

```
import pandas as pd
import sklearn
import numpy as np
```

```
#read in data
pd.set_option('display.max_columns', None)
p = pd.read_csv("personality_train.csv")
p
```

	#AUIDH	STATUS	sEXT	sNEU	sAGR	sCON	sEXT	cNEU	cAGR	cCON	cOPN	DATE	NETWORK
0	ecbdfbf90e0f83cfd802a7999464c	is stuck on Band-Aid brand, cuz Band-Aid's stu...	4.30	2.15	3.60	3.30	4.10	y	n	y	n	y	12/31/09 10:42 PM
1	b7b7764cfa1c523e4e93ab2a7999464c	likes the sound of thunder.	2.65	3.00	3.15	3.25	4.40	n	y	n	n	y	12/25/09 06:18 PM
2	db39f7b2aad360b1033ec1f8fcd5799c	Back from vacation and tired	4.65	3.20	3.05	3.65	4.75	y	y	n	y	y	12/17/09 02:46 PM
3	4d035bd3fd8d99595d15cea9e388964be	had a great day at church...	3.70	2.90	3.40	3.35	4.05	y	y	n	n	y	11/20/09 01:35 AM
4	5489ed38556fa0f50dc6a93e5d27b95dfb	fiel Time, Inc. has blocked gchat... does this...	4.15	3.10	3.20	3.60	3.80	y	y	n	y	y	11/15/09 04:16 PM
...	...	...	...	...	...	...	...	...	...	...	...	...	...
167	a764ca41dca158d7a191505dccc8ce47f	Red	3.70	2.50	4.20	4.10	3.60	y	n	y	y	y	01/12/10 07:48 PM
168	deb899e426c1a5c66c24eeb0d7df6257	About mornings and winter and magic.	2.15	2.15	4.10	2.90	4.60	n	n	y	n	y	01/11/10 04:19 AM
169	ea2ba927cb6663480ea33ca917c3c8ba	is wishing it was Saturday.	4.05	3.35	3.80	3.95	4.50	y	y	y	y	y	01/09/10 01:01 AM
170	5532642937eb3497a43e15fdb2b3a9d2d	snipers get more head	1.40	4.05	3.30	3.40	3.95	n	y	n	n	y	01/08/10 01:50 AM
171	a286bf7286b1247d4a7851709ef931e1e	Last night was amazing! Not only did I see "PR..."	4.25	3.00	3.25	3.50	4.00	y	y	n	y	y	01/06/10 06:25 PM

172 rows x 20 columns

```
#setup features and labels for the tree

#dependent variables (needed for sorting output later)

labels = p['cNEU']
labels
```

