()
```

Out[13]:

	Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Age	Education	Urba
0	9.50	138	73	11	276	120	0	42	17	
1	11.22	111	48	16	260	83	1	65	10	
2	10.06	113	35	10	269	80	2	59	12	
3	7.40	117	100	4	466	97	2	55	14	
4	4.15	141	64	3	340	128	0	38	13	

Model building

In [14]:

```
x=sales_data.iloc[:,1:11]
y=sales_data.iloc[:,11:12]
```

In [15]:

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2)
```

Model training before selecting important feature

In [16]:

```
from sklearn.tree import DecisionTreeClassifier
dt_model=DecisionTreeClassifier()
```

In [17]:

```
dt_model.fit(x_train,y_train)
y_pred=dt_model.predict(x_test)
```

In [18]:

```
from sklearn.metrics import classification_report,confusion_matrix,accuracy_score
```

In [19]:

```
print(classification_report(y_test,y_pred))
```

```
              precision    recall  f1-score   support

         Low       0.78      0.62      0.69        29
      Medium       0.68      0.78      0.73        41
        high       0.60      0.60      0.60        10

    accuracy                           0.70        80
   macro avg       0.69      0.67      0.67        80
weighted avg       0.71      0.70      0.70        80
```

In [20]:

```
print(accuracy_score(y_test,y_pred))
```

```
0.7
```

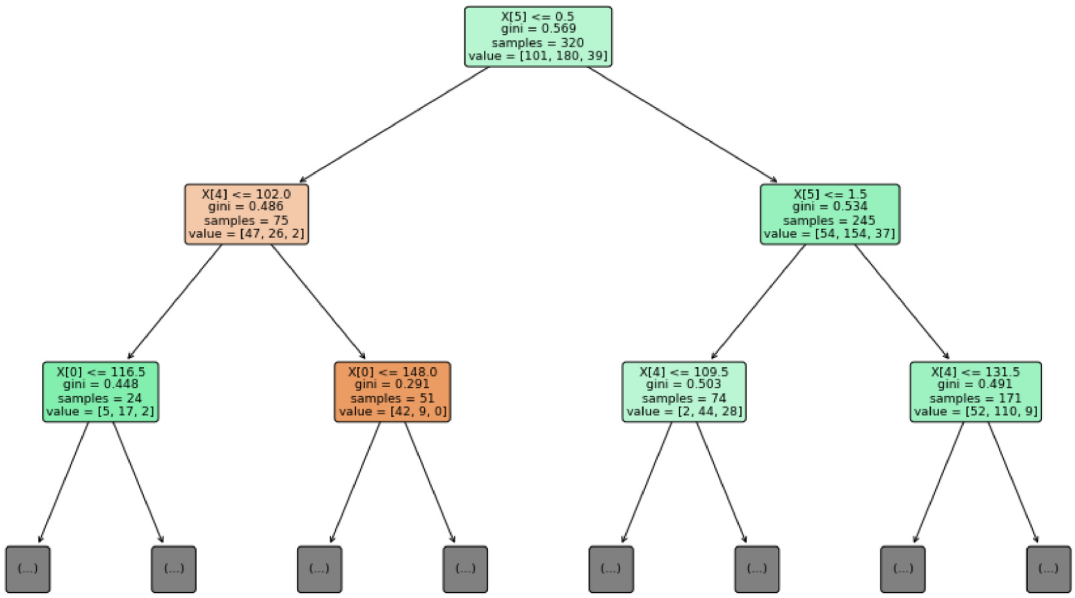
In [21]:

```
print(confusion_matrix(y_test,y_pred))
```

```
[[18 11  0]
 [ 5 32  4]
 [ 0  4  6]]
```

In [26]:

```
from sklearn import tree
plt.figure(figsize=(16,10))
tree.plot_tree(dt_model,rounded=True,filled=True,max_depth=2)
plt.show()
```



Hyperparameter tweaking by using GridsearchCV

In [22]:

```
from sklearn.model_selection import GridSearchCV
grid_model=GridSearchCV(estimator = dt_model,param_grid={'criterion':['entropy','gini'],
                                                         'max_depth':[2,4,8,10],
                                                          'min_samples_split':[2,4,6,8],
                                                          'min_samples_leaf':[1,2,3,4]})
grid_model.fit(x_train,y_train)
print(grid_model.best_params_)
print(grid_model.best_score_)
```

```
{'criterion': 'gini', 'max_depth': 4, 'min_samples_leaf': 4, 'min_samples_sp
lit': 2}
0.665625
```

