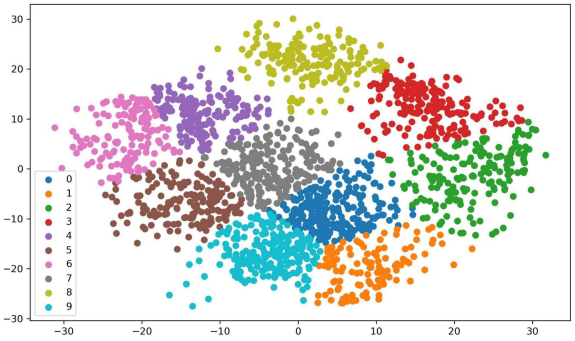
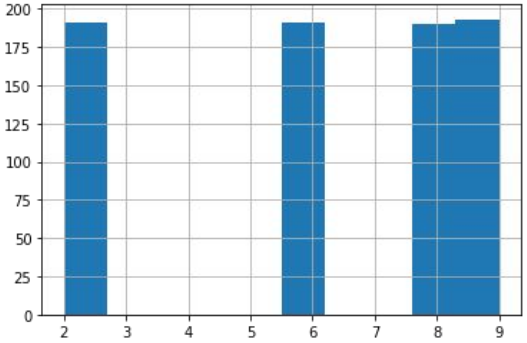
Assignment 3

Contributions:-
Q3a&b: Ananya
Q3c: Jahnvi+Prachi
Q3d: Manvi+Ananya
Q3e: Avishi+Prachi

Dataset

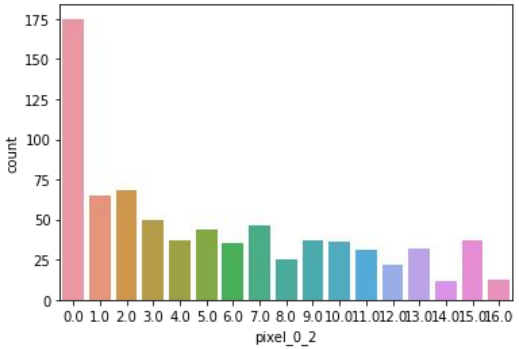
We selected the **digits dataset** for our anomaly detection tasks. It has 10 classes in it out of which we chose the digit classes **2,6,8,9** as our **target classes**. Then we selected few points from **other classes** to be introduced as **outliers**. These points were given labels from one of the 4 classes used for classification.

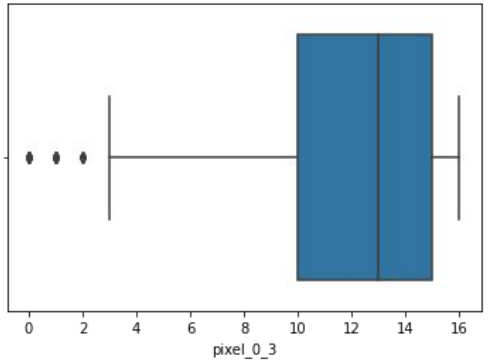


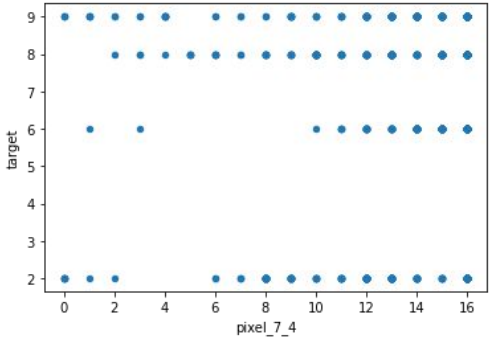


Statistical Tests

Different plots - Countplot, Boxplot, Scatterplot







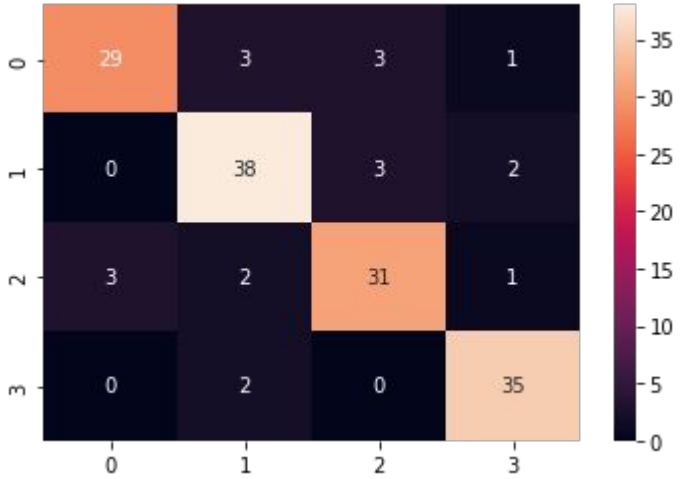
Baseline and Results.