1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026, 2027, 2028, 2029, 2030, 2031, 2032, 2033, 2034, 2035, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2045, 2046, 2047, 2048, 2049, 2050, 2051, 2052, 2053, 2054, 2055, 2056, 2057, 2058, 2059, 2060, 2061, 2062, 2063, 2064, 2065, 2066, 2067, 2068, 2069, 2070, 2071, 2072, 2073, 2074, 2075, 2076, 2077, 2078, 2079, 2080, 2081, 2082, 2083, 2084, 2085, 2086, 2087, 2088, 2089, 2090, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2100, 2101, 2102, 2103, 2104, 2105, 2106, 2107, 2108, 2109, 2110, 2111, 2112, 2113, 2114, 2115, 2116, 2117, 2118, 2119, 2120, 2121, 2122, 2123, 2124, 2125, 2126, 2127, 2128, 2129, 2130, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2143, 2144, 2145, 2146, 2147, 2148, 2149, 2150, 2151, 2152, 2153, 2154, 2155, 2156, 2157, 2158, 2159, 2160, 2161, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2176, 2177, 2178, 2179, 2180, 2181, 2182, 2183, 2184, 2185, 2186, 2187, 2188, 2189, 2190, 2191, 2192, 2193, 2194, 2195, 2196, 2197, 2198, 2199, 2200, 2201, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 2221, 2222, 2223, 2224, 2225, 2226, 2227, 2228, 2229, 2230, 2231, 2232, 2233, 2234, 2235, 2236, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 2248, 2249, 2250, 2251, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266, 2267, 2268, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277, 2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288, 2289, 2290, 2291, 2292, 2293, 2294, 2295, 2296, 2297, 2298, 2299, 2300, 2301, 2302, 2303, 2304, 2305, 2306, 2307, 2308, 2309, 2310, 2311, 2312, 2313, 2314, 2315, 2316, 2317, 2318, 2319, 2320, 2321, 2322, 2323, 2324, 2325, 2326, 2327, 2328, 2329, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2337, 2338, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2347, 2348, 2349, 2350, 2351, 2352, 2353, 2354, 2355, 2356, 2357, 2358, 2359, 2360, 2361, 2362, 2363, 2364, 2365, 2366, 2367, 2368, 2369, 2370, 2371, 2372, 2373, 2374, 2375, 2376, 2377, 2378, 2379, 2380, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2390, 2391, 2392, 2393, 2394, 2395, 2396, 2397, 2398, 2399, 2400, 2401, 2402, 2403, 2404, 2405, 2406, 2407, 2408, 2409, 2410, 2411, 2412, 2413, 2414, 2415, 2416, 2417, 2418, 2419, 2420, 2421, 2422, 2423, 2424, 2425, 2426, 2427, 2428, 2429, 2430, 2431, 2432, 2433, 2434, 2435, 2436, 2437, 2438, 2439, 2440, 2441, 2442, 2443, 2444, 2445, 2446, 2447, 2448, 2449, 2450, 2451, 2452, 2453, 2454, 2455, 2456, 2457, 2458, 2459, 2460, 2461, 2462, 2463, 2464, 2465, 2466, 2467, 2468, 2469, 2470, 2471, 2472, 2473, 2474, 2475, 2476, 2477, 2478, 2479, 2480, 2481, 2482, 2483, 2484, 2485, 2486, 2487, 2488, 2489, 2490, 2491, 2492, 2493, 2494, 2495, 2496, 2497, 2498, 2499, 2500, 2501, 2502, 2503, 2504, 2505, 2506, 2507, 2508, 2509, 2510, 2511, 2512, 2513, 2514, 2515, 2516, 2517, 2518, 2519, 2520, 2521, 2522, 2523, 2524, 2525, 2526, 2527, 2528, 2529, 2530, 2531, 2532, 2533, 2534, 2535, 2536, 2537, 2538, 2539, 2540, 2541, 2542, 2543, 2544, 2545, 2546, 2547, 2548, 2549, 2550, 2551, 2552, 2553, 2554, 2555, 2556, 2557, 2558, 2559, 2560, 2561, 2562, 2563, 2564, 2565, 2566, 2567, 2568, 2569, 2570, 2571, 2572, 2573, 2574, 2575, 2576, 2577, 2578, 2579, 2580, 2581, 2582, 2583, 2584, 2585, 2586, 2587, 2588, 2589, 2590, 2591, 2592, 2593, 2594, 2595, 2596, 2597, 2598, 2599, 2600, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2608, 2609, 2610, 2611, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 2619, 2620, 2621, 2622, 2623, 2624, 2625, 2626, 2627, 2628, 2629, 2630, 2631, 2632, 2633, 2634, 2635, 2636, 2637, 2638, 2639, 2640, 2641, 2642, 2643, 2644, 2645, 2646, 2647, 2648, 2649, 2650, 2651, 2652, 2653, 2654, 2655, 2656, 2657, 2658, 2659, 2660, 2661, 2662, 2663, 2664, 2665, 2666, 2667, 2668, 2669, 2670, 2671, 2672, 2673, 2674, 2675, 2676, 2677, 2678, 2679, 26

```

k8 n
k8 y
T0 y
T1 y
#set.seed(12345) #random seed

#features/independent variables

features =
  ("NETWORKSIZE", "BETWEENNESS", "NBETWEENNESS", "DENSITY", "BROKERAGE", "NBROKERAGE", "T1")

#features =
  ("NETWORKSIZE", "BETWEENNESS", "NBETWEENNESS", "DENSITY", "BROKERAGE", "NBROKERAGE", "T1")

#features = ["sEXTI", "sAGR", "sCON", "sOPN"]

