# Decision Tree: Income Prediction

In this lab, we will build a decision tree to predict the income of a given population, which is labelled as <= $50K and >$50K. The attributes (predictors) are age, working class type, marital status, gender, race etc.

In the following sections, we'll:

- clean and prepare the data,
- build a decision tree with default hyperparameters,
- understand all the hyperparameters that we can tune, and finally
- choose the optimal hyperparameters using grid search cross-validation.

---

## Understanding and Cleaning the Data

In [2]:
```
# Importing the required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [3]:
```
# To ignore warnings
import warnings
warnings.filterwarnings("ignore")
```

In [4]:
```
# Reading the csv file and putting it into 'df' object.
df = pd.read_csv('adult_dataset.csv')
```

In [5]:
```
# Let's understand the type of values in each column of our dataframe 'df'.
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
age               32561 non-null int64
workclass         32561 non-null object
fnlwgt            32561 non-null int64
education         32561 non-null object
education.num     32561 non-null int64
marital.status    32561 non-null object
occupation        32561 non-null object
relationship      32561 non-null object
race              32561 non-null object
sex               32561 non-null object
capital.gain      32561 non-null int64
capital.loss      32561 non-null int64
hours.per.week    32561 non-null int64
native.country    32561 non-null object
income            32561 non-null object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
```

In [6]:
```
# Let's understand the data, how it look like.
df.head()
```

Out[6]:

| | age | workclass | fnlwgt | education | education.num | marital.status | occupation | relatio |
|---|---|---|---|---|---|---|---|---|
| **0** | 90 | ? | 77053 | HS-grad | 9 | Widowed | ? | Not-in-f |
| **1** | 82 | Private | 132870 | HS-grad | 9 | Widowed | Exec-managerial | Not-in-f |
| **2** | 66 | ? | 186061 | Some-college | 10 | Widowed | ? | Unmarr |
| **3** | 54 | Private | 140359 | 7th-8th | 4 | Divorced | Machine-op-inspct | Unmarr |
| **4** | 41 | Private | 264663 | Some-college | 10 | Separated | Prof-specialty | Own-ch |

You can observe that the columns workclass and occupation consist of missing values which are represented as '?' in the dataframe.

On looking a bit more closely, you will also find that whenever workclass is having a missing value, occupation is also missing in that row. Let's check how may rows are missing.

In [7]:
```
# rows with missing values represented as'?'.
df_1 = df[df.workclass == '?']
df_1
```

Out[7]:

| | age | workclass | fnlwgt | education | education.num | marital.status | occupation | re |
|---|---|---|---|---|---|---|---|---|
| **0** | 90 | ? | 77053 | HS-grad | 9 | Widowed | ? | Nc |
| **2** | 66 | ? | 186061 | Some-college | 10 | Widowed | ? | Ur |
| **14** | 51 | ? | 172175 | Doctorate | 16 | Never-married | ? | Nc |
| **24** | 61 | ? | 135285 | HS-grad | 9 | Married-civ-spouse | ? | Hu |
| **44** | 71 | ? | 100820 | HS-grad | 9 | Married-civ-spouse | ? | Hu |
| **48** | 68 | ? | 192052 | Some-college | 10 | Married-civ-spouse | ? | Wi |
| **49** | 67 | ? | 174995 | Some-college | 10 | Married-civ-spouse | ? | Hu |
| **76** | 41 | ? | 27187 | Assoc-voc | 11 | Married-civ-spouse | ? | Hu |
| **114** | 72 | ? | 118902 | Doctorate | 16 | Married-civ-spouse | ? | Hu |
| **133** | 65 | ? | 240857 | Bachelors | 13 | Married-civ-spouse | ? | Hu |
| **136** | 68 | ? | 257269 | Bachelors | 13 | Married-civ-spouse | ? | Hu |
| **153** | 43 | ? | 152569 | Assoc-voc | 11 | Widowed | ? | Nc |
| **202** | 65 | ? | 143118 | HS-grad | 9 | Widowed | ? | Ur |
| **213** | 63 | ? | 234083 | HS-grad | 9 | Divorced | ? | Nc |
| **227** | 63 | ? | 83043 | Bachelors | 13 | Married-civ-spouse | ? | Hu |
| **230** | 66 | ? | 177351 | Bachelors | 13 | Married-civ-spouse | ? | Hu |
| **237** | 60 | ? | 141221 | Bachelors | 13 | Married-civ-spouse | ? | Hu |
| **291** | 26 | ? | 131777 | Bachelors | 13 | Married-civ-spouse | ? | Hu |
| **301** | 19 | ? | 241616 | HS-grad | 9 | Never-married | ? | Ur |
| **310** | 55 | ? | 123382 | HS-grad | 9 | Separated | ? | Nc |
| **320** | 21 | ? | 40052 | Some-college | 10 | Never-married | ? | Nc |

| | age | workclass | fnlwgt | education | education.num | marital.status | occupation | re |
|---|---|---|---|---|---|---|---|---|
| **511** | 26 | ? | 370727 | Bachelors | 13 | Married-civ-spouse | ? | Wi |
| **646** | 31 | ? | 85077 | Bachelors | 13 | Married-civ-spouse | ? | Wi |
| **691** | 61 | ? | 202106 | HS-grad | 9 | Married-civ-spouse | ? | Hu |