The Baseline model that we run for the classification task is Decision Tree classifier with its default parameter values and it gave us an accuracy of 86.92%. The classification report is as follows:

Class	precision	recall	f1-score	support
2	0.90	0.86	0.88	42
6	0.94	0.89	0.92	38
8	0.79	0.76	0.77	29
9	0.82	0.91	0.86	44
accuracy			0.86	153
macro avg	0.86	0.85	0.86	153
weighted avg	0.87	0.86	0.86	153

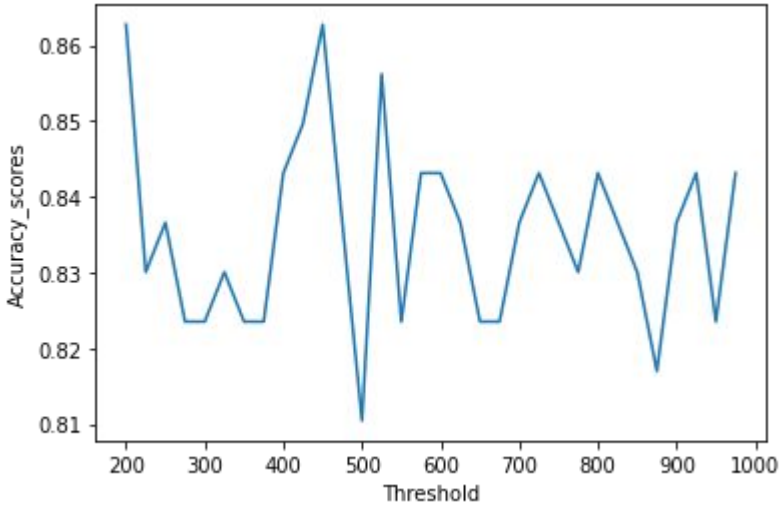
Using Dimensionality Reduction

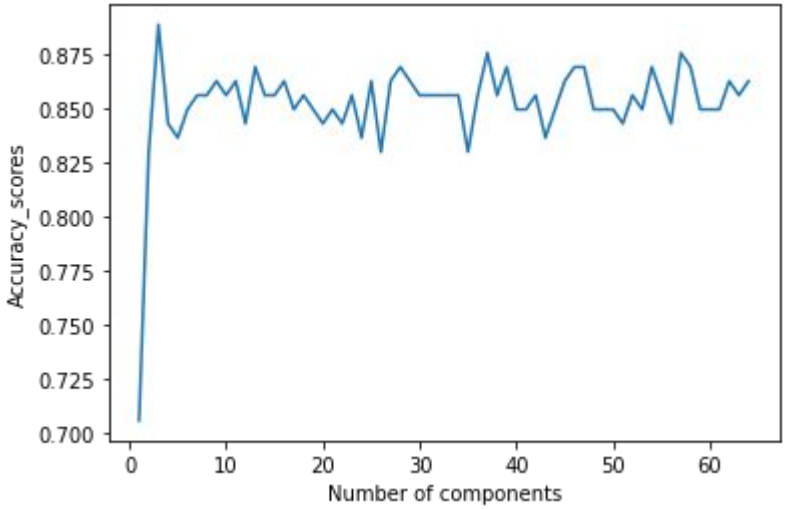
Dimensionality reduction techniques are used in reconstruction-based approach to anomaly detection. We used PCA to find the principal components and then measured the reconstruction error of each object. The objects with large reconstruction errors were classified as anomalies.