```

```
features

['NETWORKSIZE',
 'BETWEENNESS',
 'BETWEENNESS',
 'DENSITY',
 'BROKERSHAGE',
 'BROKERSHAGE',
 'TRANSITIVITY',
 'sEXT',
 'sACR',
 'sCCM',
 'sOPM']

#get dataframe of just features

#get all rows and just the columns that match our features

X = p.loc[:,features]
y
```

[illegible]

	471	50135.0	31.18	0.02	37733	0.49	0.12	3.18	4.28	4.18	3.08
168	36	185.71	31.21	0.40	377	0.32	0.63	2.15	4.10	2.90	4.60
169	83	2935.76	88.40	0.08	3120	0.47	0.26	4.05	3.80	3.95	4.50
170	154	11424.50	98.25	0.02	11510	0.49	0.05	1.40	3.30	3.40	3.95
171	539	138337.00	95.77	0.02	142460	0.49	0.13	4.25	3.25	3.50	4.00

172 rows x 11 columns

```
#setup plot for the confusion matrix and decision tree
import matplotlib.pyplot as plt
print(plt.rcParams.get('figure.figsize'))
```

```
Figure size (100, 100)
```

```
#setup figure size
fig_size = plt.rcParams["figure.figsize"]
fig_size[0] = 10
fig_size[1] = 10
plt.rcParams["figure.figsize"] = fig_size
```

```
#output/labels once more for naming
Y = p["c2NU"]
```

114

1	n
2	y
3	y
4	y
5	y
6	y
7	y
8	y
9	y
10	y
11	y
12	y
13	y
14	y
15	y
16	y
17	y
18	y
19	y
20	y
21	y
22	y
23	y
24	y
25	y
26	y
27	y
28	y
29	y
30	y
31	y
32	y
33	y
34	y
35	y
36	y
37	y
38	y
39	y
40	y
41	y
42	y
43	y
44	y
45	y
46	y
47	y
48	y
49	y
50	y
51	y
52	y
53	y
54	y
55	y
56	y
57	y
58	y
59	y
60	y
61	y
62	y
63	y
64	y
65	y
66	y
67	y
68	y
69	y
70	y
71	y
72	y
73	y
74	y
75	y
76	y
77	y
78	y
79	y
80	y
81	y
82	y
83	y
84	y
85	y
86	y
87	y
88	y
89	y
90	y
91	y
92	y
93	y
94	y
95	y
96	y
97	y
98	y
99	y
100	y

[illegible]

