Adult1

December 19, 2020

Adding Data from google DRIVE

```python
import pandas as pd

from google.colab import drive
drive.mount('/content/drive')
```

Mounted at /content/drive

```python
bd_data = pd.read_csv("drive/My Drive/Colab Notebooks/adult1.csv")
bd_data
```

[19]:

	age	workclass	education	...	sex	Hours-per-week	class
0	39	State-gov	Bachelors	...	Male	40	<=50K
1	50	Self-emp-not-inc	Bachelors	...	Male	13	<=50K
2	38	Private	HS-grad	...	Male	40	<=50K
3	53	Private	11th	...	Male	40	<=50K
4	28	Private	Bachelors	...	Female	40	<=50K
5	37	Private	Masters	...	Female	40	<=50K
6	49	Private	9th	...	Female	16	<=50K
7	52	Self-emp-not-inc	HS-grad	...	Male	45	>50K
8	31	Private	Masters	...	Female	50	>50K
9	42	Private	Bachelors	...	Male	40	>50K
10	30	State-gov	Bachelors	...	Male	40	>50K
11	23	Private	Bachelors	...	Female	30	<=50K
12	32	Private	Assoc-acdm	...	Male	50	<=50K
13	40	Private	Assoc-voc	...	Male	40	>50K
14	34	Private	7th-8th	...	Male	45	<=50K
15	32	Private	HS-grad	...	Male	40	<=50K
16	38	Private	11th	...	Male	50	<=50K
17	43	Self-emp-not-inc	Masters	...	Female	45	>50K
18	40	Private	Doctorate	...	Male	60	>50K
19	54	Private	HS-grad	...	Female	20	<=50K
20	35	Federal-gov	9th	...	Male	40	<=50K
21	43	Private	11th	...	Male	40	<=50K
22	59	Private	HS-grad	...	Female	40	<=50K
23	56	Local-gov	Bachelors	...	Male	40	>50K
24	19	Private	HS-grad	...	Male	40	<=50K

```
25  39           Private        HS-grad  ...    Male          80  <=50K
26  49           Private        HS-grad  ...    Male          40  <=50K
27  23         Local-gov      Assoc-acdm  ...    Male          52  <=50K
28  20           Private    Some-college  ...    Male          44  <=50K
29  45           Private        Bachelors  ...   Male          40  <=50K
30  30       Federal-gov    Some-college  ...    Male          40  <=50K
31  48           Private           11th  ...    Male          40  <=50K
32  21           Private    Some-college  ...    Male          40  <=50K
33  19           Private        HS-grad  ...  Female          25  <=50K
34  31           Private    Some-college  ...    Male          38   >50K
35  48   Self-emp-not-inc    Assoc-acdm  ...    Male          40  <=50K
36  31           Private            9th  ...    Male          43  <=50K
37  53   Self-emp-not-inc     Bachelors  ...    Male          40  <=50K
38  24           Private      Bachelors  ...    Male          50  <=50K
39  25           Private        HS-grad  ...    Male          35  <=50K
40  57       Federal-gov      Bachelors  ...    Male          40   >50K
41  53           Private        HS-grad  ...    Male          38  <=50K
42  44           Private        Masters  ...  Female          40  <=50K
43  41         State-gov      Assoc-voc  ...    Male          40  <=50K
44  29           Private      Assoc-voc  ...    Male          43  <=50K

[45 rows x 8 columns]
```

Encoding data :

Some attributes type is "object", so before encoding data we will need to look for those attributes and change their type to string.

```python
[20]: from sklearn.preprocessing import LabelEncoder
      bd_data_num = bd_data


      cat_features=[x for x in bd_data_num.columns if bd_data_num[x].dtype=="object"]
      cat_features
      le=LabelEncoder()
      for col in cat_features:
          if col in bd_data_num.columns:
              i = bd_data_num.columns.get_loc(col)
              bd_data_num.iloc[:,i] = bd_data_num.apply(lambda i:le.fit_transform(i.
       →astype(str)), axis=0, result_type='expand')


      bd_data = pd.read_csv("drive/My Drive/Colab Notebooks/adult1.csv")          ␣
       ↳     #to not confuse between "bd_data" and "bd_data_num"
      bd_data_num
```

```
[20]:     age  workclass  education  ...  sex  Hours-per-week  class
      0    39          4          5  ...    1              40      0
      1    50          3          5  ...    1              13      0
      2    38          2          7  ...    1              40      0
```

3	53	2	0	...	1	40	0
4	28	2	5	...	0	40	0
5	37	2	8	...	0	40	0
6	49	2	2	...	0	16	0
7	52	3	7	...	1	45	1
8	31	2	8	...	0	50	1
9	42	2	5	...	1	40	1
10	30	4	5	...	1	40	1
11	23	2	5	...	0	30	0
12	32	2	3	...	1	50	0
13	40	2	4	...	1	40	1
14	34	2	1	...	1	45	0
15	32	2	7	...	1	40	0
16	38	2	0	...	1	50	0
17	43	3	8	...	0	45	1
18	40	2	6	...	1	60	1
19	54	2	7	...	0	20	0
20	35	0	2	...	1	40	0
21	43	2	0	...	1	40	0
22	59	2	7	...	0	40	0
23	56	1	5	...	1	40	1
24	19	2	7	...	1	40	0
25	39	2	7	...	1	80	0
26	49	2	7	...	1	40	0
27	23	1	3	...	1	52	0
28	20	2	9	...	1	44	0
29	45	2	5	...	1	40	0
30	30	0	9	...	1	40	0
31	48	2	0	...	1	40	0
32	21	2	9	...	1	40	0
33	19	2	7	...	0	25	0
34	31	2	9	...	1	38	1
35	48	3	3	...	1	40	0
36	31	2	2	...	1	43	0
37	53	3	5	...	1	40	0
38	24	2	5	...	1	50	0
39	25	2	7	...	1	35	0
40	57	0	5	...	1	40	1
41	53	2	7	...	1	38	0
42	44	2	8	...	0	40	0
43	41	4	4	...	1	40	0
44	29	2	4	...	1	43	0

[45 rows x 8 columns]

Spliting data to train and test datasets

```
[41]: import numpy as np
      from sklearn.model_selection import train_test_split

      data = np.array(bd_data_num.iloc[:, :7])
      target = np.array(bd_data_num.iloc[:, 7])

      data_train, data_test, target_train, target_test = train_test_split(data,␣
       ↪target, test_size = 0.3)
```

Training model

```
[45]: from sklearn import tree

      bd_tree = tree.DecisionTreeClassifier()

      bd_tree.fit(data_train, target_train);
```
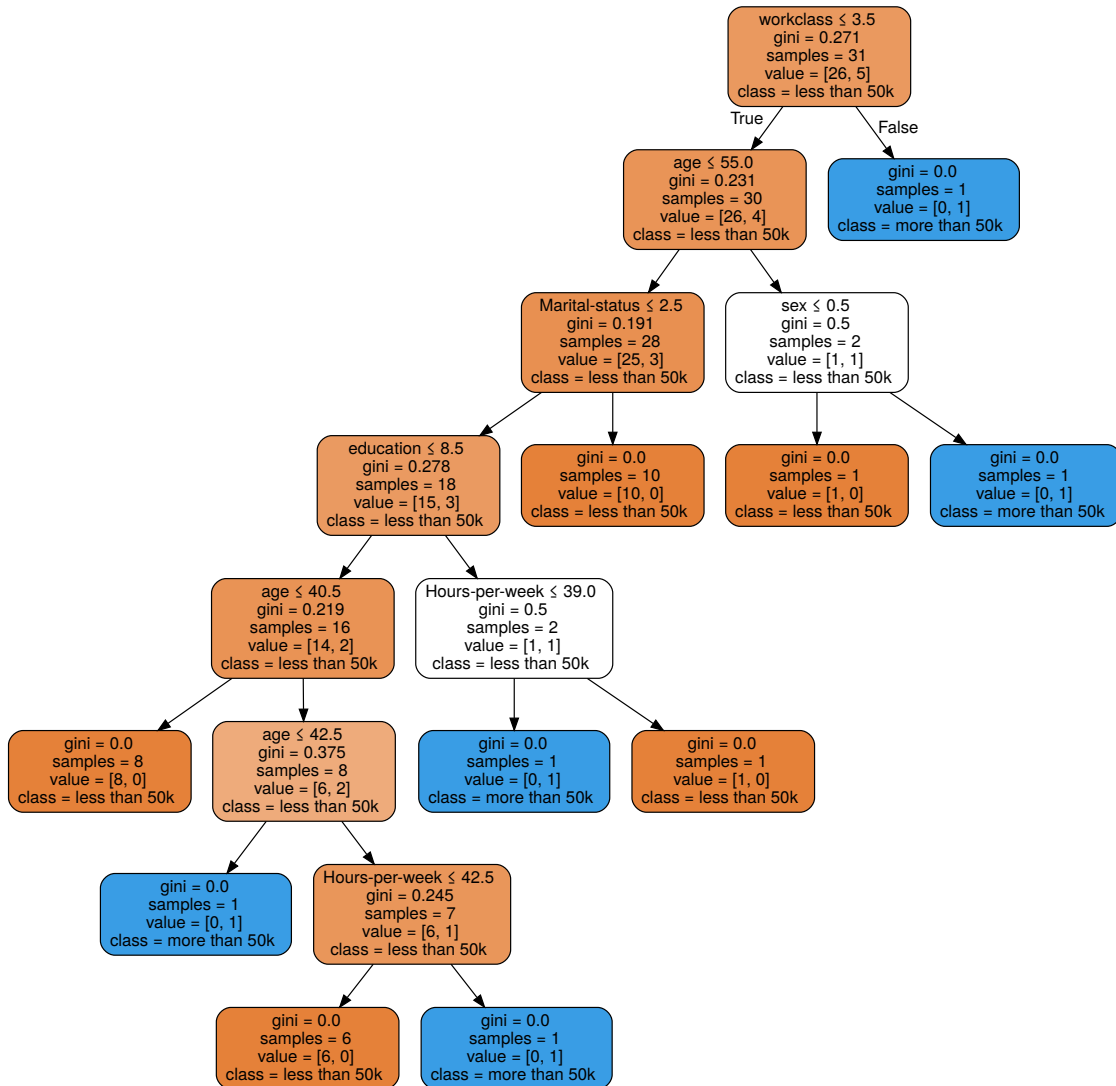
Plotting the decision tree

```
[80]: import graphviz

      dot = tree.export_graphviz(bd_tree, out_file=None,
                                 feature_names=bd_data.columns.values[:7],
                                 class_names=["less than 50k", "more than 50k"],
                                 filled=True, rotate=False, rounded=True,
                                 special_characters=True)

      graph = graphviz.Source(dot)
      graph
```

[80]:

EXPORTING THE GRAPHIC TO AN IMAGE

```
[ ]: graph.render('DecisionTree', view=True)
```

```
[47]: !pip install pdf2image
```

```
Collecting pdf2image
  Downloading https://files.pythonhosted.org/packages/03/62/089030fd16ab3e5c2453
15d63c80b29250b8f9e4579b5a09306eb7e7539c/pdf2image-1.14.0-py3-none-any.whl
Requirement already satisfied: pillow in /usr/local/lib/python3.6/dist-packages
(from pdf2image) (7.0.0)
Installing collected packages: pdf2image
Successfully installed pdf2image-1.14.0
```

```
[48]: !apt-get install poppler-utils
```

```
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following NEW packages will be installed:
  poppler-utils
0 upgraded, 1 newly installed, 0 to remove and 14 not upgraded.
Need to get 154 kB of archives.
After this operation, 613 kB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 poppler-utils
amd64 0.62.0-2ubuntu2.12 [154 kB]
Fetched 154 kB in 1s (296 kB/s)
Selecting previously unselected package poppler-utils.
(Reading database ... 144865 files and directories currently installed.)
Preparing to unpack .../poppler-utils_0.62.0-2ubuntu2.12_amd64.deb ...
Unpacking poppler-utils (0.62.0-2ubuntu2.12) ...
Setting up poppler-utils (0.62.0-2ubuntu2.12) ...
Processing triggers for man-db (2.8.3-2ubuntu0.1) ...
```

[50]:
```python
from pdf2image import convert_from_path
images = convert_from_path('/content/DecisionTree.pdf')
for img in images:
    img.save('DecisionTree.jpg', 'JPEG')
```

Model Accuracy

[51]:
```python
from sklearn.metrics import accuracy_score


prediction = bd_tree.predict(data_test)
score = accuracy_score(target_test, prediction)
score
```