Accuracy = 0.8888888888888888

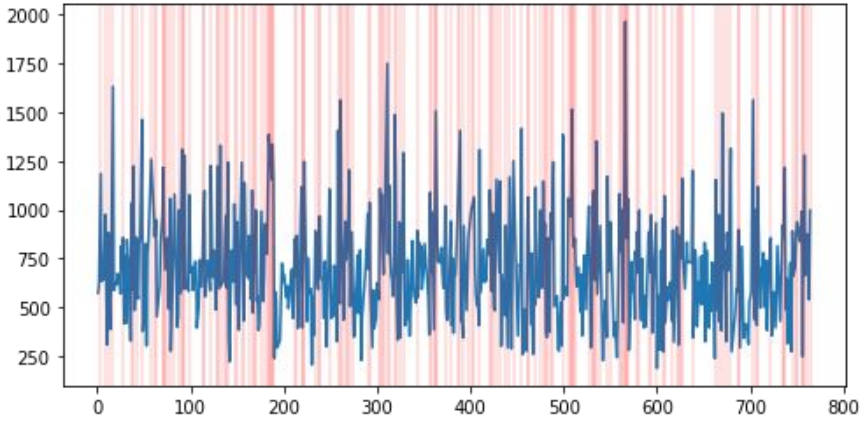
	precision	recall	f1-score	support
2	0.91	0.86	0.88	35
6	0.93	0.95	0.94	44
8	0.76	0.81	0.78	31
9	0.93	0.91	0.92	43
accuracy			0.89	153
macro avg	0.88	0.88	0.88	153
weighted avg	0.89	0.89	0.89	153





166 data points were removed from the training set to get 89% accuracy.

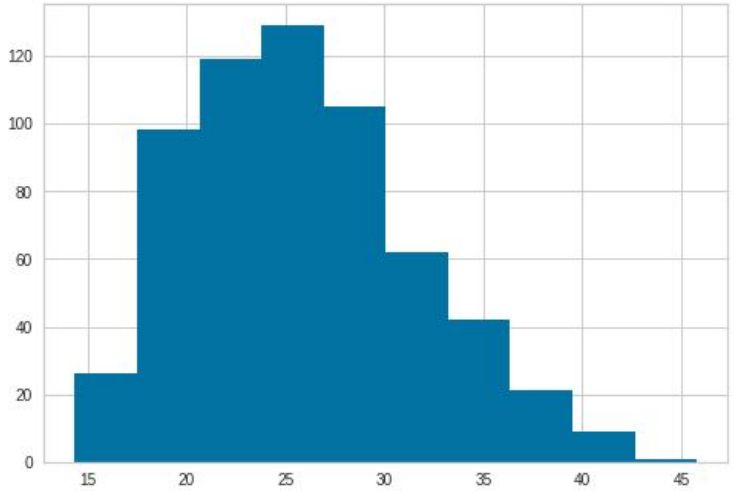
We infer from the above plots that the accuracy saturates with an increase in the number of components because the higher eigenvectors contain lesser information than the earlier ones