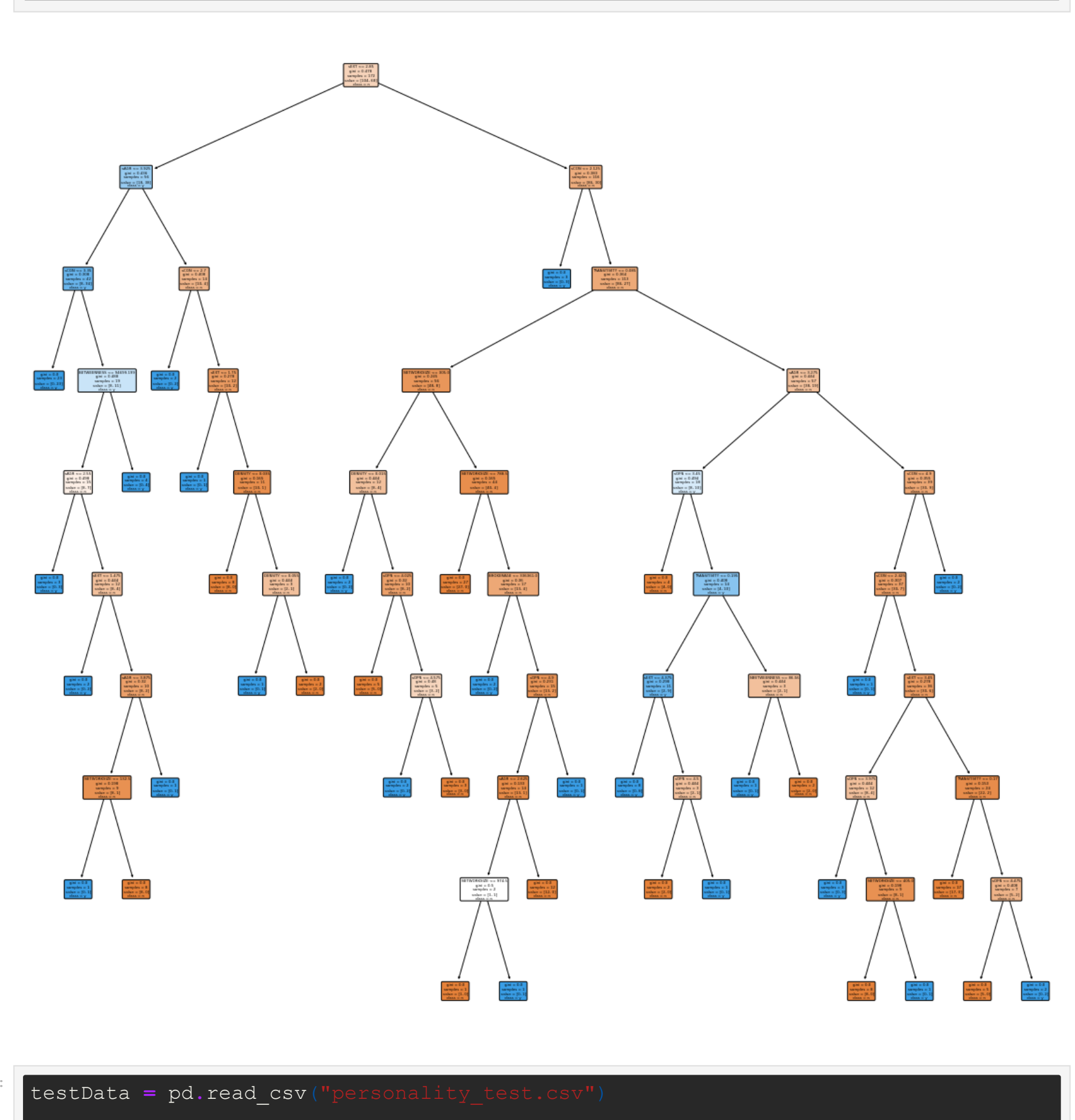
```
clf = tree.DecisionTreeClassifier(random_state=)
clf = clf.fit(X, Y)

sorted = labels.unique()
```

```
sorted = np.sort(sorted)
sorted = list(map(str, sorted))
sorted
```

```
from pandas.plotting import scatter_matrix
#X =
tree.plot_tree(clf,feature_names=features,class_names=labels.astype(str),rounded=0)
```

```
x =
tree.plot_tree(clf,rounded=True,filled=True,class_names=sorted,feature_names=feature_names)
plt.savefig('out.pdf')
```



	#AUTHID	STATUS	sEXT	sNEU	sAGR	sC
0	200255966ca6e2636535b5b93ac04497	Four day camping trip!!! Holy Crap! this	3.15	1.90	4.15	4

1	527ed53d2ba3bc417b8402d5b2f556	put in work again last night at the Pig N Whis...	3.35	2.75	2.85	3.10	4.15	n	n	n	n	y	12/25/09 04:47 PM
2	318f8822d4f2bd3920367560218619c0	has bed bugs.... ewwwwww!	4.50	4.00	3.00	4.50	3.75	y	y	n	y	n	12/23/09 10:06 PM
3	2badc4f7503a8766e89c6626d1130969a	NYC on 8/18!!!!!!!	3.55	2.30	3.65	4.65	4.60	n	n	y	y	y	12/08/09 01:27 AM
4	1c10cc0852579d21a000c9c3327bb98	for the united fans reading, dont worry be hap...	2.60	2.65	2.10	2.20	3.50	n	n	n	n	n	12/06/09 08:40 PM
...	...	...	...	...	...	...	...	...	...	...	...	...	...
69	06b055f8e2bca96496514891057913c3	is enjoying the cricket...comfy boxes and rail...	2.85	2.35	3.35	4.70	3.35	n	n	n	y	n	06/16/09 06:22 AM
70	138ac63ec2b55b84f8d19c300720cae	Can anyone out there tell me if Jesus was EVER...	1.95	3.45	3.05	2.50	3.95	n	y	n	n	y	05/15/09 11:54 PM
71	3fe44fab3eb561ae418a22182ec75fad	Debating on whether or not to drink...alcohol...	4.00	3.75	3.25	2.00	3.75	y	y	n	n	n	02/15/10 03:36 AM
72	3fe44fab3eb561ae418a22182ec75fad	is really into the Coteaux Twins and Depeche M...	4.00	3.75	3.25	2.00	3.75	y	y	n	n	n	02/12/10 10:47 PM
73	35eft99775d5ee7e83c7f912591984d5	Facebook me marea. Me hates it long time T-T	2.45	4.00	2.85	2.35	4.10	n	y	n	n	y	01/15/10 02:19 AM

