Univarient Feature selection

In [1]:
```
#feature extraction with univarient stastical test(chi sqaure for classification)
from  sklearn import datasets
import pandas as pd
from sklearn.feature_selection import chi2,SelectKBest
```

In [2]:
```
data=pd.read_csv('diabetes.csv')
data.head()
```

Out[2]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcom
0	6	148	72	35	0	33.6	0.627	50	
1	1	85	66	29	0	26.6	0.351	31	
2	8	183	64	0	0	23.3	0.672	32	
3	1	89	66	23	94	28.1	0.167	21	
4	0	137	40	35	168	43.1	2.288	33	

In [3]:
```
array=data.values
array #change data frame in array format
```

Out[3]:
```
array([[  6.   , 148.   ,  72.   , ...,   0.627,  50.   ,   1.   ],
       [  1.   ,  85.   ,  66.   , ...,   0.351,  31.   ,   0.   ],
       [  8.   , 183.   ,  64.   , ...,   0.672,  32.   ,   1.   ],
       ...,
       [  5.   , 121.   ,  72.   , ...,   0.245,  30.   ,   0.   ],
       [  1.   , 126.   ,  60.   , ...,   0.349,  47.   ,   1.   ],
       [  1.   ,  93.   ,  70.   , ...,   0.315,  23.   ,   0.   ]])
```

In [4]:
```
x=array[:,0:8]
y=array[:,8]  #divide data in input and output columns
```

In [5]:
```
test=SelectKBest(score_func=chi2,k=4) #selectbest will find out best value of k
score=test.fit(x,y)  #fit the data
```

In [6]:
```
print(score.scores_)
#will print 4 result which having higest result.
# here 5 th column have high probability and impact high on result.
#5th column has higest probability its isuline columns
```

```
[ 111.51969064 1411.88704064   17.60537322   53.10803984 2175.56527292
  127.66934333    5.39268155  181.30368904]
```