**CSE523 Machine Learning**

**Prof. Mehul Raval**

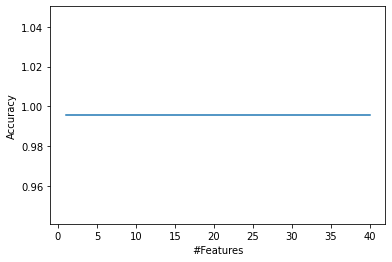
**Anomaly detection in computer networks to identify unusual activity or potential security threats**

**Week 4 Report**

| **Name** | **Enrolment Number** |
| --- | --- |
| Vatsal Patel | AU2040043 |
| Shreya Karia | AU2040076 |
| Digant Patel | AU2040086 |
| Samarth Chauhan | AU2040097 |

We had implemented kNN with different values of features obtained through PCA(Principal Component Analysis).

The line chart for the number of features vs. model accuracy on the test data for the kNN algorithm:

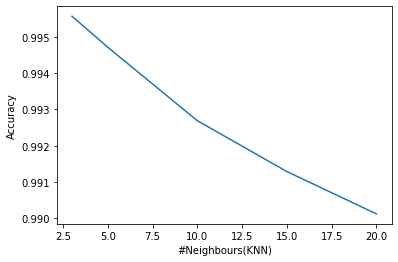


The accuracy of k-Nearest Neighbors (kNN) does not change with different numbers of features after performing Principal Component Analysis (PCA); it could be due to the following reason:

* The first few principal components may capture most of the variability in the data: PCA aims to reduce the dimensionality of the data while retaining most of the information. If the first few principal components already capture most of the variability in the data, reducing the number of features may not significantly impact the accuracy of kNN.
* kNN is a non-parametric algorithm that works by finding the k nearest neighbors to a given point. It does not make any assumptions about the data distribution, and the number of features may not significantly impact the algorithm's accuracy.

Due to this observation, we concluded to change the value of k in the kNN algorithm to different values to see how it affects the accuracy.

The line chart for the different number of k values on the test data for the kNN algorithm:



We can see that the accuracy decreases with the increasing value of k. This can be attributed to the fact that with the increase in the value of k, the model becomes too generalized, and the important details are ignored. And decreasing the value to less than three would lead to overfitting.

Hence, three seems to be the optimal value of k as it would balance the tradeoff between underfitting and overfitting.