| **713** | 50 | ? | 204577 | Bachelors | 13 | Married-civ-spouse | ? | Hu |
| **739** | 28 | ? | 123147 | Some-college | 10 | Married-civ-spouse | ? | Wi |
| **822** | 42 | ? | 212206 | Masters | 14 | Married-civ-spouse | ? | Wi |
| **946** | 60 | ? | 191118 | HS-grad | 9 | Married-civ-spouse | ? | Hu |
| **987** | 66 | ? | 213149 | Some-college | 10 | Married-civ-spouse | ? | Hu |
| **994** | 20 | ? | 114746 | 11th | 7 | Married-spouse-absent | ? | Ov |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **32084** | 62 | ? | 178764 | HS-grad | 9 | Married-civ-spouse | ? | Hu |
| **32103** | 24 | ? | 108495 | HS-grad | 9 | Never-married | ? | Ov |
| **32121** | 26 | ? | 375313 | Some-college | 10 | Never-married | ? | Ov |
| **32128** | 18 | ? | 97474 | HS-grad | 9 | Never-married | ? | Ov |
| **32131** | 65 | ? | 192825 | 7th-8th | 4 | Married-civ-spouse | ? | Hu |
| **32133** | 27 | ? | 147638 | Masters | 14 | Never-married | ? | No |
| **32138** | 21 | ? | 155697 | 9th | 5 | Never-married | ? | Ov |
| **32141** | 51 | ? | 43909 | HS-grad | 9 | Divorced | ? | Ur |
| **32146** | 21 | ? | 78374 | HS-grad | 9 | Never-married | ? | Ot rel |
| **32149** | 62 | ? | 263374 | Assoc-voc | 11 | Married-civ-spouse | ? | Hu |
| **32158** | 34 | ? | 330301 | 7th-8th | 4 | Separated | ? | Ur |

| | age | workclass | fnlwgt | education | education.num | marital.status | occupation | re |
|---|---|---|---|---|---|---|---|---|
| **32232** | 19 | ? | 204868 | HS-grad | 9 | Married-civ-spouse | ? | Wi |
| **32243** | 49 | ? | 113913 | HS-grad | 9 | Married-civ-spouse | ? | Wi |
| **32247** | 72 | ? | 96867 | 5th-6th | 3 | Widowed | ? | No |
| **32303** | 20 | ? | 99891 | Some-college | 10 | Never-married | ? | Ov |
| **32319** | 59 | ? | 120617 | Some-college | 10 | Never-married | ? | No |
| **32335** | 21 | ? | 205939 | Some-college | 10 | Never-married | ? | Ov |
| **32341** | 63 | ? | 126540 | Some-college | 10 | Divorced | ? | No |
| **32359** | 18 | ? | 156608 | 11th | 7 | Never-married | ? | Ov |
| **32366** | 66 | ? | 93318 | HS-grad | 9 | Widowed | ? | Ur |
| **32440** | 20 | ? | 203992 | HS-grad | 9 | Never-married | ? | Ov |
| **32483** | 49 | ? | 114648 | 12th | 8 | Divorced | ? | Ot rel |
| **32496** | 60 | ? | 134152 | 9th | 5 | Divorced | ? | No |
| **32500** | 82 | ? | 403910 | HS-grad | 9 | Never-married | ? | No |
| **32528** | 81 | ? | 120478 | Assoc-voc | 11 | Divorced | ? | Ur |
| **32533** | 35 | ? | 320084 | Bachelors | 13 | Married-civ-spouse | ? | Wi |
| **32534** | 30 | ? | 33811 | Bachelors | 13 | Never-married | ? | No |
| **32541** | 71 | ? | 287372 | Doctorate | 16 | Married-civ-spouse | ? | Hu |
| **32543** | 41 | ? | 202822 | HS-grad | 9 | Separated | ? | No |
| **32544** | 72 | ? | 129912 | HS-grad | 9 | Married-civ-spouse | ? | Hu |

1836 rows × 15 columns

Now we can check the number of rows in df_1.

In [8]:
```
df_1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1836 entries, 0 to 32544
Data columns (total 15 columns):
age               1836 non-null int64
workclass         1836 non-null object
fnlwgt            1836 non-null int64
education         1836 non-null object
education.num     1836 non-null int64
marital.status    1836 non-null object
occupation        1836 non-null object
relationship      1836 non-null object
race              1836 non-null object
sex               1836 non-null object
capital.gain      1836 non-null int64
capital.loss      1836 non-null int64
hours.per.week    1836 non-null int64
native.country    1836 non-null object
income            1836 non-null object
dtypes: int64(6), object(9)
memory usage: 229.5+ KB
```

There are 1836 rows with missing values, which is about 5% of the total data. We choose to simply drop these rows.

In [9]:
```
# dropping the rows having missing values in workclass
df = df[df['workclass'] != '?']
df.head()
```

Out[9]:

| | age | workclass | fnlwgt | education | education.num | marital.status | occupation | relatio |
|---|---|---|---|---|---|---|---|---|
| **1** | 82 | Private | 132870 | HS-grad | 9 | Widowed | Exec-managerial | Not-in-f |
| **3** | 54 | Private | 140359 | 7th-8th | 4 | Divorced | Machine-op-inspct | Unmarr |
| **4** | 41 | Private | 264663 | Some-college | 10 | Separated | Prof-specialty | Own-ch |
| **5** | 34 | Private | 216864 | HS-grad | 9 | Divorced | Other-service | Unmarr |
| **6** | 38 | Private | 150601 | 10th | 6 | Separated | Adm-clerical | Unmarr |

Let's see whether any other columns contain a "?". Since "?" is a string, we can apply this check only on the categorical columns.

In [10]:
```
# select all categorical variables
df_categorical = df.select_dtypes(include=['object'])