74 rows × 20 columns

```
XTest = testData.loc[:, :features
```

XTest

	NETWORKSIZE	BETWEENNESS	NBETWEENNESS	D
0	75	2650.67	98.14	
1	789	303058.00	97.74	
2	318	49024.80	97.88	

	379	66420.90	93.22	0.03	69191	0.49	0.27	3.55	3.65	4.65	4.60
4	139	9100.53	96.27	0.04	9251	0.49	0.11	2.60	2.10	2.20	3.50
...	...	...	...	...	...	...	...	...	...	...	...
69	194	18123.10	97.81	0.02	18313	0.49	0.06	2.85	3.35	4.70	3.35
70	415	84621.10	98.98	0.01	84969	0.50	0.04	1.95	3.05	2.50	3.95
71	329	49454.70	92.22	0.03	52282	0.49	0.17	4.00	3.25	2.00	3.75
72	329	49454.70	92.22	0.03	52282	0.49	0.17	4.00	3.25	2.00	3.75
73	65	1696.85	84.17	0.10	1876	0.47	0.28	2.45	2.85	2.35	4.10

74 rows x 11 columns

```
YTest = testData["cNEU"]
YTest
```

```

0      n
1      y
2      n
3      n
4      .
5      .
6      .
7      .
8      .
9      .
10     y
11     y
12     y
13     y
14     y

```

Data source: Stanford University

11

```
YPredicted
```

```
array(['n', 'y', 'n', 'y', 'y',  
      'n', 'n', 'y', 'n', 'n',  
      'y', 'n', 'n', 'n', 'n',  
      'n', 'n', 'n', 'n', 'y'])
```

[illegible]

7	y
8	n
9	n
10	n
11	n
12	y
13	y
14	y
15	y
16	y
17	y
18	y
19	n
20	y
21	y
22	y

```
from sklearn import metrics
accuracy = metrics.accuracy_score(YTest, YPredicted)
accuracy
```

```
#setup plots for confusion matrix
from sklearn.metrics import plot_confusion_matrix as matrix
```

```
figSize = plt.rcParams["figure.figsize"]
figSize[0] = 30
figSize[1] = 3
plt.rcParams["figure.figsize"]=figSize
```