In [24]:

```
dt_model_cv=DecisionTreeClassifier(max_depth=4,min_samples_leaf=4)
dt_model_cv.fit(x_train,y_train)
y_pred=dt_model_cv.predict(x_test)
print(accuracy_score(y_test,y_pred))
```

```
0.65
```

Feature importance plot

In [26]:

```
feature_dt=x_train.columns
len(feature_dt)
```

Out[26]:

10

In [28]:

```
imp_feature_dt=dt_model.feature_importances_
len(imp_feature_dt)
```

Out[28]:

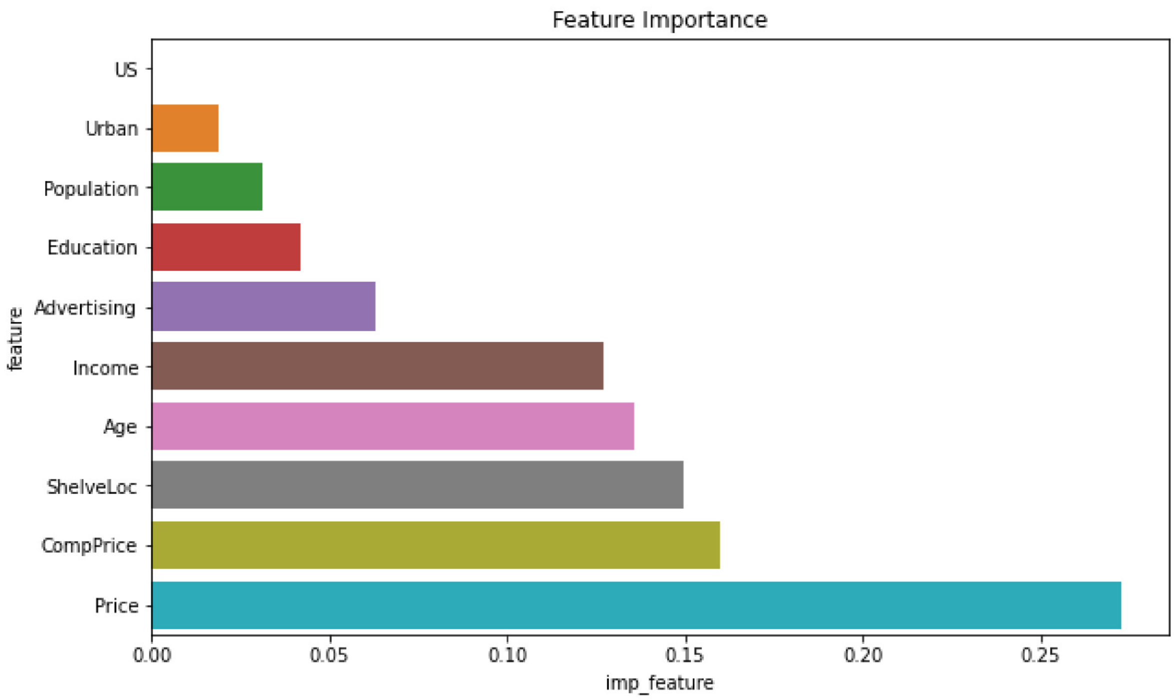
10

In [31]:

```
data_dt=pd.DataFrame({'feature':feature_dt,'imp_feature':imp_feature_dt})
data_dt=data_dt.sort_values('imp_feature')
```

In [32]:

```
import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(10,6))
plt.title('Feature Importance')
sns.barplot(y='feature', x='imp_feature', data=data_dt)
plt.show()
```



The plot clearly shows that some of the features donot contribute much for model building

Hence the accuracy can be imporved by droping those insignificant features

Feature selection by feature importance for Decision tree

In [33]:

```
from sklearn.feature_selection import SelectFromModel
```

In [34]:

```
selector=SelectFromModel(estimator=DecisionTreeClassifier())
selector.fit(x_train,y_train)
```

Out[34]:

```
SelectFromModel(estimator=DecisionTreeClassifier())
```

In [35]:

```
selector.get_support()
```

Out[35]:

```
array([ True,  True, False, False,  True,  True,  True, False, False,
       False])
```

In [36]:

```
len(x_train.columns)
```

Out[36]:

```
10
```

In [37]:

```
feature=x_train.columns[selector.get_support()]
```

In [38]:

```
len(x_train.columns[selector.get_support()])
```

Out[38]:

```
5
```

Out of 10 features only 5 features is selscted for building models

In [39]:

```
x_train_dt=selector.transform(x_train)
x_test_dt=selector.transform(x_test)
```