# checking whether any other columns contain a "?"
df_categorical.apply(lambda x: x=="?", axis=0).sum()
```

Out[10]:
```
workclass           0
education           0
marital.status      0
occupation          7
relationship        0
race                0
sex                 0
native.country    556
income              0
dtype: int64
```

Thus, the columns occupation and native.country contain some "?"s. Let's get rid of them.

In [11]:
```
# dropping the "?"s
df = df[df['occupation'] != '?']
df = df[df['native.country'] != '?']
```

Now we have a clean dataframe which is ready for model building.

In [12]:
```
# clean dataframe
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 30162 entries, 1 to 32560
Data columns (total 15 columns):
age               30162 non-null int64
workclass         30162 non-null object
fnlwgt            30162 non-null int64
education         30162 non-null object
education.num     30162 non-null int64
marital.status    30162 non-null object
occupation        30162 non-null object
relationship      30162 non-null object
race              30162 non-null object
sex               30162 non-null object
capital.gain      30162 non-null int64
capital.loss      30162 non-null int64
hours.per.week    30162 non-null int64
native.country    30162 non-null object
income            30162 non-null object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
```

# Data Preparation

There are a number of preprocessing steps we need to do before building the model.

Firstly, note that we have both categorical and numeric features as predictors. In previous models such as linear and logistic regression, we had created **dummy variables** for categorical variables, since those models (being mathematical equations) can process only numeric variables.

All that is not required in decision trees, since they can process categorical variables easily. However, we still need to **encode the categorical variables** into a standard format so that sklearn can understand them and build the tree. We'll do that using the `LabelEncoder()` class, which comes with `sklearn.preprocessing`.

You can read the documentation of `LabelEncoder` here (http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html).

In [13]:
```
from sklearn import preprocessing


# encode categorical variables using Label Encoder

# select all categorical variables
df_categorical = df.select_dtypes(include=['object'])
df_categorical.head()
```

Out[13]:

| | workclass | education | marital.status | occupation | relationship | race | sex | native.c |
|---|---|---|---|---|---|---|---|---|
| **1** | Private | HS-grad | Widowed | Exec-managerial | Not-in-family | White | Female | United-S |
| **3** | Private | 7th-8th | Divorced | Machine-op-inspct | Unmarried | White | Female | United-S |
| **4** | Private | Some-college | Separated | Prof-specialty | Own-child | White | Female | United-S |
| **5** | Private | HS-grad | Divorced | Other-service | Unmarried | White | Female | United-S |
| **6** | Private | 10th | Separated | Adm-clerical | Unmarried | White | Male | United-S |

In [14]:
```
# apply Label encoder to df_categorical

le = preprocessing.LabelEncoder()
df_categorical = df_categorical.apply(le.fit_transform)
df_categorical.head()
```

Out[14]:

| | workclass | education | marital.status | occupation | relationship | race | sex | native.countr |
|---|---|---|---|---|---|---|---|---|
| **1** | 2 | 11 | 6 | 3 | 1 | 4 | 0 | 38 |
| **3** | 2 | 5 | 0 | 6 | 4 | 4 | 0 | 38 |
| **4** | 2 | 15 | 5 | 9 | 3 | 4 | 0 | 38 |
| **5** | 2 | 11 | 0 | 7 | 4 | 4 | 0 | 38 |
| **6** | 2 | 0 | 5 | 0 | 4 | 4 | 1 | 38 |

In [15]:
```
# concat df_categorical with original df
df = df.drop(df_categorical.columns, axis=1)
df = pd.concat([df, df_categorical], axis=1)
df.head()
```

Out[15]:

| | age | fnlwgt | education.num | capital.gain | capital.loss | hours.per.week | workclass | edu |
|---|---|---|---|---|---|---|---|---|
| **1** | 82 | 132870 | 9 | 0 | 4356 | 18 | 2 | 11 |
| **3** | 54 | 140359 | 4 | 0 | 3900 | 40 | 2 | 5 |
| **4** | 41 | 264663 | 10 | 0 | 3900 | 40 | 2 | 15 |
| **5** | 34 | 216864 | 9 | 0 | 3770 | 45 | 2 | 11 |
| **6** | 38 | 150601 | 6 | 0 | 3770 | 40 | 2 | 0 |

In [16]:
```
# look at column types
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 30162 entries, 1 to 32560
Data columns (total 15 columns):
age               30162 non-null int64
fnlwgt            30162 non-null int64
education.num     30162 non-null int64
capital.gain      30162 non-null int64
capital.loss      30162 non-null int64
hours.per.week    30162 non-null int64
workclass         30162 non-null int64
education         30162 non-null int64
marital.status    30162 non-null int64
occupation        30162 non-null int64
relationship      30162 non-null int64
race              30162 non-null int64
sex               30162 non-null int64
native.country    30162 non-null int64
income            30162 non-null int64
dtypes: int64(15)
memory usage: 3.7 MB
```

In [17]:
```
# convert target variable income to categorical
df['income'] = df['income'].astype('category')
```

Now all the categorical variables are suitably encoded. Let's build the model.

---

## Model Building and Evaluation

Let's first build a decision tree with default hyperparameters. Then we'll use cross-validation to tune them.

In [18]:
```
# Importing train-test-split
from sklearn.model_selection import train_test_split
```

In [19]:
```
# Putting feature variable to X
X = df.drop('income',axis=1)

