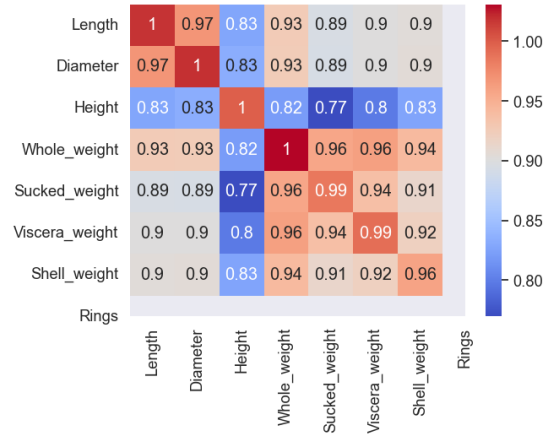
Q1. Representation Learning

For Abalone Dataset: Raw Dataset
Only dataset normalization and applying KNN. Accuracy achieved is 27.05% at K=68.

Next we applied, PCA as a pre-processing step.

Step 1: calculate the co-variance between the features and cumulative covariance produced by each principal components

```
Variance explained by each principal component:
 [0.9078731478516083, 0.039918908993422 65, 0.023906381975154992, 0.016295977883821613, 0.009236274060776192, 0.0018182993981407
179, 0.0009510098370754434]
Cumulative variance explained by each principal component:
 [0.90787315 0.94779206 0.97169844 0.98799442 0.99723069 0.99904899
 1.        ]
Number of principal components needed to explain 95% of the variance: 3
```



Step 2: Calculate the eigen values for each principal components

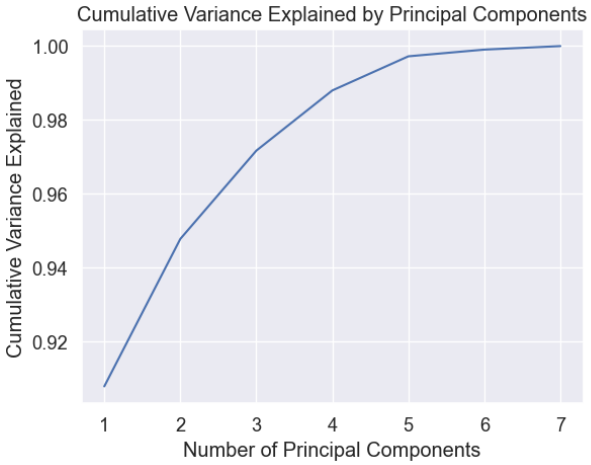
Eigenvalues of the principal components:
6.3566
0.2795
0.1674

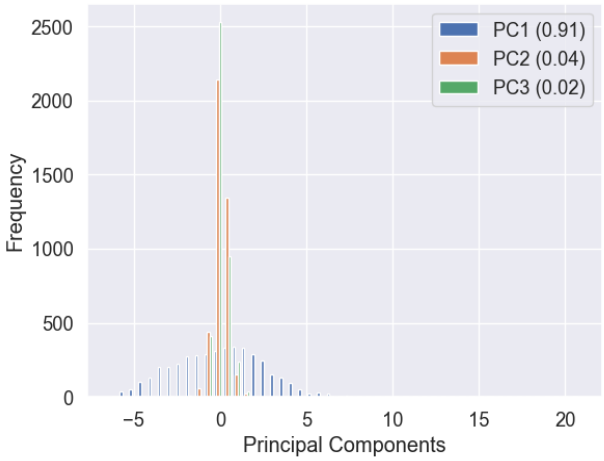
Step 3: print the Eigen vectors of the principal components

Eigenvectors of the principal components:

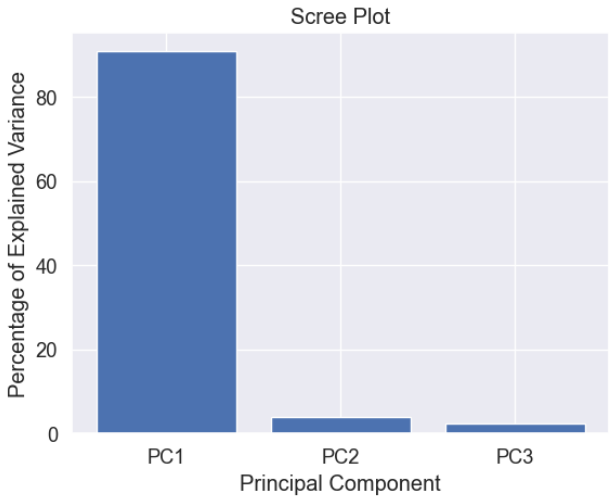
0.3833	0.3836	0.3481	0.3907	0.3782	0.3815	0.3789
0.0379	0.0653	0.8668	-0.2333	-0.3480	-0.2529	-0.0584
-0.5933	-0.5854	0.3149	0.2308	0.2316	0.2703	0.1621