In [40]:

```
dt_model_imp=DecisionTreeClassifier().fit(x_train_dt,y_train)
y_pred_imp=dt_model_imp.predict(x_test_dt)
```

In [41]:

```
print(accuracy_score(y_test,y_pred_imp))
```

```
0.6875
```

In [42]:

```
print(confusion_matrix(y_test,y_pred_imp))
```

```
[[16  6  0]
 [11 36  4]
 [ 0  4  3]]
```

In [43]:

```
print(classification_report(y_test,y_pred_imp))
```

```
              precision    recall  f1-score   support

         Low       0.59      0.73      0.65        22
      Medium       0.78      0.71      0.74        51
        high       0.43      0.43      0.43         7

    accuracy                           0.69        80
   macro avg       0.60      0.62      0.61        80
weighted avg       0.70      0.69      0.69        80
```

In [44]:

```
imp_feature=dt_model_imp.feature_importances_
imp_feature
```

Out[44]:

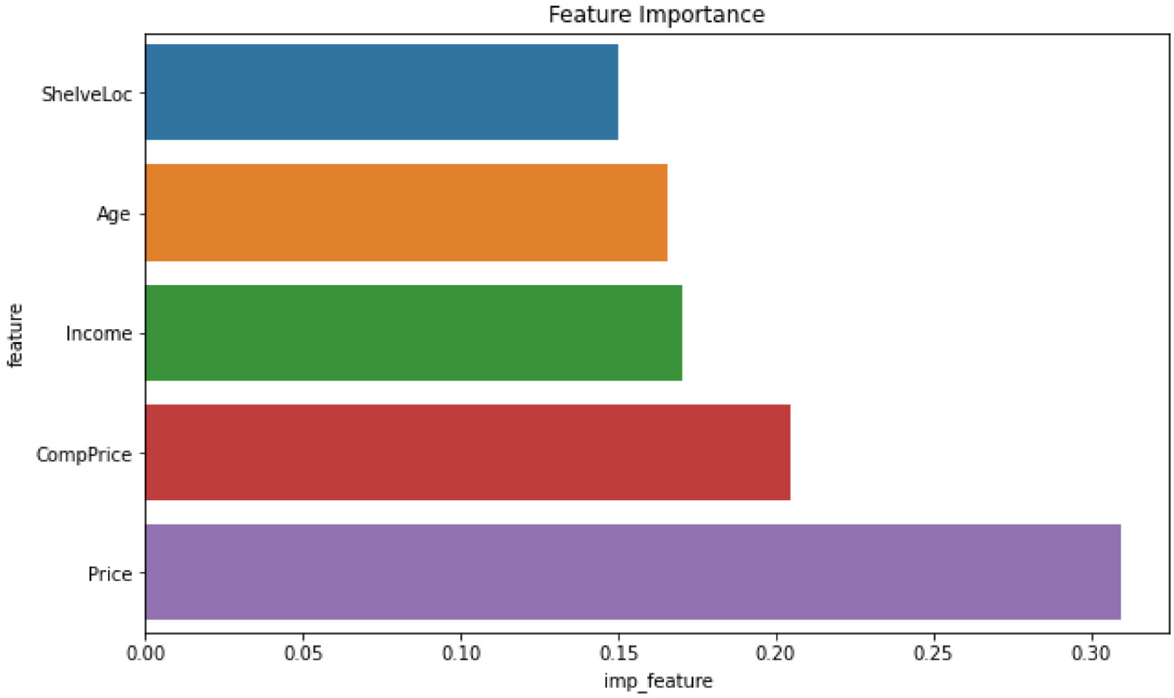
```
array([0.20483117, 0.17044253, 0.30939202, 0.1497078 , 0.16562648])
```

In [45]:

```
data_imp=pd.DataFrame({'feature':feature,'imp_feature':imp_feature})
data_imp=data_imp.sort_values('imp_feature')
```

In [46]:

```
import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(10,6))
plt.title('Feature Importance')
sns.barplot(y='feature', x='imp_feature', data=data_imp)
plt.show()
```



Recursive feature elimination

In [47]:

```
from sklearn.feature_selection import RFE
```

In [48]:

```
selector_rfe=RFE(DecisionTreeClassifier())
selector_rfe.fit(x_train,y_train)
```

Out[48]:

```
RFE(estimator=DecisionTreeClassifier())
```

In [49]:

```
selector_rfe.get_support()
```

Out[49]:

```
array([ True,  True, False, False,  True,  True,  True, False, False,
       False])
```

In [50]:

```
feature_rfe=x_train.columns[selector_rfe.get_support()]
feature_rfe
```