# Putting response variable to y
y = df['income']
```

In [20]:
```
# Splitting the data into train and test
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=0.30,
                                                    random_state = 99)
X_train.head()
```

Out[20]:

| | age | fnlwgt | education.num | capital.gain | capital.loss | hours.per.week | workclass |
|---|---|---|---|---|---|---|---|
| **24351** | 42 | 289636 | 9 | 0 | 0 | 46 | 2 |
| **15626** | 37 | 52465 | 9 | 0 | 0 | 40 | 1 |
| **4347** | 38 | 125933 | 14 | 0 | 0 | 40 | 0 |
| **23972** | 44 | 183829 | 13 | 0 | 0 | 38 | 5 |
| **26843** | 35 | 198841 | 11 | 0 | 0 | 35 | 2 |

In [21]:
```
# Importing decision tree classifier from sklearn library
from sklearn.tree import DecisionTreeClassifier

# Fitting the decision tree with default hyperparameters, apart from
# max_depth which is 5 so that we can plot and read the tree.
dt_default = DecisionTreeClassifier(max_depth=5)
dt_default.fit(X_train, y_train)
```

Out[21]:
```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=5,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')
```

In [24]:
```
# Let's check the evaluation metrics of our default model

# Importing classification report and confusion matrix from sklearn metrics
from sklearn.metrics import classification_report, confusion_matrix, accuracy_
score

# Making predictions
y_pred_default = dt_default.predict(X_test)

# Printing classification report
print(classification_report(y_test, y_pred_default))
```

```
             precision    recall  f1-score   support

          0       0.86      0.95      0.91      6867
          1       0.78      0.52      0.63      2182

avg / total       0.84      0.85      0.84      9049
```

In [25]:
```
# Printing confusion matrix and accuracy
print(confusion_matrix(y_test,y_pred_default))
print(accuracy_score(y_test,y_pred_default))
```

```
[[6553  314]
 [1039 1143]]
0.850480716101
```

---

## Plotting the Decision Tree

To visualise decision trees in python, you need to install certain external libraries. You can read about the process in detail here: [http://scikit-learn.org/stable/modules/tree.html (http://scikit-learn.org/stable/modules/tree.html)](http://scikit-learn.org/stable/modules/tree.html)

We need the `graphviz` library to plot a tree.

In [26]:
```
# Importing required packages for visualization
from IPython.display import Image
from sklearn.externals.six import StringIO
from sklearn.tree import export_graphviz
import pydot, graphviz

# Putting features
features = list(df.columns[1:])
features
```

Out[26]:
```
['fnlwgt',
 'education.num',
 'capital.gain',
 'capital.loss',
 'hours.per.week',
 'workclass',
 'education',
 'marital.status',
 'occupation',
 'relationship',
 'race',
 'sex',
 'native.country',
 'income']
```

**Note**:
Python requires library pydot and an external software graphviz to visualize the decision tree. If you are on wondows, you'll need to specify the path for the pydot library to access dot file from graphviz.

Please read the downloadable instructions to install graphviz. For Mac users, one way is to:

- Install the python graphviz module: `conda install python-graphviz`
- Then install the Graphviz software on Mac, you do this using homebrew:
  - Install homebrew: https://docs.brew.sh/Installation (https://docs.brew.sh/Installation)
  - `brew install graphviz`

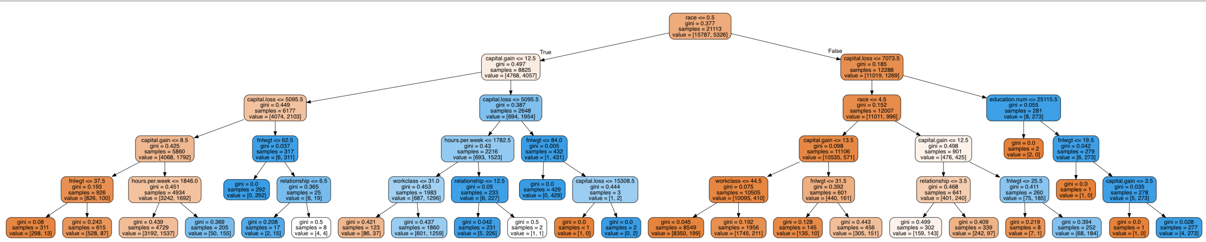
In [ ]:

```
# If you're on windows:
# Specifing path for dot file.
# import os
# os.environ["PATH"] += os.pathsep + 'C:/Program Files (x86)/Graphviz2.38/bi
n/'
```

In [27]:

```
# plotting tree with max_depth=3
dot_data = StringIO()
export_graphviz(dt_default, out_file=dot_data,
                feature_names=features, filled=True,rounded=True)

graph = pydot.graph_from_dot_data(dot_data.getvalue())
Image(graph[0].create_png())
```

Out[27]:



## Hyperparameter Tuning

The default tree is quite complex, and we need to simplify it by tuning the hyperparameters.

First, let's understand the parameters in a decision tree. You can read this in the documentation using `help(DecisionTreeClassifier)`.

- **criterion** (Gini/IG or entropy): It defines the function to measure the quality of a split. Sklearn supports “gini” criteria for Gini Index & “entropy” for Information Gain. By default, it takes the value “gini”.
- **splitter**: It defines the strategy to choose the split at each node. Supports “best” value to choose the best split & “random” to choose the best random split. By default, it takes “best” value.
- **max_features**: It defines the no. of features to consider when looking for the best split. We can input integer, float, string & None value.
  - If an integer is inputted then it considers that value as max features at each split.
  - If float value is taken then it shows the percentage of features at each split.
  - If “auto” or “sqrt” is taken then max_features=sqrt(n_features).
  - If “log2” is taken then max_features= log2(n_features).
  - If None, then max_features=n_features. By default, it takes “None” value.
- **max_depth**: The max_depth parameter denotes maximum depth of the tree. It can take any integer value or None. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples. By default, it takes “None” value.
- **min_samples_split**: This tells above the minimum no. of samples reqd. to split an internal node. If an integer value is taken then consider min_samples_split as the minimum no. If float, then it shows percentage. By default, it takes “2” value.
- **min_samples_leaf**: The minimum number of samples required to be at a leaf node. If an integer value is taken then consider - -min_samples_leaf as the minimum no. If float, then it shows percentage. By default, it takes “1” value.
- **max_leaf_nodes**: It defines the maximum number of possible leaf nodes. If None then it takes an unlimited number of leaf nodes. By default, it takes “None” value.
- **min_impurity_split**: It defines the threshold for early stopping tree growth. A node will split if its impurity is above the threshold otherwise it is a leaf.

## Tuning max_depth

Let's first try to find the optimum values for max_depth and understand how the value of max_depth affects the decision tree.

Here, we are creating a dataframe with max_depth in range 1 to 80 and checking the accuracy score corresponding to each max_depth.

To reiterate, a grid search scheme consists of:

```
- an estimator (classifier such as SVC() or decision tree)
- a parameter space
- a method for searching or sampling candidates (optional)
- a cross-validation scheme, and
- a score function (accuracy, roc_auc etc.)
```

In [28]:

```python
# GridSearchCV to find optimal max_depth
from sklearn.model_selection import KFold
from sklearn.model_selection import GridSearchCV


# specify number of folds for k-fold CV
n_folds = 5

# parameters to build the model on
parameters = {'max_depth': range(1, 40)}

# instantiate the model
dtree = DecisionTreeClassifier(criterion = "gini",
                               random_state = 100)

# fit tree on training data
tree = GridSearchCV(dtree, parameters,
                    cv=n_folds,
                   scoring="accuracy")
tree.fit(X_train, y_train)
```

Out[28]:

```
GridSearchCV(cv=5, error_score='raise',
       estimator=DecisionTreeClassifier(class_weight=None, criterion='gini',
max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=100,
            splitter='best'),
       fit_params=None, iid=True, n_jobs=1,
       param_grid={'max_depth': range(1, 40)}, pre_dispatch='2*n_jobs',
       refit=True, return_train_score='warn', scoring='accuracy',
       verbose=0)
```

In [29]:
```
# scores of GridSearch CV
scores = tree.cv_results_
pd.DataFrame(scores).head()
```

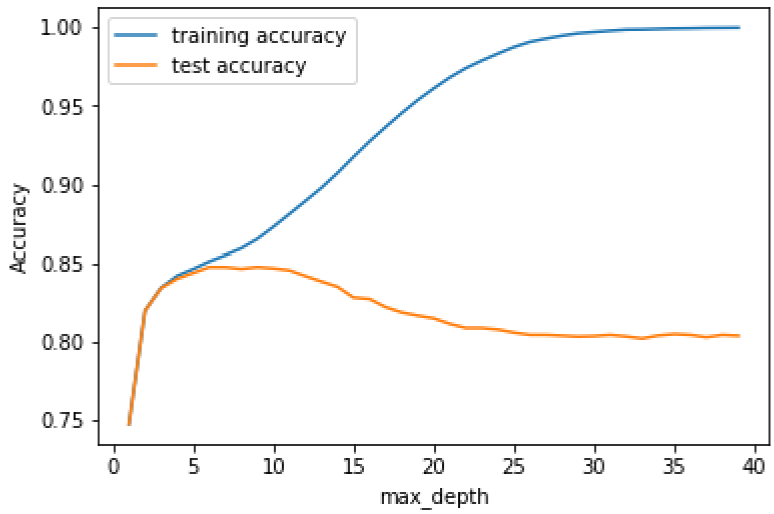
Out[29]:

| | mean_fit_time | mean_score_time | mean_test_score | mean_train_score | param_max_de |
|---|---|---|---|---|---|
| **0** | 0.011095 | 0.001951 | 0.747738 | 0.747738 | 1 |
| **1** | 0.014384 | 0.001419 | 0.819969 | 0.819969 | 2 |
| **2** | 0.019318 | 0.001629 | 0.834273 | 0.834510 | 3 |
| **3** | 0.027667 | 0.001569 | 0.840193 | 0.842088 | 4 |
| **4** | 0.036721 | 0.002027 | 0.843888 | 0.846386 | 5 |

5 rows × 21 columns

Now let's visualize how train and test score changes with max_depth.

In [30]:
```
# plotting accuracies with max_depth
plt.figure()
plt.plot(scores["param_max_depth"],
         scores["mean_train_score"],
         label="training accuracy")
plt.plot(scores["param_max_depth"],
         scores["mean_test_score"],
         label="test accuracy")
plt.xlabel("max_depth")
plt.ylabel("Accuracy")
plt.legend()
plt.show()
```




You can see that as we increase the value of max_depth, both training and test score increase till about max-depth = 10, after which the test score gradually reduces. Note that the scores are average accuracies across the 5-folds.

Thus, it is clear that the model is overfitting the training data if the max_depth is too high. Next, let's see how the model behaves with other hyperparameters.

---

## Tuning min_samples_leaf

The hyperparameter **min_samples_leaf** indicates the minimum number of samples required to be at a leaf.

So if the values of min_samples_leaf is less, say 5, then the will be constructed even if a leaf has 5, 6 etc. observations (and is likely to overfit).

Let's see what will be the optimum value for min_samples_leaf.

In [31]:
```
# GridSearchCV to find optimal max_depth
from sklearn.model_selection import KFold
from sklearn.model_selection import GridSearchCV