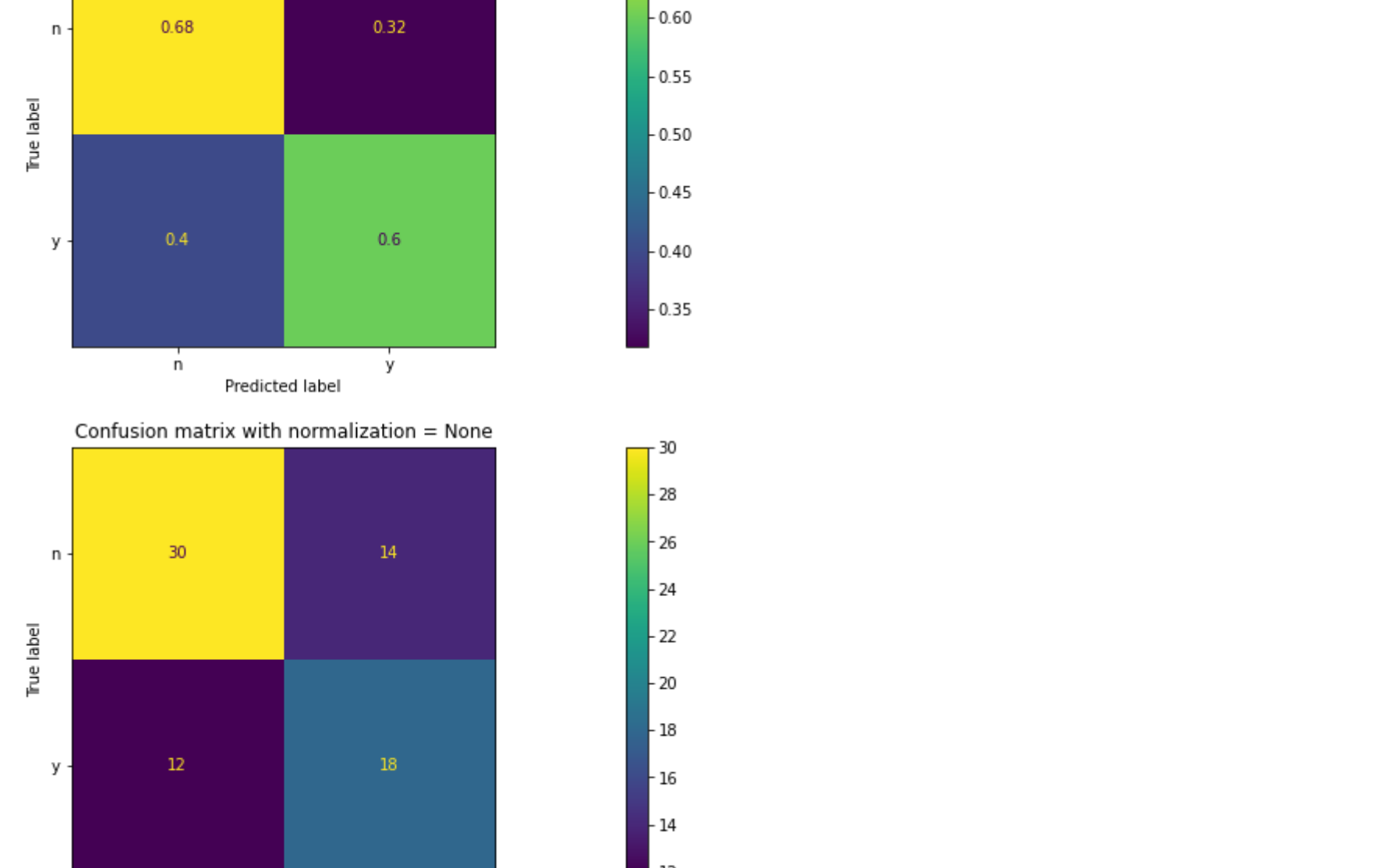
```
print(pic.parameters.get('figure(figsize',

#plot the confusion matrices 1 for normalized the other un-normalized
values = ['true',None]
#open figure
```

```
for x in values:
    disp = matrix(clf.XTest.YTest,display_labels=sorted,normalize=x)
    disp.ax_.set_title("Confusion matrix with normalization = "+str(x))
print(disp.confusion_matrix)
```

```
[[0.0, 0.0]
 [0.0, 1.0]
 [0.0, 0.0]]
```

Confusion matrix with normalization = true



```
#get false positives
#pd.set_option('display.max_rows',100)
```

