**Abstract**

This report analyzes Dimensionality Reduction and Clustering algorithms over two datasets.

GitHub Readme and repo – <https://github.com/Maimoons/7641/blob/master/README.md>

**Introduction**

Two labelled datasets were used to analyze Kmeans and Expectation Maximization clustering algorithms.

Moreover, 4 dimensionality reduction algorithms, namely PCA, ICA, Randomized Projection and Random Forest were further run on the same datasets.

**Datasets and Pre-processing**

Preprocessing is the first step in machine learning algorithms. Raw and unclean data do not offer as well as inferences as clean, standardized data do. Preprocessing was done to remove inconsistencies and noise and fill up the missing values.

1. **Breast Cancer dataset**: This dataset was taken from sklearn. This was an interesting dataset because of the large number of features = 30

Table

Description automatically generated with medium confidence

*Statistics after preprocessing*

1. **Titanic dataset**:

The titanic dataset is a famous dataset taken from Kaggle’s list of datasets.

Table

Description automatically generated with medium confidence

*Statistics after preprocessing*

**Techniques for preprocessing–**

**Correlation analysis**: The correlation of the features was studied, and features with over 95% correlation were dropped to prevent overfitting. This was necessary since the amount of data was small and suspectable to overfitting and the breast cancer dataset had a lot of highly correlated features.

The columns dropped for breast cancer dataset were ['mean perimeter', 'mean area', 'perimeter error', 'area error', 'worst radius', 'worst perimeter', 'worst area']

**Feature Engineering:** The dataset was preprocessed to remove any null values. The breast cancer did not have any null values. The titanic dataset had two features with null values – Age and Fare. The Age had over 100 missing values whereas the Fare feature had a couple of missing values. They both had their mean filled out in the null rows to extrapolate from the existing data.

**Hot encoding**: Breast cancer had numerical values for all features and did not require non-numerical to numerical encoding. For titanic dataset, since there were not as many features, instead of dropping non numerical feature – sex, it was hot encoded into a 1/0 numerical value.

**Normalization:** For both the datasets, all features were normalized to model the data correctly and use a common scale across all features. This was also helpful in the case of neural networks which prevent the weights from over shooting

Kmeans - Breast Cancer

Performance

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated Chart, line chart

Description automatically generated

Speed

Chart, line chart

Description automatically generated Chart, scatter chart

Description automatically generated

Best Cluster: size 7

Chart, line chart

Description automatically generated Chart

Description automatically generated Chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, scatter chart, bubble chart

Description automatically generated Chart, scatter chart

Description automatically generated

Kmeans - Titanic

Performance

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

Speed

Chart, line chart

Description automatically generated Chart, scatter chart

Description automatically generated

Best Cluster: size 7

Chart, line chart

Description automatically generated Chart

Description automatically generated Chart, bar chart

Description automatically generated

Chart, scatter chart

Description automatically generated Chart, bubble chart

Description automatically generated Chart, scatter chart

Description automatically generated

Expectation Maximization

Breast Cancer

Performance

Chart, bar chart

Description automatically generated Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

Time

Chart, line chart

Description automatically generatedChart, scatter chart

Description automatically generated

Best Cluster 3

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated Chart

Description automatically generated

Expectation Maximization

Titanic

Performance

Chart, bar chart

Description automatically generated Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

Time

Chart, line chart

Description automatically generated Scatter chart

Description automatically generated

Best Cluster 13

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated Chart, bar chart, histogram

Description automatically generated

Dimensionality Reduction

|  |  |  |  |
| --- | --- | --- | --- |
| Breast Cancer dataset | Mean Reconstruction Error | Time to run | Number of components |
| PCA | 1.48999 e-24 | 0.0137708 | 30 |
| ICA | 0.987522 | 0.0234218 | 27 |
| Randomized | 3.91066 e-23 | 0.00243401 | 30 |
| Random Forest | NA | 2.45245 | 24 |

|  |  |  |  |
| --- | --- | --- | --- |
| Titanic dataset | Mean Reconstruction Error | Time to run | Number of components |
| PCA | 5.648538 e-22 | 0.00683975 | 7 |
| ICA | 4.58859 e-24 | 0.00523472 | 7 |
| Randomized | 43.9812 | 0.00215983 | 5 |
| Random Forest | NA | 0.659480 | 7 |

Breast Cancer

PCA ICA Randomized Projection Random Forest

Chart

Description automatically generated Chart

Description automatically generated Chart

Description automatically generated Chart, bar chart

Description automatically generated

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Titanic

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated Chart, bar chart, funnel chart

Description automatically generated

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated Graphical user interface

Description automatically generated

Clustering

|  |  |  |
| --- | --- | --- |
| K Means - Breast Cancer | Best K | Time to run |
| Original | 7 | 0.02999997 |
| PCA | 6 | 0.03223395 |
| ICA | 11 | 0.03854393 |
| Randomized | 6 | 0.08760285 |
| Random Forest | 6 | 0.03558493 |

|  |  |  |
| --- | --- | --- |
| EM - Breast Cancer | Best K | Time to run |
| Original | 3 | 0.1026802 |
| PCA | 3 | 0.0337829 |
| ICA | 10 | 0.1067998 |
| Randomized | 3 | 0.0657670 |
| Random Forest | 3 | 0.0583450 |

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Chart, waterfall chart

Description automatically generated Chart, bar chart

Description automatically generated Chart, waterfall chart

Description automatically generated Chart, bar chart

Description automatically generated

|  |  |  |
| --- | --- | --- |
| K Means - Titanic dataset | Best K | Time to run |
| Original | 7 | 0.03259516 |
| PCA | 8 | 0.030057907 |
| ICA | 6 | 0.039527893 |
| Randomized | 5 | 0.038379192 |
| Random Forest | 7 | 0.0428228378 |

|  |  |  |
| --- | --- | --- |
| EM – Titanic dataset | Best K | Time to run |
| Original | 13 | 0.0853328 |
| PCA | 15 | 0.83065390 |
| ICA | 15 | 0.45504999 |
| Randomized | 15 | 0.64911818 |
| Random Forest | 13 | 0.16853284 |

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Neural Network with Dimensionality Reduction

Original. PCA ICA Randomized Projection Random Forest

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated Chart, line chart

Description automatically generated Chart

Description automatically generated with medium confidence Chart

Description automatically generated

Performance

Chart

Description automatically generatedChart

Description automatically generatedChart

Description automatically generatedChart

Description automatically generatedChart

Description automatically generated

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Speed

Chart, bar chart

Description automatically generated

Neural Network with Clustering

Performance

Original. PCA ICA Randomized Projection Random Forest

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generatedChart

Description automatically generated

Chart

Description automatically generatedChart

Description automatically generated Chart

Description automatically generated Chart

Description automatically generatedChart

Description automatically generated

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Time

Chart, bar chart

Description automatically generated

References