In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

import warnings
warnings.simplefilter('ignore')
```

In [2]:
```
df=pd.read_csv('penguins_size.csv')
df.head()
```

Out[2]:

	species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	FE
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	FE
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	FE

In [3]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 344 entries, 0 to 343
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   species            344 non-null    object 
 1   island             344 non-null    object 
 2   culmen_length_mm   342 non-null    float64
 3   culmen_depth_mm    342 non-null    float64
 4   flipper_length_mm  342 non-null    float64
 5   body_mass_g        342 non-null    float64
 6   sex                334 non-null    object 
dtypes: float64(4), object(3)
memory usage: 18.9+ KB
```

In [4]:
```
df.isnull().sum()
```

Out[4]:
```
species              0
island               0
culmen_length_mm     2
culmen_depth_mm      2
flipper_length_mm    2
body_mass_g          2
sex                 10
dtype: int64
```

In [5]:
```
df.shape
```

Out[5]:
```
(344, 7)
```

In [6]:
```
df['species'].unique()
```

Out[6]:
```
array(['Adelie', 'Chinstrap', 'Gentoo'], dtype=object)
```

In [7]:
```
df['island'].unique()
```

Out[7]:
```
array(['Torgersen', 'Biscoe', 'Dream'], dtype=object)
```

In [8]:
```
df['sex'].value_counts()
```

Out[8]:
```
MALE      168
FEMALE    165
.           1
Name: sex, dtype: int64
```

In [9]:
```
df=df[df['sex']!='.']
df.shape
```

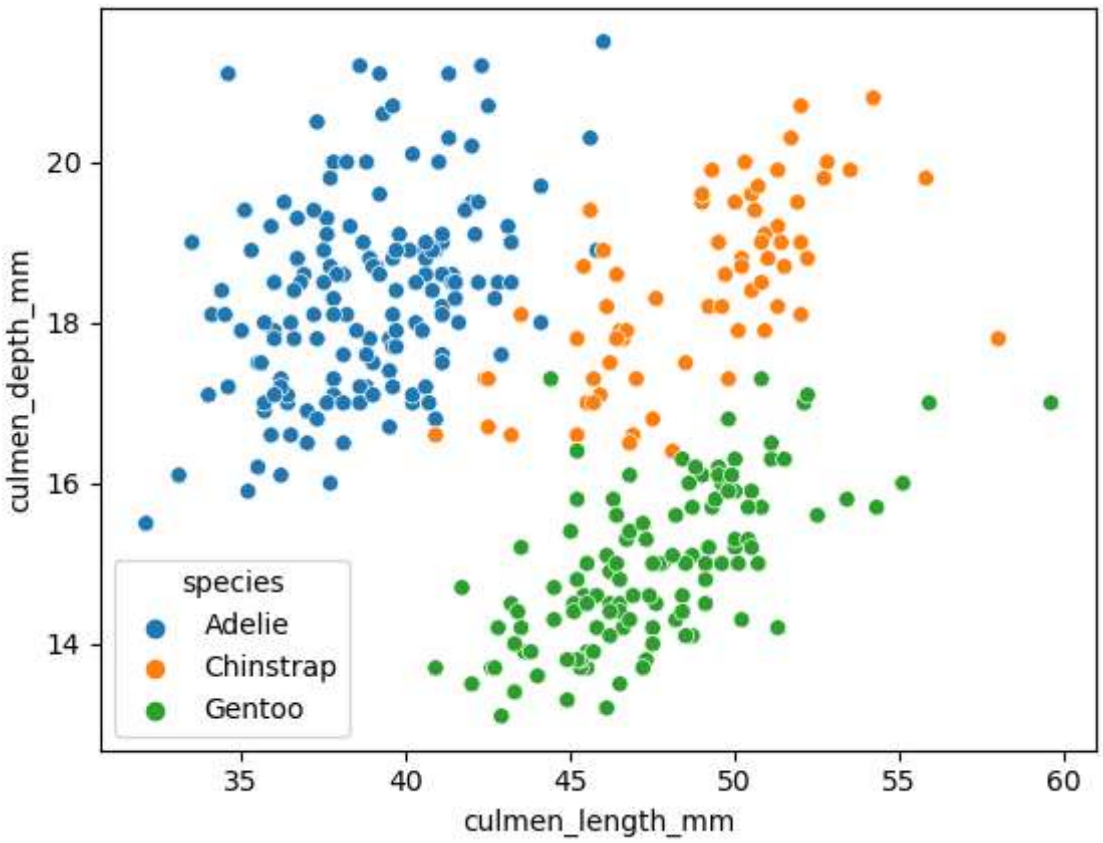
Out[9]:
```
(343, 7)
```

In [10]:
```
df.isnull().sum()
```

Out[10]:
```
species               0
island                0
culmen_length_mm      2
culmen_depth_mm       2
flipper_length_mm     2
body_mass_g           2
sex                  10
dtype: int64
```