Out[50]:

```
Index(['CompPrice', 'Income', 'Price', 'ShelveLoc', 'Age'], dtype='object')
```

In [51]:

```
len(x_train.columns[selector_rfe.get_support()])
```

Out[51]:

```
5
```

Here 5 out of 10 feature is selscted as an important feature

In [52]:

```
x_train_rfe=selector_rfe.transform(x_train)
x_test_rfe=selector_rfe.transform(x_test)
```

In [53]:

```
dt_model_rfe=DecisionTreeClassifier().fit(x_train_rfe,y_train)
```

In [54]:

```
y_pred_rfe=dt_model_rfe.predict(x_test_rfe)
```

In [55]:

```
print(accuracy_score(y_test,y_pred_rfe))
```

```
0.7
```

In [56]:

```
print(confusion_matrix(y_test,y_pred_rfe))
```

```
[[16  6  0]
 [ 9 37  5]
 [ 0  4  3]]
```

In [57]:

```
print(classification_report(y_test,y_pred_rfe))
```

```
              precision    recall  f1-score   support

         Low       0.64      0.73      0.68        22
      Medium       0.79      0.73      0.76        51
        high       0.38      0.43      0.40         7

    accuracy                           0.70        80
   macro avg       0.60      0.63      0.61        80
weighted avg       0.71      0.70      0.70        80
```

In [58]:

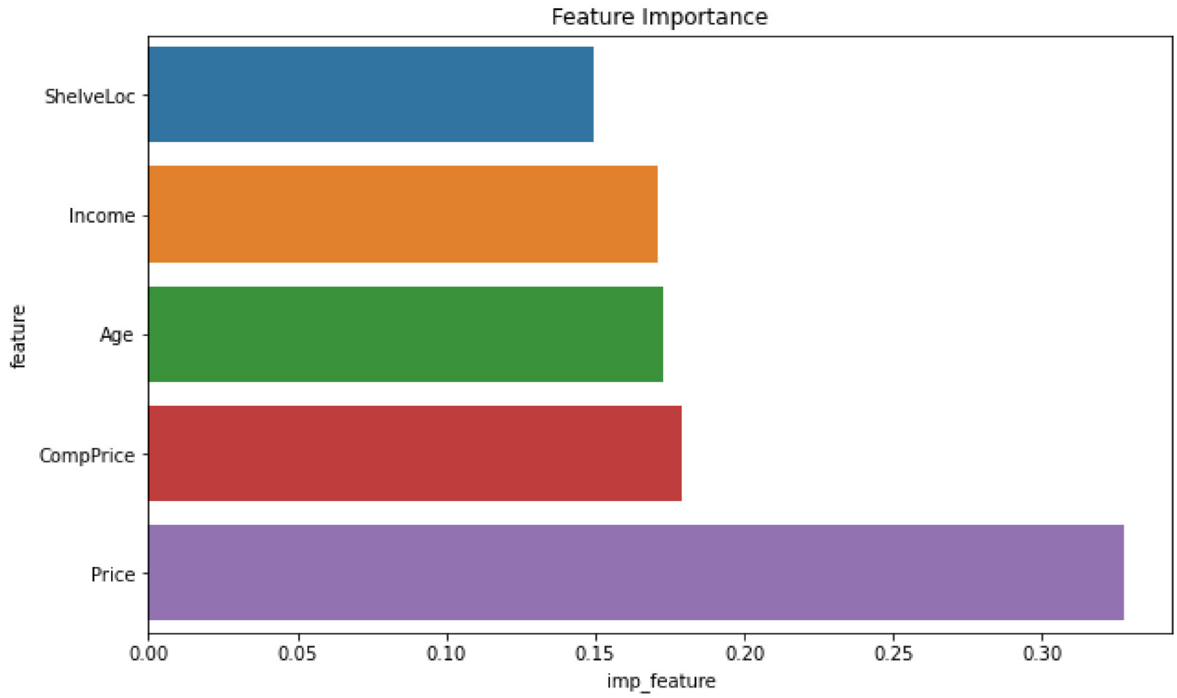
```
imp_feature_rfe=dt_model_rfe.feature_importances_
```

In [59]:

```
data_rfe=pd.DataFrame({'feature':feature_rfe,'imp_feature':imp_feature_rfe})
data_rfe=data_rfe.sort_values('imp_feature')
```

In [60]:

```
plt.figure(figsize=(10,6))
plt.title('Feature Importance')
sns.barplot(y='feature', x='imp_feature', data=data_rfe)
plt.show()
```



Inference

Price is the feature which affect/contibute more for the sales

Competerior price fallows price which affect the most