

Decision Tree

```
In [ ]: import pandas as pd, numpy as np, matplotlib.pyplot as plt
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import OneHotEncoder
from sklearn.preprocessing import LabelEncoder
import warnings
warnings.filterwarnings("ignore")
```

```
In [ ]: file_path = "C:/Users/Owner/source/vsc_repo/machine_learn_cookbook/Logistic_Regress
hr_data = pd.read_csv(file_path, sep=",", engine="python", encoding="utf-8", encoding="utf-8", encoding="utf-8")
hr_data.rename(columns={'Departments ': 'Departments', 'salary' : 'Salary' }, inplace=True)
# hr_data.rename(str.title, axis='columns', inplace=True)
hr_data['Departments'] = [s.title() for s in hr_data['Departments']]
# hr_data.drop('Departments ', axis=1, inplace=True)
hr_data['Salary'] = [s.title() for s in hr_data['Salary']]
hr_data['Departments'] = hr_data['Departments'].replace(["Hr", "It", "Mng", "Randd"])
```

Exploratory Data Analysis

```
In [ ]: hr_data.head(5)
```

Out[]:

	Satisfaction Level	Last Evaluation	Number of Projects	Monthly Hours	Total Time at the Company	Work Accidents	Quit the Company	Promoted in Last 5 yrs
0	0.38	0.53	2	157	3	0	1	0
1	0.80	0.86	5	262	6	0	1	0
2	0.11	0.88	7	272	4	0	1	0
3	0.72	0.87	5	223	5	0	1	0
4	0.37	0.52	2	159	3	0	1	0



```
In [ ]: hr_data.tail(5)
```

```
Out[ ]:
```

	Satisfaction Level	Last Evaluation	Number of Projects	Monthly Hours	Total Time at the Company	Work Accidents	Quit the Company	Promoted in Last 5 yrs
14994	0.40	0.57	2	151	3	0	1	
14995	0.37	0.48	2	160	3	0	1	
14996	0.37	0.53	2	143	3	0	1	
14997	0.11	0.96	6	280	4	0	1	
14998	0.37	0.52	2	158	3	0	1	



```
In [ ]: hr_data.sample(n=5, random_state=3)
```

```
Out[ ]:
```

	Satisfaction Level	Last Evaluation	Number of Projects	Monthly Hours	Total Time at the Company	Work Accidents	Quit the Company	Promoted in Last 5 yrs
11584	0.65	0.74	4	233	4	1	0	
5034	0.47	0.51	6	190	5	0	0	
12282	0.83	0.92	5	267	6	0	1	
28	0.41	0.46	2	128	3	0	1	
9702	0.54	0.72	6	222	5	0	0	



```
In [ ]: hr_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14999 entries, 0 to 14998
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Satisfaction Level    14999 non-null   float64 
 1   Last Evaluation      14999 non-null   float64 
 2   Number of Projects   14999 non-null   int64  
 3   Monthly Hours        14999 non-null   int64  
 4   Total Time at the Company 14999 non-null   int64  
 5   Work Accidents       14999 non-null   int64  
 6   Quit the Company     14999 non-null   int64  
 7   Promoted in Last 5 yrs 14999 non-null   int64  
 8   Departments          14999 non-null   object  
 9   Salary               14999 non-null   object  
 10  Management           14999 non-null   int64  
dtypes: float64(2), int64(7), object(2)
memory usage: 1.3+ MB
```

```
In [ ]: hr_data.dtypes
```

```
Out[ ]: Satisfaction Level      float64
         Last Evaluation       float64
         Number of Projects     int64
         Monthly Hours          int64
         Total Time at the Company int64
         Work Accidents         int64
         Quit the Company        int64
         Promoted in Last 5 yrs  int64
         Departments             object
         Salary                  object
         Management              int64
         dtype: object
```

```
In [ ]: hr_data.index
```

```
Out[ ]: RangeIndex(start=0, stop=14999, step=1)
```

```
In [ ]: hr_data.shape
```

```
Out[ ]: (14999, 11)
```

```
In [ ]: hr_data.ndim
```

```
Out[ ]: 2
```

```
In [ ]: hr_data.columns
```

```
Out[ ]: Index(['Satisfaction Level', 'Last Evaluation', 'Number of Projects',
               'Monthly Hours', 'Total Time at the Company', 'Work Accidents',
               'Quit the Company', 'Promoted in Last 5 yrs', 'Departments', 'Salary',
               'Management'],
               dtype='object')
```

```
In [ ]: hr_data["Departments"].unique()
```

```
Out[ ]: array(['Sales', 'Accounting', 'HR', 'Technical', 'Support', 'Management',
               'IT', 'Product_MNG', 'Marketing', 'R&D'], dtype=object)
```

```
In [ ]: hr_data["Salary"].unique()
```

```
Out[ ]: array(['Low', 'Medium', 'High'], dtype=object)
```

```
In [ ]: hr_data.describe()
```

```
Out[ ]:
```