In [11]:
```
sns.scatterplot(x='culmen_length_mm',y='culmen_depth_mm',data=df,hue='species')
```

Out[11]:
```
<AxesSubplot:xlabel='culmen_length_mm', ylabel='culmen_depth_mm'>
```



In [12]:
```
df=df.dropna()
df.shape
```

Out[12]:
```
(333, 7)
```

In [13]:
```
df1=df.drop('species',axis=1)
```

In [14]:
```
df1.head()
```

Out[14]:

	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex
0	Torgersen	39.1	18.7	181.0	3750.0	MALE
1	Torgersen	39.5	17.4	186.0	3800.0	FEMALE
2	Torgersen	40.3	18.0	195.0	3250.0	FEMALE
4	Torgersen	36.7	19.3	193.0	3450.0	FEMALE
5	Torgersen	39.3	20.6	190.0	3650.0	MALE

In [15]:
```
x=pd.get_dummies(df1,drop_first=True)
```

In [16]:
```
x.head()
```

Out[16]:

	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	island_Dream	island_
0	39.1	18.7	181.0	3750.0	0	
1	39.5	17.4	186.0	3800.0	0	
2	40.3	18.0	195.0	3250.0	0	
4	36.7	19.3	193.0	3450.0	0	
5	39.3	20.6	190.0	3650.0	0	

In [17]:
```
x.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 333 entries, 0 to 343
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   culmen_length_mm   333 non-null    float64
 1   culmen_depth_mm    333 non-null    float64
 2   flipper_length_mm  333 non-null    float64
 3   body_mass_g        333 non-null    float64
 4   island_Dream       333 non-null    uint8  
 5   island_Torgersen   333 non-null    uint8  
 6   sex_MALE           333 non-null    uint8  
dtypes: float64(4), uint8(3)
memory usage: 14.0 KB
```

In [18]:
```
df['island'].value_counts()
```

Out[18]:
```
Biscoe       163
Dream        123
Torgersen     47
Name: island, dtype: int64
```

In [19]:
```
y=df['species']
```

In [20]:
```
y.info()
```

```
<class 'pandas.core.series.Series'>
Int64Index: 333 entries, 0 to 343
Series name: species
Non-Null Count  Dtype
--------------  -----
333 non-null    object
dtypes: object(1)
memory usage: 5.2+ KB
```

In [21]:
```
y.value_counts()
```

Out[21]:
```
Adelie       146
Gentoo       119
Chinstrap     68
Name: species, dtype: int64
```

In [22]:
```
x.head()
```

Out[22]:

	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	island_Dream	island_
0	39.1	18.7	181.0	3750.0	0	
1	39.5	17.4	186.0	3800.0	0	
2	40.3	18.0	195.0	3250.0	0	
4	36.7	19.3	193.0	3450.0	0	
5	39.3	20.6	190.0	3650.0	0	

In [23]:
```
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=42)
```

In [24]:
```
from sklearn.tree import DecisionTreeClassifier
dt_default=DecisionTreeClassifier(random_state=0)
dt_default.fit(x_train,y_train)

train_pred=dt_default.predict(x_train)
test_pred=dt_default.predict(x_test)
```

In [25]:
```
dt_default.score(x_train,y_train)
dt_default.score(x_test,y_test)
```