Step 4: Plot the Cumulative variance explained by principal components and frequency of each PC

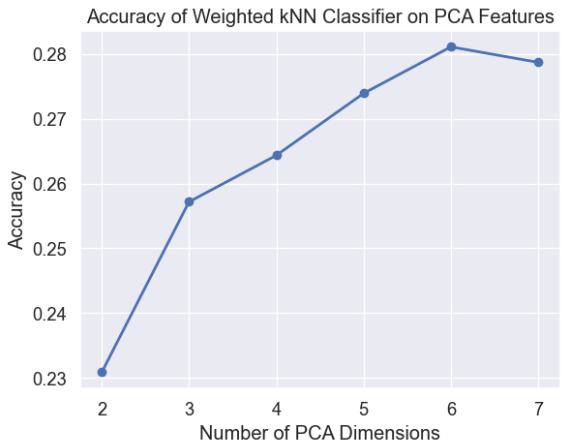




Step 5: Scree plot explaining the percentage of explained variance



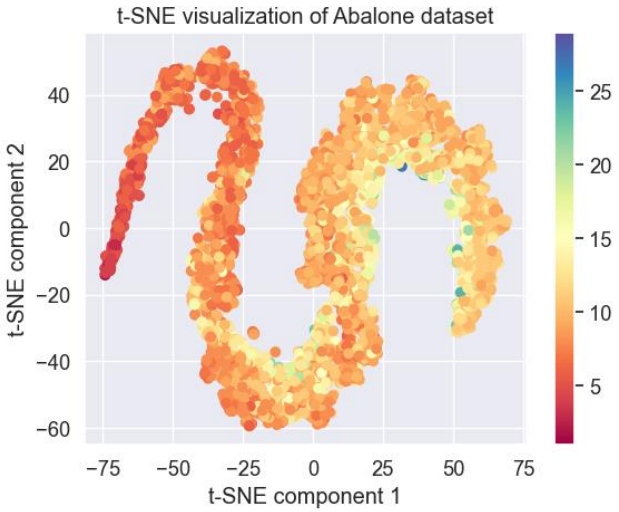
Step 6: use PCA as a pre-processing step and calculate the classification test using KNN K =68.



Step 7: using 3 principal components, we got test accuracy of 25.717%

Using LDA as a pre-processing step for abalone dataset we got the following accuracies:
Mean CV accuracy: 0.6692637147230204
Test Accuracy: 0.6961722488038278
CV Accuracy scores: [0.66367713 0.68263473 0.69011976 0.67215569 0.65568862]
CV Average accuracy: 0.6728551864880105

T-SNE Plot:



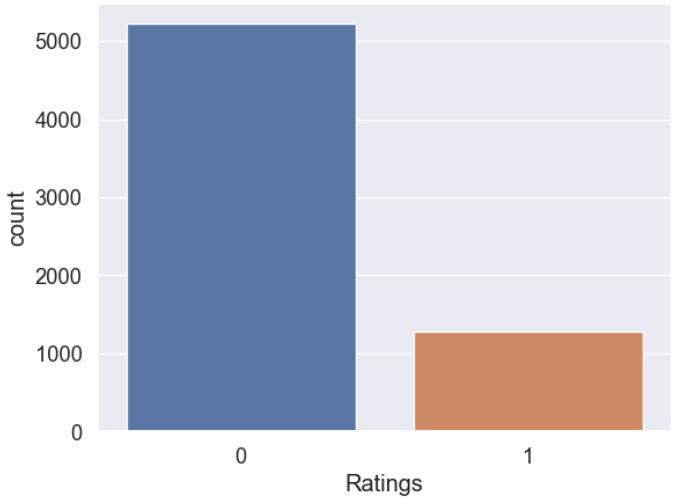
T-SNE Summary

1. The t-SNE plot shows the abalone samples projected onto a two-dimensional space based on their similarity in the original high-dimensional space.
2. Each point in the plot represents an abalone sample, and the color of the point corresponds to the number of rings in the abalone (an indicator of age).
3. The t-SNE plot reveals that the abalone samples with similar numbers of rings tend to cluster together, indicating that age is an important factor in the variability of the data.
4. The plot also shows that the length and diameter measurements of the abalone are strongly correlated, as points that are close together in the plot tend to have similar values for these variables.
5. There is some overlap between the clusters corresponding to different numbers of rings, indicating that other variables in the dataset also contribute to the variability of the data.
6. Overall, the t-SNE visualization provides an intuitive way to explore the structure of the abalone dataset and can reveal interesting patterns and relationships between the variables.

