Data_Processing

June 21, 2017

In [ ]:
```
#Imporing Library
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
```

In [5]:
```
#Importing Dataset
data_frame=pd.read_csv('/home/tanmoy/Desktop/MachineLearning/pima-data.csv')
```

In [7]:
```
#Display the data set size row and col
data_frame.shape
```

Out[7]: (768, 10)

In [9]:
```
#display first 3 row of the dataset
data_frame.head(3)
```

Out[9]:

	num_preg	glucose_conc	diastolic_bp	thickness	insulin	bmi	diab_pred	age	skin	diabetes
0	6	148	72	35	0	33.6	0.627	50	1.3790	True
1	1	85	66	29	0	26.6	0.351	31	1.1426	False
2	8	183	64	0	0	23.3	0.672	32	0.0000	True

In [11]:
```
#display last 4 row of the dataset
data_frame.tail(4)
```

Out[11]:

	num_preg	glucose_conc	diastolic_bp	thickness	insulin	bmi	diab_pred	age	skin	diabetes
764	2	122	70	27	0	36.8	0.340	27	1.0638	False
765	5	121	72	23	112	26.2	0.245	30	0.9062	False
766	1	126	60	0	0	30.1	0.349	47	0.0000	True
767	1	93	70	31	0	30.4	0.315	23	1.2214	False

In [13]:
```
#If there is any empty cell in dataset then isnull value is true otherwise false in thi
data_frame.isnull()
```

Out[13]:

	num_preg	glucose_conc	diastolic_bp	thickness	insulin	bmi	\
0	False	False	False	False	False	False	
1	False	False	False	False	False	False	
2	False	False	False	False	False	False	
3	False	False	False	False	False	False	
4	False	False	False	False	False	False	
5	False	False	False	False	False	False	
6	False	False	False	False	False	False	
7	False	False	False	False	False	False	
8	False	False	False	False	False	False	
9	False	False	False	False	False	False	
10	False	False	False	False	False	False	
11	False	False	False	False	False	False	
12	False	False	False	False	False	False	
13	False	False	False	False	False	False	
14	False	False	False	False	False	False	
15	False	False	False	False	False	False	
16	False	False	False	False	False	False	
17	False	False	False	False	False	False	
18	False	False	False	False	False	False	
19	False	False	False	False	False	False	
20	False	False	False	False	False	False	
21	False	False	False	False	False	False	
22	False	False	False	False	False	False	
23	False	False	False	False	False	False	
24	False	False	False	False	False	False	
25	False	False	False	False	False	False	
26	False	False	False	False	False	False	
27	False	False	False	False	False	False	
28	False	False	False	False	False	False	
29	False	False	False	False	False	False	
..	...	...	...	...	...	...	
738	False	False	False	False	False	False	
739	False	False	False	False	False	False	
740	False	False	False	False	False	False	
741	False	False	False	False	False	False	
742	False	False	False	False	False	False	
743	False	False	False	False	False	False	
744	False	False	False	False	False	False	
745	False	False	False	False	False	False	
746	False	False	False	False	False	False	
747	False	False	False	False	False	False	
748	False	False	False	False	False	False	
749	False	False	False	False	False	False	
750	False	False	False	False	False	False	

751	False	False	False	False	False	False
752	False	False	False	False	False	False
753	False	False	False	False	False	False
754	False	False	False	False	False	False
755	False	False	False	False	False	False
756	False	False	False	False	False	False
757	False	False	False	False	False	False
758	False	False	False	False	False	False
759	False	False	False	False	False	False
760	False	False	False	False	False	False
761	False	False	False	False	False	False
762	False	False	False	False	False	False
763	False	False	False	False	False	False
764	False	False	False	False	False	False
765	False	False	False	False	False	False
766	False	False	False	False	False	False
767	False	False	False	False	False	False