	#AUTHID	STATUS	sEXT	sNEU	sAGR	sCON	sOPN	cEXT	cNEU	cAGR	cCON	cOPN	DATE	NETWORK
1	527ed53d2b3a3bc417b8402d5b27b5f56	put in work again last night at the Ptg N Whis...	3.35	2.75	2.85	3.10	4.15	n	n	n	n	y	12/25/09 04:47 PM	
3	2badb47503a98766c89e266d1130969a	NYC on 8/18!!!!!!!	3.55	2.30	3.65	4.65	4.60	n	n	y	y	y	12/08/09 01:27 AM	
4	1c10cc0852579d2fa00e3fc3327b2b98	for the united fans reading, dont worry be hap...	2.60	2.65	2.10	2.20	3.50	n	n	n	n	n	12/06/09 08:40 PM	
8	526ac26353b3f5ce0ee5d742d48e83e9107	A vegetarian delight	2.45	2.70	3.10	4.00	4.30	n	n	n	y	y	10/29/09 06:04 AM	
9	370a8295d2f28b9069e75423c37d2b639	IS SOOOOOO EXCITED AT THE HOLDS	4.00	1.65	3.10	3.10	4.50	y	n	n	n	y	10/23/09 03:15 AM	
19	225c97c90103cc04cda7f10845f2733e	has been asked by several friends about our tr...	2.70	1.45	3.80	4.35	3.30	n	n	y	y	n	08/24/09 12:30 PM	
22	1bd281623f6a26d08caa394cdad75c7d	Nothing to do in lab because the TAs messed up...	2.10	2.70	3.40	2.05	3.65	n	n	n	n	n	08/15/09 04:35 PM	
23	0bfa3952f9ed50f25011b128e73ab820	is glad "PROPNAM" has taken responsibility fo...	2.80	1.60	3.65	3.25	4.15	n	n	n	n	y	08/09/09 10:59 PM	
37	3b47ab4bc02985674f2b64e46a939b3bd	Me + Exams = Epic Fail -..	3.60	1.95	4.20	3.60	4.80	y	n	y	y	y	07/10/09 09:43 PM	
38	301e1788a595203d0da9f3fed10afe1c9	Its way to early on a Saturday to be heading t...	3.00	2.50	4.00	3.50	3.75	n	n	y	y	n	07/10/09 02:33 PM	
44	0ea88660e23db79e2a3c897a54900df	is wondering how many people know that the Uni...	4.25	2.00	3.25	4.75	3.50	y	n	n	n	n	07/05/09 10:28 PM	
45	1d6d222b3fb4c0af35466042cb82d78	Michael Jackson is dead! What the world com...	1.95	2.40	3.80	2.80	2.70	n	n	y	n	n	07/04/09 02:28 PM	
48	5880081cd3bde1619cd431a75d9052dfc	ran 19 miles today Shower a big glass of wat...	2.90	2.20	3.05	3.90	4.10	n	n	n	y	y	06/28/09 11:20 PM	
52	4e5bc97d9f3aca05a420bdfda3ca2639	can't wait to have all the stuff moved into th...	4.65	2.20	3.70	2.10	3.85	y	n	y	n	y	06/25/09 02:42 AM	
<pre>#get false negative #pd.set_option('display.max_rows',100) testData[(YTest!=YPredicted &amp; YPredicted=="n")]</pre>														
	#AUTHID	STATUS	sEXT	sNEU	sAGR	sCON	sOPN	cEXT	cNEU	cAGR	cCON	cOPN	DATE	NETWORK
2	318bf822d4f22bd3920367560218619c	has bed bugs.... ewwww!!!	4.50	4.00	3.00	4.50	3.75	y	y	n	y	n	12/23/09 10:06 PM	
7	172400f46880b309ca5e97d322b68f01	I have no excuses, least of all for God. Like...	3.45	2.85	2.80	2.70	4.15	n	y	n	n	y	11/18/09 04:56 AM	
12	450c787001b004a6f9428e267c7a4ca1	Writing, then 2 chapters homework, then bed. O...	2.30	3.50	4.50	2.85	4.50	n	y	y	n	y	09/25/09 09:53 AM	
14	4cac659f923d6f3b4605f38477a04458	I have no water right now. WHY	2.80	3.15	3.65	3.45	4.35	n	y	y	n	y	09/20/09 12:06 AM	
17	448084546da4ae45e47f3a8f338ade56	ive come to the conclusion that some people ar...	3.70	3.15	4.30	3.75	3.85	y	y	y	y	y	08/31/09 05:03 AM	
25	259f8f9c95b214dc3924af48bcdf0	is back from D.C. and is very tired...	3.30	2.85	4.20	3.40	4.75	n	y	y	n	y	08/06/09 07:35 PM	1
30	03e6c4eca2693c1839f0e1780f73faba	*Those who criticize our generation forget who...	3.20	3.60	3.85	4.35	4.80	n	y	y	y	y	07/23/09 05:55 PM	
31	3cc2cbf4bc8c5c9f005a092a9e9acab	Damn, U.S. beat Spain? We should be proud	2.75	3.75	3.75	4.75	4.25	n	n	y	y	y	07/19/09 05:40 PM	
39	1187ed8a8b100eb9b864ac30df6ad29	why does it seem like im always	4.00	4.00	2.25	4.50	3.25	y	y	n	y	n	07/09/09 07:16	

