Tkagg – used to display plots in separate window

Pca – used to reduce number of columns, to remove unwanted noise

Higher eigen value = more important the principal component

Eigen value = 1 is the cutoff according to kaiser rule

Factor analysis – similar to pca but used to find hidden factors

Mann-whiteney u test –

H0(null hypothesis) : there is no significant difference between phishing and legitimate url = p>0.05

H1(alternate hypothesis): there is difference between phishing and legitimate url = p<0.05

Phishing – 1 , legitimate – 0

If p value < 0.05 we reject the null hypothesis

Clustering – high risk url are scattered because the patterns of high risk url are distinct they are made intentionally to attack using diverse tactics