	diab_pred	age	skin	diabetes
0	False	False	False	False
1	False	False	False	False
2	False	False	False	False
3	False	False	False	False
4	False	False	False	False
5	False	False	False	False
6	False	False	False	False
7	False	False	False	False
8	False	False	False	False
9	False	False	False	False
10	False	False	False	False
11	False	False	False	False
12	False	False	False	False
13	False	False	False	False
14	False	False	False	False
15	False	False	False	False
16	False	False	False	False
17	False	False	False	False
18	False	False	False	False
19	False	False	False	False
20	False	False	False	False
21	False	False	False	False
22	False	False	False	False
23	False	False	False	False
24	False	False	False	False
25	False	False	False	False
26	False	False	False	False
27	False	False	False	False
28	False	False	False	False

```
29       False  False  False     False
..         ...    ...    ...       ...
738      False  False  False     False
739      False  False  False     False
740      False  False  False     False
741      False  False  False     False
742      False  False  False     False
743      False  False  False     False
744      False  False  False     False
745      False  False  False     False
746      False  False  False     False
747      False  False  False     False
748      False  False  False     False
749      False  False  False     False
750      False  False  False     False
751      False  False  False     False
752      False  False  False     False
753      False  False  False     False
754      False  False  False     False
755      False  False  False     False
756      False  False  False     False
757      False  False  False     False
758      False  False  False     False
759      False  False  False     False
760      False  False  False     False
761      False  False  False     False
762      False  False  False     False
763      False  False  False     False
764      False  False  False     False
765      False  False  False     False
766      False  False  False     False
767      False  False  False     False

[768 rows x 10 columns]
```

In [15]:
```
#This function returns an array of the result
data_frame.isnull().values
```

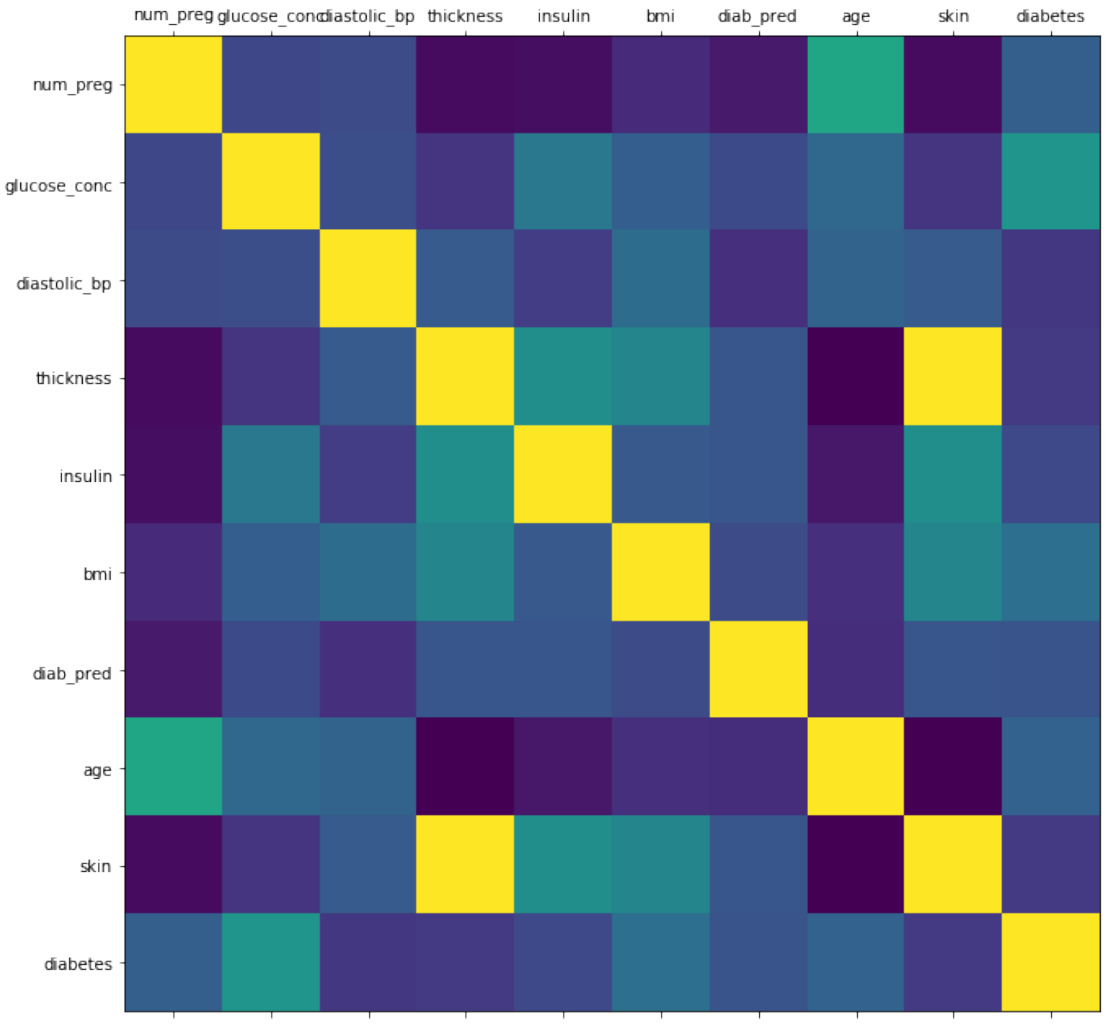
Out[15]:
```
array([[False, False, False, ..., False, False, False],
       [False, False, False, ..., False, False, False],
       [False, False, False, ..., False, False, False],
       ...,
       [False, False, False, ..., False, False, False],
       [False, False, False, ..., False, False, False],
       [False, False, False, ..., False, False, False]], dtype=bool)
```