Using Wine Dataset:

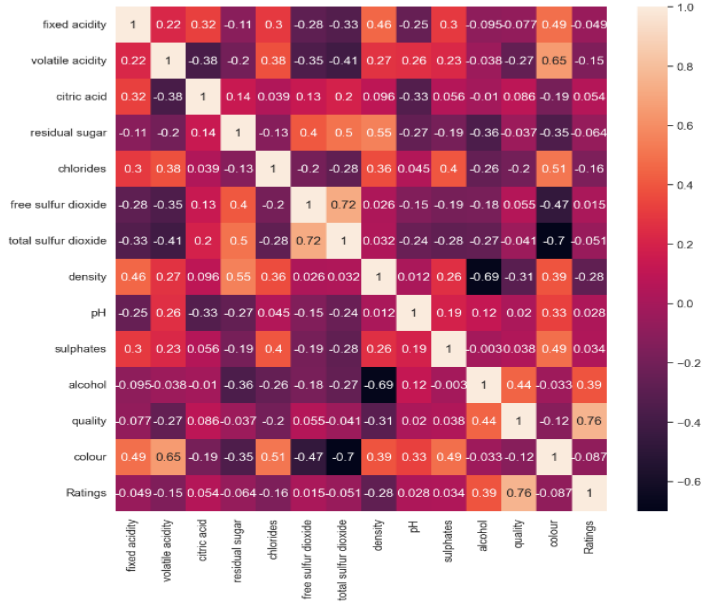
In Wine dataset, There are no missing values, so now we can start with EDA. More samples of quality 5 or 6 have been observed in the dataset, which shows that it is not a balanced dataset. The standard deviation for most features vary over a range and hence, we require normalization of the features before applying PCA.

Most of the wines in this dataset has a quality score of 5 or 6. We will now add a feature called 'rating' depending on the quality score of each wine data point. If quality is <5, we assign them as 'Bad' (value of 0) and if quality is >=5, we assign it as 'Good' (value of 1).



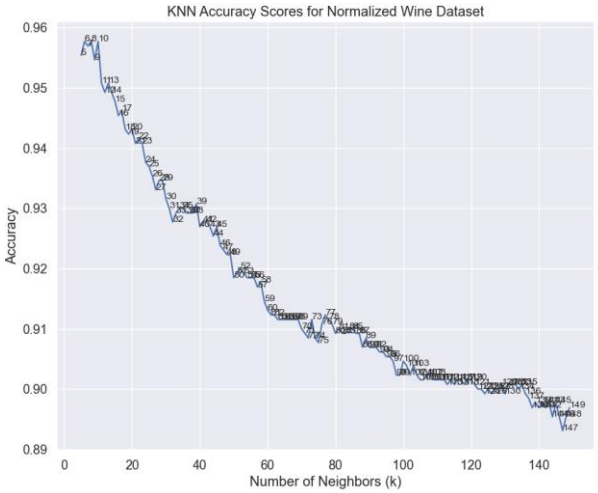
	mean	median	variance	skew	kurtosis
fixed acidity	7.215307	7.00000	1.680740	1.723290	5.061161
volatile acidity	0.339666	0.29000	0.027105	1.495097	2.825372
citric acid	0.318633	0.31000	0.021117	0.471731	2.397239
residual sugar	5.443235	3.00000	22.636696	1.435404	4.359272
chlorides	0.056034	0.04700	0.001227	5.399828	50.898051
free sulfur dioxide	30.525319	29.00000	315.041192	1.220066	7.906238
total sulfur dioxide	115.744574	118.00000	3194.720039	-0.001177	-0.371664
density	0.994697	0.99489	0.000009	0.503602	6.606067
pH	3.218501	3.21000	0.025853	0.386839	0.367657
sulphates	0.531268	0.51000	0.022143	1.797270	8.653699
alcohol	10.491801	10.30000	1.422561	0.565718	-0.531687
quality	5.818378	6.00000	0.762575	0.189623	0.232322
colour	0.246114	0.00000	0.185570	1.179095	-0.609922
Ratings	0.196552	0.00000	0.157944	1.527553	0.333522

Almost 5200 of the total number of wines seem to be "Bad" and the remaining 1297 wines "Good".



Alcohol has the maximum correlation with quality followed by sulphates and citric acid and then fixed acidity. We can also observe that residual sugar has a significant positive correlation with density and total sulfur dioxide is strongly correlated with the type of wine.