# specify number of folds for k-fold CV
n_folds = 5

# parameters to build the model on
parameters = {'min_samples_leaf': range(5, 200, 20)}

# instantiate the model
dtree = DecisionTreeClassifier(criterion = "gini",
                               random_state = 100)

# fit tree on training data
tree = GridSearchCV(dtree, parameters,
                    cv=n_folds,
                   scoring="accuracy")
tree.fit(X_train, y_train)
```

Out[31]:
```
GridSearchCV(cv=5, error_score='raise',
       estimator=DecisionTreeClassifier(class_weight=None, criterion='gini',
max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=100,
            splitter='best'),
       fit_params=None, iid=True, n_jobs=1,
       param_grid={'min_samples_leaf': range(5, 200, 20)},
       pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
       scoring='accuracy', verbose=0)
```

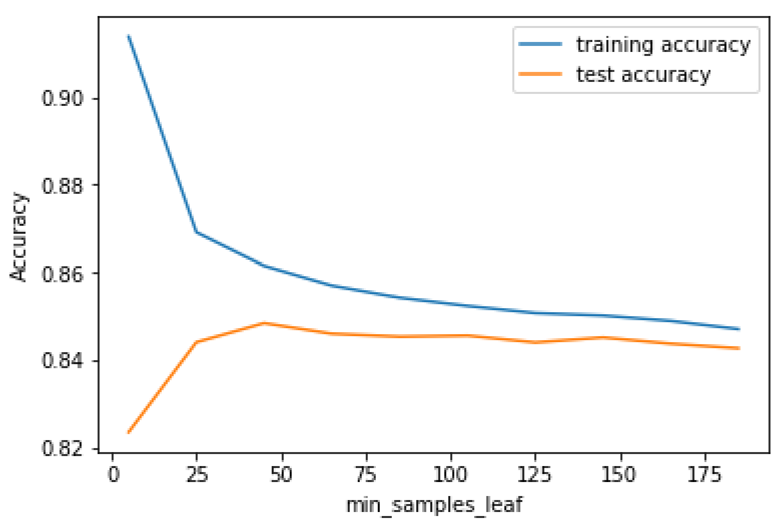
In [32]:
```
# scores of GridSearch CV
scores = tree.cv_results_
pd.DataFrame(scores).head()
```

Out[32]:

| | mean_fit_time | mean_score_time | mean_test_score | mean_train_score | param_min_sa |
|---|---|---|---|---|---|
| **0** | 0.076940 | 0.002072 | 0.823663 | 0.913785 | 5 |
| **1** | 0.070042 | 0.002049 | 0.844172 | 0.869180 | 25 |
| **2** | 0.055783 | 0.001711 | 0.848529 | 0.861554 | 45 |
| **3** | 0.061370 | 0.002087 | 0.846114 | 0.857067 | 65 |
| **4** | 0.051622 | 0.002022 | 0.845451 | 0.854320 | 85 |

5 rows × 21 columns

In [33]:
```
# plotting accuracies with min_samples_leaf
plt.figure()
plt.plot(scores["param_min_samples_leaf"],
         scores["mean_train_score"],
         label="training accuracy")
plt.plot(scores["param_min_samples_leaf"],
         scores["mean_test_score"],
         label="test accuracy")
plt.xlabel("min_samples_leaf")
plt.ylabel("Accuracy")
plt.legend()
plt.show()
```



You can see that at low values of min_samples_leaf, the tree gets a bit overfitted. At values > 100, however, the model becomes more stable and the training and test accuracy start to converge.

---

## Tuning min_samples_split

The hyperparameter **min_samples_split** is the minimum no. of samples required to split an internal node. Its default value is 2, which means that even if a node is having 2 samples it can be furthur divided into leaf nodes.

In [34]:
```
# GridSearchCV to find optimal min_samples_split
from sklearn.model_selection import KFold
from sklearn.model_selection import GridSearchCV


# specify number of folds for k-fold CV
n_folds = 5

# parameters to build the model on
parameters = {'min_samples_split': range(5, 200, 20)}

# instantiate the model
dtree = DecisionTreeClassifier(criterion = "gini",
                               random_state = 100)

# fit tree on training data
tree = GridSearchCV(dtree, parameters,
                    cv=n_folds,
                   scoring="accuracy")
tree.fit(X_train, y_train)
```

Out[34]:
```
GridSearchCV(cv=5, error_score='raise',
       estimator=DecisionTreeClassifier(class_weight=None, criterion='gini',
max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=100,
            splitter='best'),
       fit_params=None, iid=True, n_jobs=1,
       param_grid={'min_samples_split': range(5, 200, 20)},
       pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
       scoring='accuracy', verbose=0)
```

In [35]:

```
# scores of GridSearch CV
scores = tree.cv_results_
pd.DataFrame(scores).head()
```

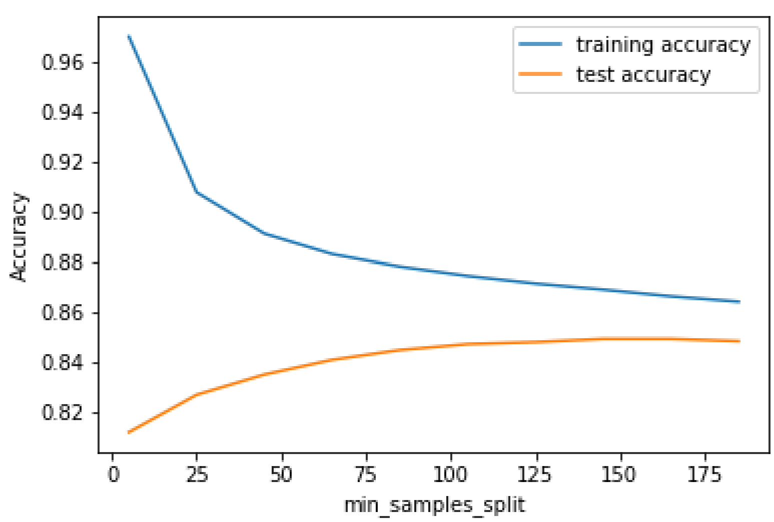
Out[35]:

| | mean_fit_time | mean_score_time | mean_test_score | mean_train_score | param_min_sa |
|---|---|---|---|---|---|
| **0** | 0.083918 | 0.001904 | 0.811775 | 0.969865 | 5 |
| **1** | 0.087892 | 0.002155 | 0.826742 | 0.907569 | 25 |
| **2** | 0.079704 | 0.002158 | 0.834841 | 0.891062 | 45 |
| **3** | 0.084147 | 0.002347 | 0.840714 | 0.882928 | 65 |
| **4** | 0.078208 | 0.002643 | 0.844645 | 0.877919 | 85 |

5 rows × 21 columns

In [36]:

```
# plotting accuracies with min_samples_leaf
plt.figure()
plt.plot(scores["param_min_samples_split"],
         scores["mean_train_score"],
         label="training accuracy")
plt.plot(scores["param_min_samples_split"],
         scores["mean_test_score"],
         label="test accuracy")
plt.xlabel("min_samples_split")
plt.ylabel("Accuracy")
plt.legend()
plt.show()
```



This shows that as you increase the min_samples_split, the tree overfits lesser since the model is less complex.

---

# Grid Search to Find Optimal Hyperparameters

We can now use GridSearchCV to find multiple optimal hyperparameters together. Note that this time, we'll also specify the criterion (gini/entropy or IG).

In [37]:
```
# Create the parameter grid
param_grid = {
    'max_depth': range(5, 15, 5),
    'min_samples_leaf': range(50, 150, 50),
    'min_samples_split': range(50, 150, 50),
    'criterion': ["entropy", "gini"]
}

n_folds = 5

# Instantiate the grid search model
dtree = DecisionTreeClassifier()
grid_search = GridSearchCV(estimator = dtree, param_grid = param_grid,
                          cv = n_folds, verbose = 1)

# Fit the grid search to the data
grid_search.fit(X_train,y_train)
```

```
Fitting 5 folds for each of 16 candidates, totalling 80 fits

[Parallel(n_jobs=1)]: Done  80 out of  80 | elapsed:    4.3s finished
```

Out[37]:
```
GridSearchCV(cv=5, error_score='raise',
       estimator=DecisionTreeClassifier(class_weight=None, criterion='gini',
max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best'),
       fit_params=None, iid=True, n_jobs=1,
       param_grid={'criterion': ['entropy', 'gini'], 'min_samples_split': ran
ge(50, 150, 50), 'min_samples_leaf': range(50, 150, 50), 'max_depth': range
(5, 15, 5)},
       pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
       scoring=None, verbose=1)
```

In [38]:
```
# cv results
cv_results = pd.DataFrame(grid_search.cv_results_)
cv_results
```

Out[38]:

| | mean_fit_time | mean_score_time | mean_test_score | mean_train_score | param_criteri |
|---|---|---|---|---|---|
| **0** | 0.038060 | 0.001683 | 0.841804 | 0.843296 | entropy |
| **1** | 0.036670 | 0.001513 | 0.841804 | 0.843296 | entropy |
| **2** | 0.042220 | 0.002639 | 0.841614 | 0.843142 | entropy |
| **3** | 0.049321 | 0.001518 | 0.841614 | 0.843142 | entropy |
| **4** | 0.062581 | 0.001753 | 0.849903 | 0.854083 | entropy |
| **5** | 0.058499 | 0.001862 | 0.849903 | 0.854083 | entropy |
| **6** | 0.053737 | 0.001614 | 0.848956 | 0.851312 | entropy |
| **7** | 0.058195 | 0.002408 | 0.848956 | 0.851312 | entropy |
| **8** | 0.028285 | 0.001407 | 0.844409 | 0.846055 | gini |
| **9** | 0.030749 | 0.001483 | 0.844409 | 0.846055 | gini |
| **10** | 0.029885 | 0.001636 | 0.843177 | 0.845143 | gini |
| **11** | 0.029095 | 0.001720 | 0.843177 | 0.845143 | gini |
| **12** | 0.059498 | 0.002442 | 0.851466 | 0.855563 | gini |

| | mean_fit_time | mean_score_time | mean_test_score | mean_train_score | param_criteri |
|---|---|---|---|---|---|
| **13** | 0.049704 | 0.001789 | 0.851466 | 0.855563 | gini |
| **14** | 0.048904 | 0.001591 | 0.846493 | 0.851336 | gini |
| **15** | 0.058167 | 0.002185 | 0.846493 | 0.851336 | gini |

16 rows × 24 columns

In [39]:
```
# printing the optimal accuracy score and hyperparameters
print("best accuracy", grid_search.best_score_)
print(grid_search.best_estimator_)
```