In [19]:
```
#This function check if there is any empty cell in the data set or not
data_frame.isnull().values.any()
```

Out[19]: False

In [31]:
```
#function for display the heatmap here yellow cell means the data in this cell are same
def corr_heatmap(data_frame, size=11):
    correlation=data_frame.corr()
    fig, heatmap = plt.subplot(figsize=(size,size))
    heatmap.matshow(correlation)
    plt.xticks(range(len(correlation.columns)),correlation.columns)
    plt.yticks(range(len(correlation.columns)),correlation.columns)
    plt.show()
```

In [34]:
```
#call the above funciton
corr_heatmap(data_frame)
```

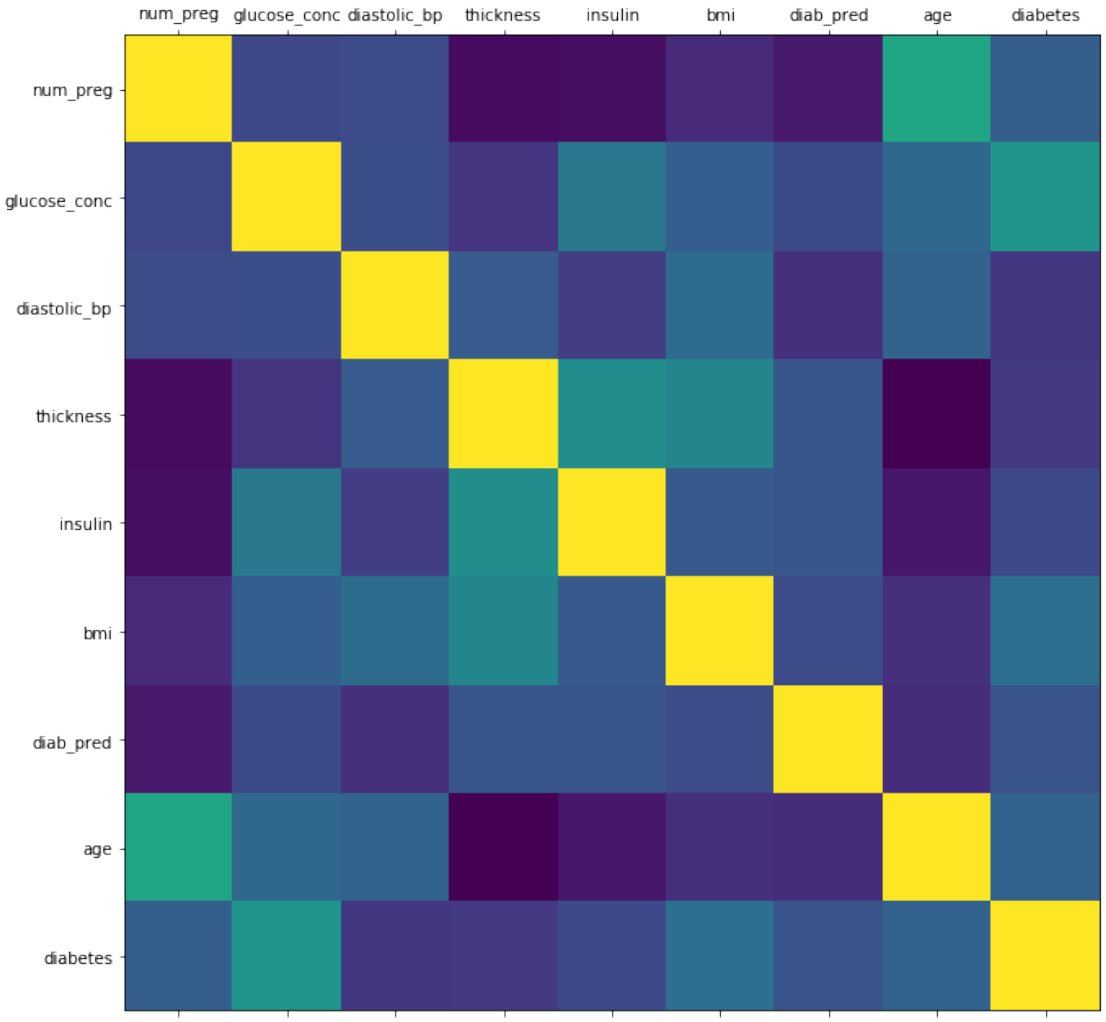


In [37]:
```
#here we delete the duplicate date col skin and preserve the thickness column
del data_frame['skin']
data_frame.head()
```

Out[37]:

	num_preg	glucose_conc	diastolic_bp	thickness	insulin	bmi	diab_pred	age	diabetes
0	6	148	72	35	0	33.6	0.627	50	True
1	1	85	66	29	0	26.6	0.351	31	False
2	8	183	64	0	0	23.3	0.672	32	True
3	1	89	66	23	94	28.1	0.167	21	False
4	0	137	40	35	168	43.1	2.288	33	True

In [39]:
```
#This is the dataset heatmap after deleting the duplicate col skin
corr_heatmap(data_frame)
```



In [41]:
```
data_frame.head()
```

Out[41]:

	num_preg	glucose_conc	diastolic_bp	thickness	insulin	bmi	diab_pred	\
0	6	148	72	35	0	33.6	0.627	
1	1	85	66	29	0	26.6	0.351	
2	8	183	64	0	0	23.3	0.672	
3	1	89	66	23	94	28.1	0.167	
4	0	137	40	35	168	43.1	2.288	

	age	diabetes
0	50	True
1	31	False
2	32	True
3	21	False
4	33	True

In [43]:
```
#Data Molding = Converting from True False to binary 0 or 1
map_diabetis={True:1,False:0}
data_frame['diabetes']=data_frame['diabetes'].map(map_diabetis)
```

In [45]:
```
data_frame.head()
```

Out[45]:

	num_preg	glucose_conc	diastolic_bp	thickness	insulin	bmi	diab_pred	\
0	6	148	72	35	0	33.6	0.627	
1	1	85	66	29	0	26.6	0.351	
2	8	183	64	0	0	23.3	0.672	
3	1	89	66	23	94	28.1	0.167	
4	0	137	40	35	168	43.1	2.288	

	age	diabetes
0	50	1
1	31	0
2	32	1
3	21	0
4	33	1

In [46]:
```
#Checking True false Ratio
cnt_true=0.0
cnt_false=0.0
for item in data_frame['diabetes']:
    if(item==1):
        cnt_true+=1
    else:
        cnt_false+=1
percent_true=(cnt_true/(cnt_true+cnt_false))*100
percent_false=(cnt_false/(cnt_true+cnt_false))*100

print "Number of True cases: {0} ({1:2.2f}%)".format(cnt_true,percent_true)
print "Number of False cases: {0} ({1:2.2f}%)".format(cnt_false,percent_false)
```

```
Number of True cases: 268.0 (34.90%)
Number of False cases: 500.0 (65.10%)
```