Out[25]:
```
0.98
```

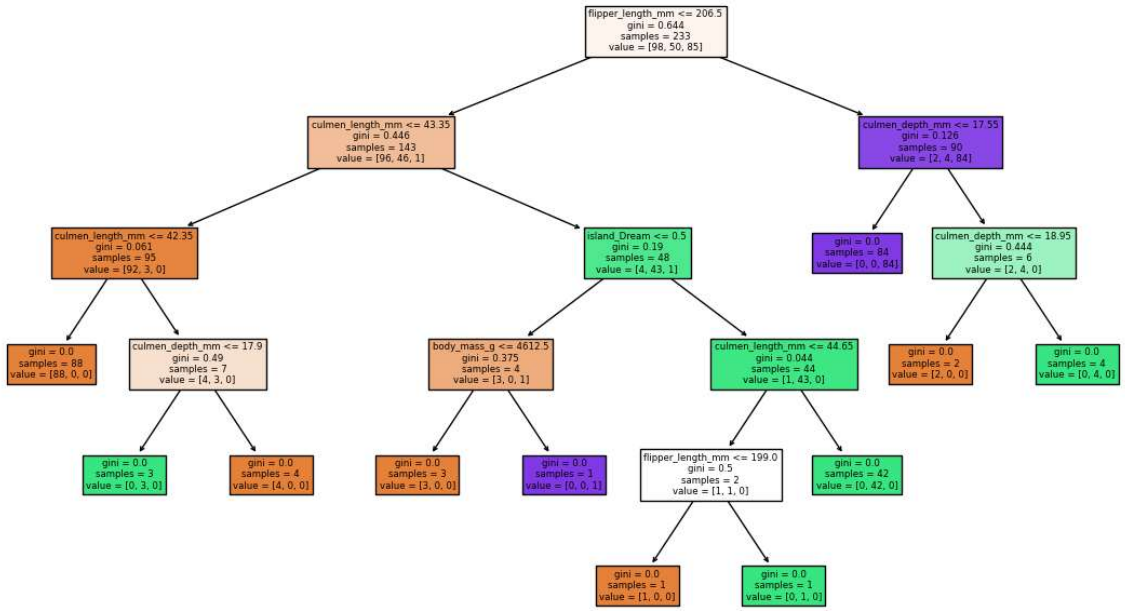
In [26]:
```
from sklearn.model_selection import cross_val_score
scores=cross_val_score(dt_default,x,y,cv=5)
scores.mean()
```

Out[26]:
```
0.9698778833107191
```

In [27]:
```
dt_default.predict([[39.1,18.7,181.0,3750.0,0,1,1]])
```

Out[27]:
```
array(['Adelie'], dtype=object)
```

In [28]:
```
from sklearn.tree import plot_tree
plt.figure(figsize=(14,8),dpi=100)
plot_tree(dt_default,filled=True,feature_names=x.columns)
plt.show()
```



In [29]:
```
dt_default.feature_importances_
```

Out[29]:
```
array([0.34756206, 0.09868076, 0.50596782, 0.00999714, 0.03779222,
       0.        , 0.        ])
```

In [30]:
```
pd.DataFrame(index=x.columns,data=dt_default.feature_importances_)
```

Out[30]:

	0
culmen_length_mm	0.347562
culmen_depth_mm	0.098681
flipper_length_mm	0.505968
body_mass_g	0.009997
island_Dream	0.037792
island_Torgersen	0.000000
sex_MALE	0.000000

In [31]:
```
x.head()
```

Out[31]:

	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	island_Dream	island_
0	39.1	18.7	181.0	3750.0	0	
1	39.5	17.4	186.0	3800.0	0	
2	40.3	18.0	195.0	3250.0	0	
4	36.7	19.3	193.0	3450.0	0	
5	39.3	20.6	190.0	3650.0	0	

In [32]:
```
x.head()
```

Out[32]:

	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	island_Dream	island_
0	39.1	18.7	181.0	3750.0	0	
1	39.5	17.4	186.0	3800.0	0	
2	40.3	18.0	195.0	3250.0	0	
4	36.7	19.3	193.0	3450.0	0	
5	39.3	20.6	190.0	3650.0	0	

In [33]:
```
x=x.drop(['island_Torgersen','sex_MALE'],axis=1)
```

In [34]:
```
x.head()
```

Out[34]:

	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	island_Dream
0	39.1	18.7	181.0	3750.0	0
1	39.5	17.4	186.0	3800.0	0
2	40.3	18.0	195.0	3250.0	0
4	36.7	19.3	193.0	3450.0	0
5	39.3	20.6	190.0	3650.0	0

In [44]:
```
from sklearn.model_selection import GridSearchCV

estimator=DecisionTreeClassifier(random_state=0)

param_grid={'criterion':['gini','entropy'],
            'max_depth':[1,2,3,4]}

grid=GridSearchCV(estimator,param_grid,scoring='accuracy',cv=5)

grid.fit(x_train,y_train)

grid.best_params_
```

Out[44]:
```
{'criterion': 'gini', 'max_depth': 4}
```

In [46]:
```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3,random_state=0)
dt_hp=DecisionTreeClassifier(criterion='gini',max_depth=3,random_state=0)
dt_hp.fit(x_train,y_train)

train_pred=dt_hp.predict(x_train)

test_pred=dt_hp.predict(x_test)

dt_hp.score(x_train,y_train)
dt_hp.score(x_test,y_test)
```

Out[46]:
```
0.95
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: