Classification Trees in python

from start to finish

* In this lesson I used scikit-learn and cost complexity pruning to build a classification tree.

Helps in predicting weather a patient has heart disease or not

We will :
1) Import data from a file
2) Identify & deal with missing data
3) Format the data for decision trees
4) Build a preliminary classification tree
5) use, cost complexity pruning to improve the tree
6) Build, Evaluate, Draw and interpret final tree

Classification Trees are an exceptionally useful machine learning method when you need to know how the decision are being made For example, if you have to justify the predictions to your boss, classification Trees are a good method because each step in the decision making process is easy to understand

① Import the Modules

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeClassifier
from sklearn.tree import plot_tree
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import plot_confusion_matrix
```

2) Import the data: (In Pandas read returns data frame)

> df = read_csv ('processed.cleveland.data', header=none)

↳ The above command loads the data into the dataframe df.

lets look at the first 5 rows using the head() function

» df.head()

Rename Columns

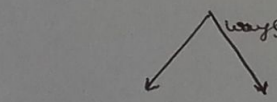
```
df.columns = [ 'age',
               'somethings',
             ]

df.head()
```

keeps column names.

3) Missing data (1)

like forgot to ask age



remove the rows of missing data

impute (educated guess)

the option depends on the importance of it.

df.dtypes # gives the data types of them

print unique values: (of a column)

Quiz [For practise]

1Q) Why do we use one-hot Encoding?

A) So that we sklearn's decision trees can use categorical data

2Q) Sklearn decision trees can handle missing data

A) False

3Q) Why do we deal with missing data before splitting the full database into 'x', just the variables or features that will make classification, and 'y', the known classifications?

A) To ensure that each row in 'x' correctly corresponds to an item in 'y'.

4Q) What's the best way to find alpha for cost complexity pruning?

A) use cross validation

5Q) Why does pruning a tree improve accuracy?

A) Smaller trees ~~are~~ ~~fit~~ donot overfit the training data

Project Name: Classification-Trees-in-python-from-start-to-finish-project

github link:

https://github.com/Suryateja-Kolla/Classification-Trees-in-python-From-Start-To-finish-project