**CSE523 Machine Learning**

**Prof. Mehul Raval**

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

**Week 5 Report**

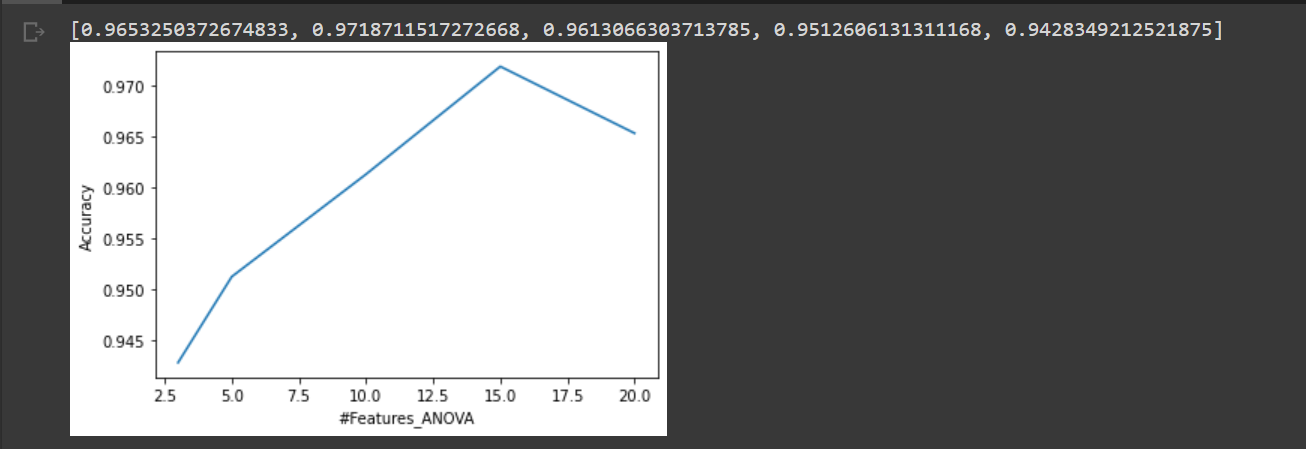
| **Name** | **Enrolment Number** |
| --- | --- |
| Vatsal Patel | AU2040043 |
| Shreya Karia | AU2040076 |
| Digant Patel | AU2040086 |
| Samarth Chauhan | AU2040097 |

We have used PCA( Principal Component Analysis) for reducing the dimensions of our dataset. This has led to high accuracy. We also wanted to use ANOVA to see which dimension reductionality technique would work best: PCA or ANOVA or a combination of both. So we implemented ANOVA this week.

**ANOVA Test:**

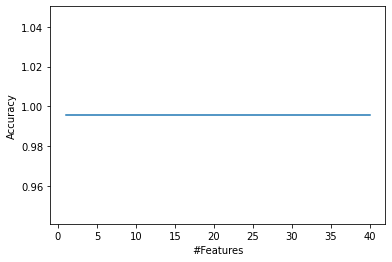
ANOVA (Analysis of Variance) is a statistical method used to select the most important features that explain the variability of the data to reduce the dimensions. This means that by using ANOVA,, we can select the features which discriminate between the classes well. This helps in reducing the features and reduces the risk of overfitting. Since ANOVA can handle a large number of features well, it is a good choice for our dataset, and it is of high dimensionality.

**Accuray vs Features for kNN using ANOVA**



Firstly, to perform ANOVA test, any column of the data should not consist of variance equal to zero. Hence, we defined a variance threshold of 0 to remove columns consisting of constant values.

Based on the score of the ANOVA test, we selected a different number of total top features. After that, we ran the kNN algorithm for a fixed number of k = 3 and plotted the graph of number of features vs accuracy.



We were getting the same accuracy for different number of features in kNN with PCA but with ANOVA there is some significant difference in accuracy. In PCA, we were creating new features by projecting existing one’s and some of the principal components captured most of the variation. But in ANOVA test, top features are picked up as they are sorted by their scores.