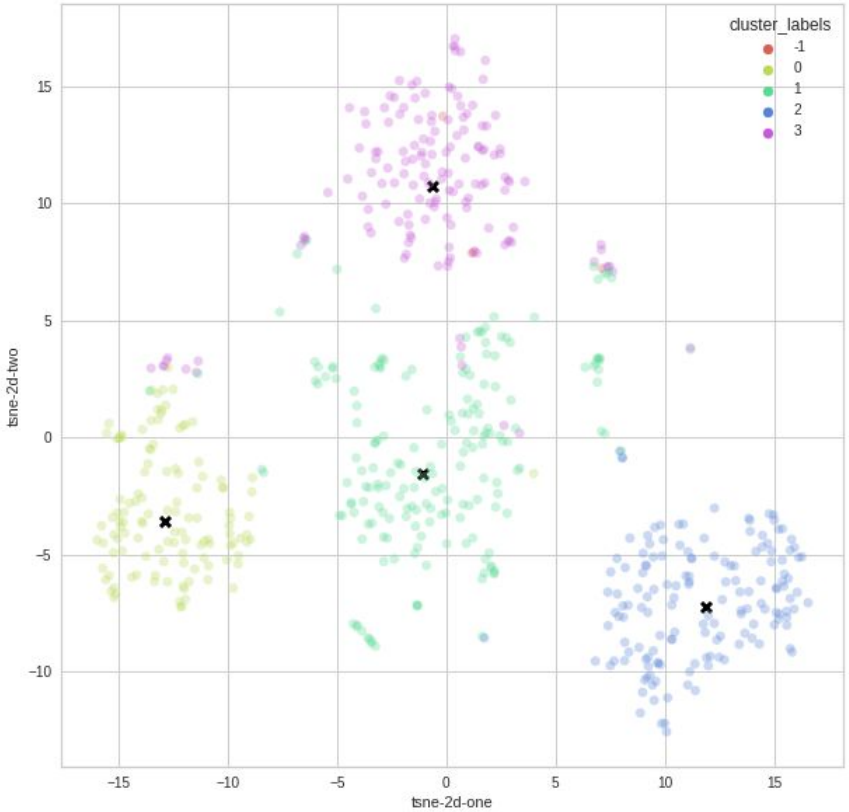


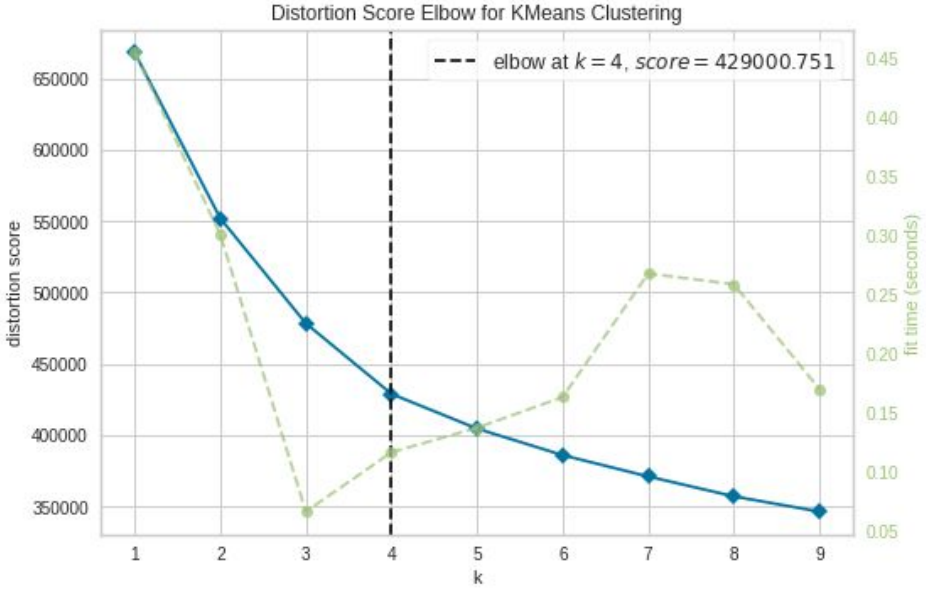
Using Clustering

Method used - KMeans Clustering

Find the distance to the closest centroid and plot a histogram. Find the points that are very far away from any centroid and are also few in number.







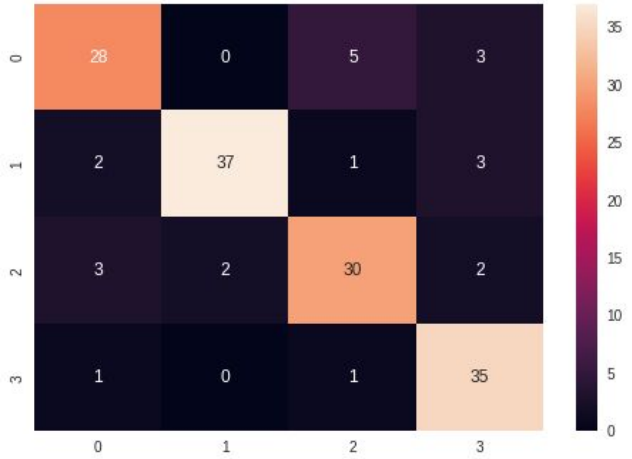
The positions of outliers using TSNE and Elbow Method

Results

We retrain the baseline to get the following results.

For clustering we get, SSE Score = 428998.3932157522

	precision	recall	f1-score	support
2	0.78	0.82	0.80	34
6	0.86	0.95	0.90	39
8	0.81	0.81	0.81	37
9	0.95	0.81	0.88	43
accuracy			0.85	153
macro avg	0.85	0.85	0.85	153
weighted avg	0.85	0.85	0.85	153



Analysis

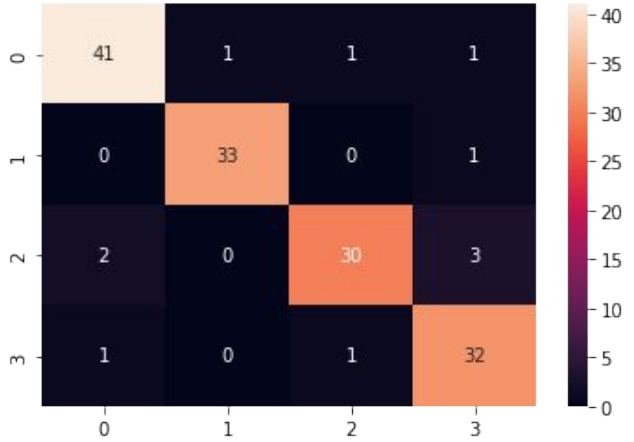
We can see that KMeans which is suitable for the dataset, is not only efficient but also good at identifying outliers in the data. From the TSNE plot we can see that many points that are at a distance from two cluster centers or lie between two clusters are mostly identified as outliers which is true for our dataset.
The same can be said using empirical measures where the F1 score and Precision improved as the model became more robust towards outliers.

Using Classification

We have used **One-Class Support Vector Machine** for outlier detection. We removed the outliers using the one-class support vector machine and then trained and tested the dataset using a vanilla decision tree classifier.

0.9251700680272109

	precision	recall	f1-score	support
2	0.93	0.93	0.93	44
6	0.97	0.97	0.97	34
8	0.86	0.94	0.90	32
9	0.94	0.86	0.90	37
accuracy			0.93	147
macro avg	0.93	0.93	0.92	147
weighted avg	0.93	0.93	0.93	147



Analysis

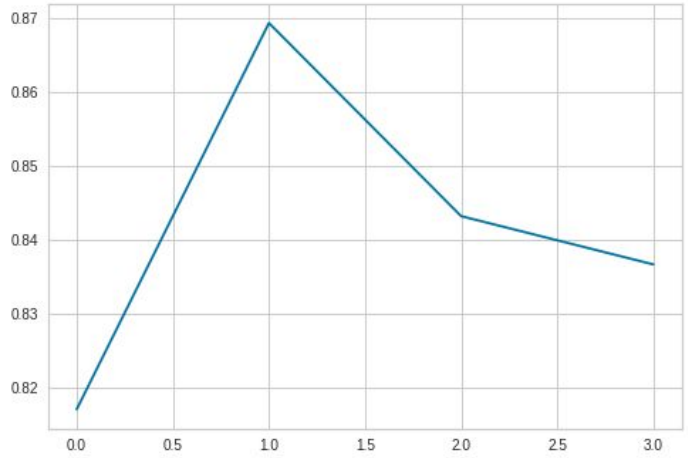
Analysis: We notice the accuracy score calculated using the Decision Tree classifier after using the one-class SVM to remove outliers is better as compared to using the Decision Tree without removing outliers. This shows that removing outliers reduces the chances of misclassification.
Before removing anomalies, the shape of data was: (765, 66), while shape after removing anomalies was: (735, 66).

Shortcomings: Vanilla decision tree trained on dataset containing anomalies gives poor test accuracy as it overfits the data. Thus, accuracy of the classifier improves after removing anomalies.

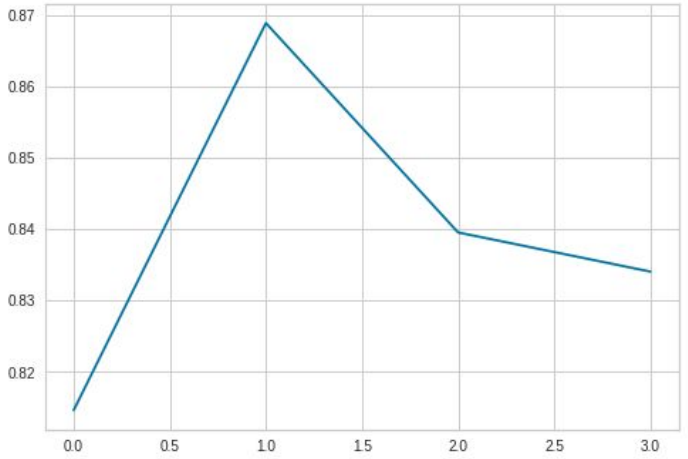
Infographics

Comparing across all the models for various metrics, we get

Accuracy



F1-Score (Macro)



Analysis

We notice a similar trend in the plots where all the models outperform the baseline model but we can notice that some models outperform others which depends both on the dataset chosen and the algorithm chosen in the category.
One thing to note is that clustering algorithm does not improve the performance by much, this can be attributed to the fact that very samples were chosen as outliers by the algorithms due to selected threshold.
In one-class SVM, 30 outliers are removed - which is high as compared to other algorithms.Thus resulting in best performance among selected algorithms.