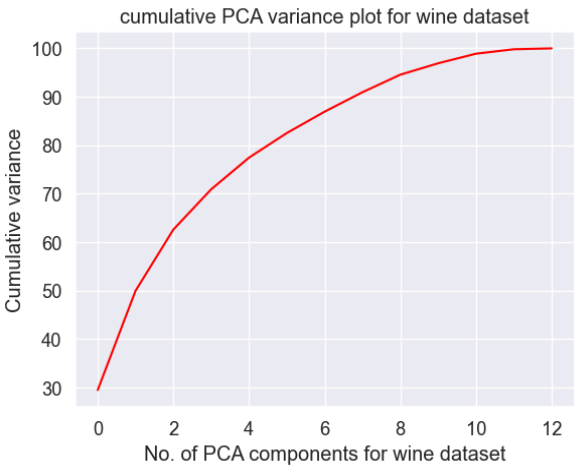
Calculating the accuracy on Wine dataset using KNN. we have selected value of K = 40 as the accuracy is decreasing when K is increasing. so, we chose the middle value i.e. K=40

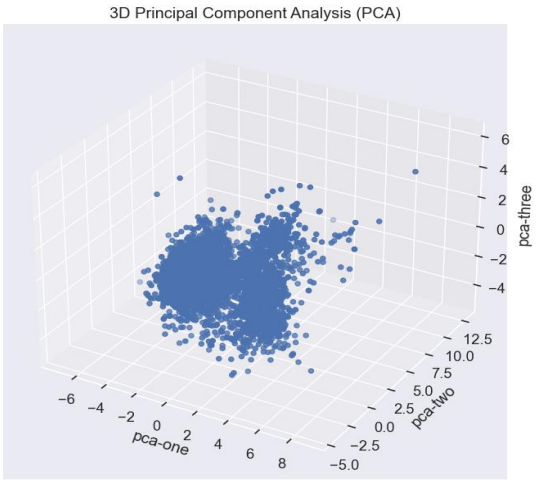


Test Accuracy: 0.9176923076923077
CV Accuracy scores: [0.66367713 0.68263473 0.690 11976 0.67215569 0.65568862]
CV Average accuracy: 0.6728551864880105

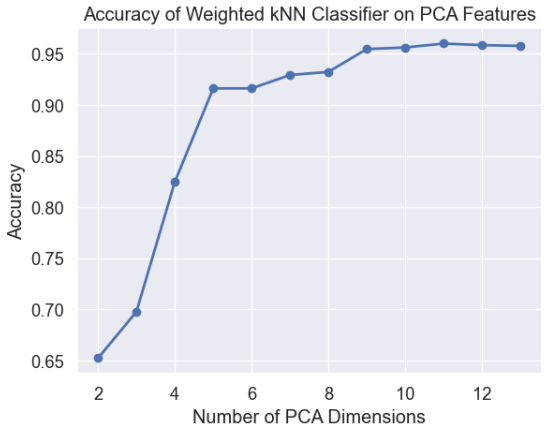
The above accuracy is for raw wine dataset with only normalization done as a pre-processing step

Applying PCA pre-processing on Wine dataset and plot the cumulative PCA Variance plot and 3D plot.





Plotting the Accuracy using weighted KNN classifier using PCA as a pre-processing step.



Significant Result:

After using PCA as a pre-processing step, we got the accuracy of 71.15%

After using LDA as a pre-processing step, we got the accuracy of 99.92 %

Conclusion for wine dataset after using PCA and LDA

The accuracy of KNN on the raw wine dataset is 98.85%, which is the highest accuracy among all settings. However, the accuracy drops significantly to 88.92% when using PCA as a preprocessing technique with three principal components. This drop in accuracy can be explained by the fact that PCA reduces the dimensionality of the dataset by projecting it onto a lower-dimensional space, which may cause some information loss.

On the other hand, when using LDA with three linear discriminants, the accuracy increases to 99.92%. This increase in accuracy can be explained by the fact that LDA seeks to find the linear combinations of features that maximize the separation between classes, leading to a better representation of the data for classification.

Conclusion for abalone dataset after using PCA and LDA

The results show that the accuracy of the KNN algorithm on the Abalone dataset is significantly improved after applying PCA and LDA. The accuracy of the KNN algorithm is 27.06% for the raw dataset, whereas it is increased to 62.68% and 69.62% after applying PCA and LDA, respectively.

The higher accuracy values obtained after applying PCA and LDA can be attributed to the fact that these techniques help to reduce the dimensionality of the dataset and select the most important features that contribute to the classification task. PCA identifies the directions of maximal variance in the data and projects the data onto a lower-dimensional subspace, while LDA finds the linear combinations of features that maximize the class separability. Both techniques help to reduce the noise in the data and make it easier for the KNN algorithm to find the correct class label for new data points.