	Satisfaction Level	Last Evaluation	Number of Projects	Monthly Hours	Total Time at the Company	Work Accidents
count	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000
mean	0.612834	0.716102	3.803054	201.050337	3.498233	0.144610
std	0.248631	0.171169	1.232592	49.943099	1.460136	0.351719
min	0.090000	0.360000	2.000000	96.000000	2.000000	0.000000
25%	0.440000	0.560000	3.000000	156.000000	3.000000	0.000000
50%	0.640000	0.720000	4.000000	200.000000	3.000000	0.000000
75%	0.820000	0.870000	5.000000	245.000000	4.000000	0.000000
max	1.000000	1.000000	7.000000	310.000000	10.000000	1.000000



```
In [ ]: hr_data["Quit the Company"].loc[(hr_data["Departments"] == "IT") & (hr_data["Salary"]
```

```
Out[ ]: Quit the Company
0    437
1    172
Name: count, dtype: int64
```

```
In [ ]: hr_data["Quit the Company"].loc[(hr_data["Departments"] == "IT") & (hr_data["Salary"]
```

```
Out[ ]: Quit the Company
0    438
1     97
Name: count, dtype: int64
```

```
In [ ]: hr_data["Quit the Company"].loc[(hr_data["Departments"] == "IT") & (hr_data["Salary"]
```

```
Out[ ]: Quit the Company
0    79
1     4
Name: count, dtype: int64
```

Preprocessing

```
In [ ]: he = OneHotEncoder(categories="auto", drop="first", handle_unknown="error")
hr_encode = he.fit_transform(hr_data)
hr_dummies = pd.get_dummies(hr_data, prefix_sep="_", dummy_na=False, dtype=int, drop=True)
hr_dummies.head(3)
```

```
Out[ ]:
```

	Satisfaction Level	Last Evaluation	Number of Projects	Monthly Hours	Total Time at the Company	Work Accidents	Quit the Company	Promoted in Last 5 yrs
0	0.38	0.53	2	157	3	0	1	0
1	0.80	0.86	5	262	6	0	1	0
2	0.11	0.88	7	272	4	0	1	0

```
In [ ]:
```

```
le = LabelEncoder()
hr_data["Departments_Encode"] = le.fit_transform(hr_data["Departments"])
hr_data["Salary_Encode"] = le.fit_transform(hr_data["Salary"])
hr_data.head(5)
```

```
Out[ ]:
```

	Satisfaction Level	Last Evaluation	Number of Projects	Monthly Hours	Total Time at the Company	Work Accidents	Quit the Company	Promoted in Last 5 yrs
0	0.38	0.53	2	157	3	0	1	0
1	0.80	0.86	5	262	6	0	1	0
2	0.11	0.88	7	272	4	0	1	0
3	0.72	0.87	5	223	5	0	1	0
4	0.37	0.52	2	159	3	0	1	0

```
In [ ]:
```

```
X = hr_dummies.drop("Quit the Company", axis=1)
y = hr_dummies["Quit the Company"]
```

```
In [ ]:
```

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.33, random_st
```

Decision Tree Classifier