```
best accuracy 0.85146592147
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=10,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=50, min_samples_split=50,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')
```

**Running the model with best parameters obtained from grid search.**

In [40]:
```
# model with optimal hyperparameters
clf_gini = DecisionTreeClassifier(criterion = "gini",
                                  random_state = 100,
                                  max_depth=10,
                                  min_samples_leaf=50,
                                  min_samples_split=50)
clf_gini.fit(X_train, y_train)
```

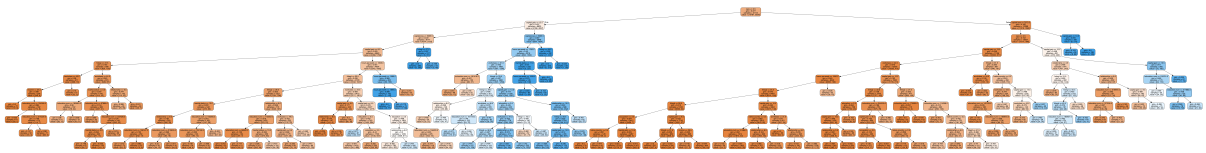
Out[40]:
```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=10,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=50, min_samples_split=50,
            min_weight_fraction_leaf=0.0, presort=False, random_state=100,
            splitter='best')
```

In [41]:
```
# accuracy score
clf_gini.score(X_test,y_test)
```

Out[41]: 0.85092275389545802

In [42]:
```
# plotting the tree
dot_data = StringIO()
export_graphviz(clf_gini, out_file=dot_data,feature_names=features,filled=True
,rounded=True)

graph = pydot.graph_from_dot_data(dot_data.getvalue())
Image(graph[0].create_png())
```

Out[42]:



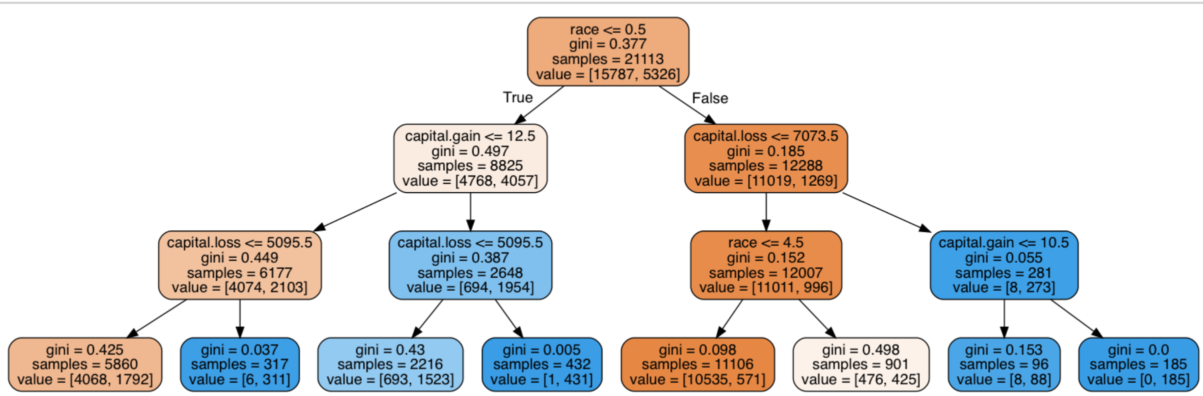
You can see that this tree is too complex to understand. Let's try reducing the max_depth and see how the tree looks.

In [43]:
```
# tree with max_depth = 3
clf_gini = DecisionTreeClassifier(criterion = "gini",
                                  random_state = 100,
                                  max_depth=3,
                                  min_samples_leaf=50,
                                  min_samples_split=50)
clf_gini.fit(X_train, y_train)

# score
print(clf_gini.score(X_test,y_test))
```

0.839319261797

In [44]:
```
# plotting tree with max_depth=3
dot_data = StringIO()
export_graphviz(clf_gini, out_file=dot_data,feature_names=features,filled=True
,rounded=True)

graph = pydot.graph_from_dot_data(dot_data.getvalue())
Image(graph[0].create_png())
```

Out[44]:



In [45]:
```
# classification metrics
from sklearn.metrics import classification_report,confusion_matrix
y_pred = clf_gini.predict(X_test)
print(classification_report(y_test, y_pred))
```

```
             precision    recall  f1-score   support

          0       0.85      0.96      0.90      6867
          1       0.77      0.47      0.59      2182

avg / total       0.83      0.84      0.82      9049
```

In [46]:
```
# confusion matrix
print(confusion_matrix(y_test,y_pred))
```

```
[[6564  303]
 [1151 1031]]
```