[51]: 0.6428571428571429

Expoting notebook to pdf

[92]:
```python
!sudo apt update && sudo apt upgrade
!sudo apt install inkscape pandoc texlive-xetex texlive-fonts-recommended␣
 ↪texlive-generic-recommended
```

```
Hit:1 http://security.ubuntu.com/ubuntu bionic-security InRelease
Ign:2 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1804/x86_64
InRelease
Hit:3 https://cloud.r-project.org/bin/linux/ubuntu bionic-cran40/ InRelease
Ign:4 https://developer.download.nvidia.com/compute/machine-
learning/repos/ubuntu1804/x86_64  InRelease
Hit:5 http://ppa.launchpad.net/c2d4u.team/c2d4u4.0+/ubuntu bionic InRelease
Hit:6 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1804/x86_64
Release
Hit:7 http://archive.ubuntu.com/ubuntu bionic InRelease
Hit:8 https://developer.download.nvidia.com/compute/machine-
```

```
learning/repos/ubuntu1804/x86_64  Release
Hit:9 http://archive.ubuntu.com/ubuntu bionic-updates InRelease
Hit:10 http://ppa.launchpad.net/graphics-drivers/ppa/ubuntu bionic InRelease
Hit:11 http://archive.ubuntu.com/ubuntu bionic-backports InRelease
Reading package lists... Done
Building dependency tree
Reading state information... Done
11 packages can be upgraded. Run 'apt list --upgradable' to see them.
Reading package lists... Done
Building dependency tree
Reading state information... Done
Calculating upgrade... Done
The following packages were automatically installed and are no longer required:
  linux-headers-4.15.0-126 linux-headers-4.15.0-126-generic
Use 'sudo apt autoremove' to remove them.
The following packages have been kept back:
  cuda-drivers libcublas-dev libcublas10 libcudnn7 libcudnn7-dev libnccl-dev
  libnccl2 libnvidia-cfg1-430 libnvidia-compute-430 libnvidia-gl-430
  r-cran-usethis
0 upgraded, 0 newly installed, 0 to remove and 11 not upgraded.
Reading package lists... Done
Building dependency tree
Reading state information... Done
inkscape is already the newest version (0.92.3-1).
pandoc is already the newest version (1.19.2.4~dfsg-1build4).
texlive-fonts-recommended is already the newest version (2017.20180305-1).
texlive-generic-recommended is already the newest version (2017.20180305-1).
texlive-xetex is already the newest version (2017.20180305-1).
The following packages were automatically installed and are no longer required:
  linux-headers-4.15.0-126 linux-headers-4.15.0-126-generic
Use 'sudo apt autoremove' to remove them.
0 upgraded, 0 newly installed, 0 to remove and 11 not upgraded.
```

[93]: `%cd /content/drive/MyDrive/Colab Notebooks`

```
/content/drive/MyDrive/Colab Notebooks
```

[91]: `!jupyter nbconvert --to pdf Adult1.ipynb`

```
[NbConvertApp] Converting notebook Adult1.ipynb to pdf
Failed to get connection
** (inkscape:18891): CRITICAL **: 13:26:36.668:
dbus_g_proxy_new_for_name: assertion 'connection != NULL' failed

** (inkscape:18891): CRITICAL **: 13:26:36.668:
dbus_g_proxy_call: assertion 'DBUS_IS_G_PROXY (proxy)' failed
```

```
** (inkscape:18891): CRITICAL **: 13:26:36.668:
dbus_g_connection_register_g_object: assertion 'connection != NULL' failed
[NbConvertApp] Support files will be in Adult1_files/
[NbConvertApp] Making directory ./Adult1_files
[NbConvertApp] Making directory ./Adult1_files
[NbConvertApp] Writing 93422 bytes to ./notebook.tex
[NbConvertApp] Building PDF
[NbConvertApp] Running xelatex 3 times: [u'xelatex', u'./notebook.tex',
'-quiet']
[NbConvertApp] Running bibtex 1 time: [u'bibtex', u'./notebook']
[NbConvertApp] WARNING | bibtex had problems, most likely because there were no
citations
[NbConvertApp] PDF successfully created
[NbConvertApp] Writing 75378 bytes to Adult1.pdf
```