```
In [ ]:
```

```
from sklearn.model_selection import GridSearchCV
param_grid = {"criterion" : ["gini", "entropy", "log_loss"], "max_depth" : [6, 8, 10],
              "random_state" : [0, 42], }
grid_search = GridSearchCV(DecisionTreeClassifier(), param_grid, cv=5, scoring="accuracy")
grid_search.fit(x_train, y_train)
print(grid_search.best_params_)

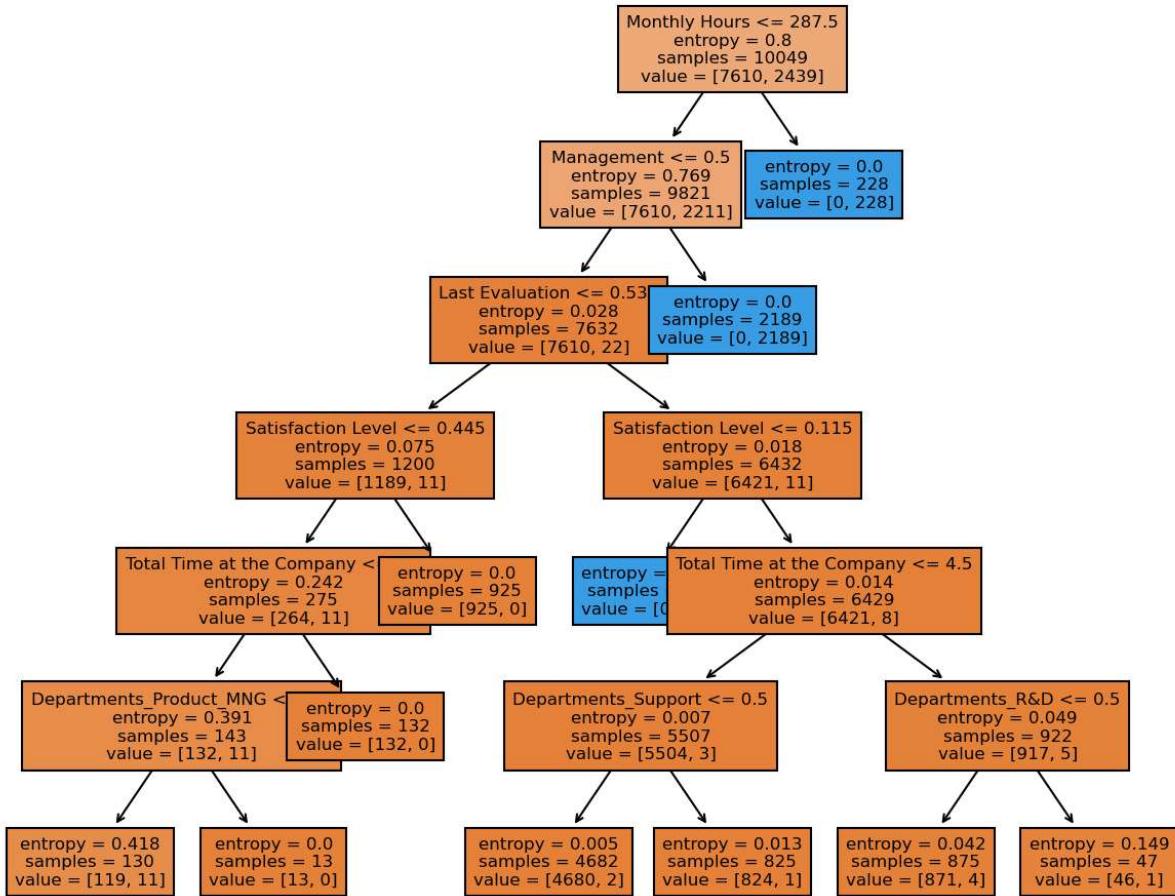
dtc = DecisionTreeClassifier(criterion="entropy", max_depth=6, max_features="auto",
                           random_state=0)
dtc.fit(x_train, y_train)
y_predict = dtc.predict(x_test)
```



```
6, random_state=42;, score=0.991 total time= 0.0s
[CV 4/5] END criterion=log_loss, max_depth=12, max_features=log2, min_samples_split=
6, random_state=42;, score=0.986 total time= 0.0s
[CV 5/5] END criterion=log_loss, max_depth=12, max_features=log2, min_samples_split=
6, random_state=42;, score=0.996 total time= 0.0s
[CV 1/5] END criterion=log_loss, max_depth=12, max_features=log2, min_samples_split=
8, random_state=0;, score=0.997 total time= 0.0s
[CV 2/5] END criterion=log_loss, max_depth=12, max_features=log2, min_samples_split=
8, random_state=0;, score=0.999 total time= 0.0s
[CV 3/5] END criterion=log_loss, max_depth=12, max_features=log2, min_samples_split=
8, random_state=0;, score=0.997 total time= 0.0s
[CV 4/5] END criterion=log_loss, max_depth=12, max_features=log2, min_samples_split=
8, random_state=0;, score=0.998 total time= 0.0s
[CV 5/5] END criterion=log_loss, max_depth=12, max_features=log2, min_samples_split=
8, random_state=0;, score=0.995 total time= 0.0s
[CV 1/5] END criterion=log_loss, max_depth=12, max_features=log2, min_samples_split=
8, random_state=42;, score=0.994 total time= 0.0s
[CV 2/5] END criterion=log_loss, max_depth=12, max_features=log2, min_samples_split=
8, random_state=42;, score=0.998 total time= 0.0s
[CV 3/5] END criterion=log_loss, max_depth=12, max_features=log2, min_samples_split=
8, random_state=42;, score=0.998 total time= 0.0s
[CV 4/5] END criterion=log_loss, max_depth=12, max_features=log2, min_samples_split=
8, random_state=42;, score=0.996 total time= 0.0s
[CV 5/5] END criterion=log_loss, max_depth=12, max_features=log2, min_samples_split=
8, random_state=42;, score=0.996 total time= 0.0s
{'criterion': 'entropy', 'max_depth': 6, 'max_features': 'auto', 'min_samples_split': 2, 'random_state': 0}
```

Visualization: Tree Plot

```
In [ ]: classNames=["No", "Quit"]
plt.figure(figsize=(10,8), dpi=150)
tree.plot_tree(dtc, feature_names=X.columns, filled=True, label="all", fontsize=8,
plt.show()
```



Decision Tree for Power BI

```
In [ ]: import pandas as pd, numpy as np, matplotlib.pyplot as plt
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split

file_path = "C:/Users/Owner/source/vsc_repo/machine_learn_cookbook/Logistic_Regress
hr_data = pd.read_csv(file_path, sep=",", engine="python", encoding="utf-8", encoding="utf-8")
hr_data.rename(columns={'Departments ': 'Departments', 'salary' : 'Salary' }, inplace=True)
# hr_data.rename(str.title, axis='columns', inplace=True)
hr_data['Departments'] = [s.title() for s in hr_data['Departments']]
# hr_data.drop('Departments ', axis=1, inplace=True)
hr_data['Salary'] = [s.title() for s in hr_data['Salary']]
hr_data['Departments'] = hr_data['Departments'].replace(["Hr", "It","Mng", "Randd"])

hr_dummies = pd.get_dummies(hr_data, prefix_sep="_", dummy_na=False, dtype=int, dropna=True)

X = hr_dummies.drop("Quit the Company", axis=1)
y = hr_dummies["Quit the Company"]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)

dtc = DecisionTreeClassifier(criterion="entropy", max_depth=6, max_features="auto",
                             min_samples_leaf=1, min_samples_split=2)
```

```

dtc.fit(X_train, y_train)
y_predict = dtc.predict(X_test)

classNames= ["No_Quit", "Quit"]
plt.figure(figsize=(10,8), dpi=150)
tree.plot_tree(dtc, feature_names=X.columns, filled=True, label="all", fontsize=8,
plt.show()

```

