Appendix: A2.1 Data Frame Configuration

Data Frame Configuration

Cell 1: Install and manage Python libraries

Cell 2: Ensure the latest python packages & Install and manage Python libraries

Cell 3: Mount the drive

Cell 4: Import various libraries and modules for data processing, ML and data visualisation

Cell 5: Convert list of integers into a numpy array and check array dimension

Cell 6: Extract data from the pandas dataframe and represent it as a numpy array

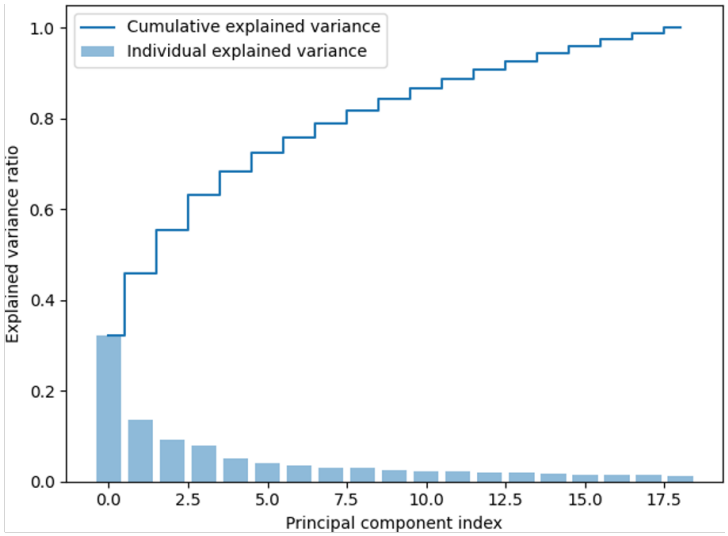
Cell 7: Standarise data features

Cell 8: Split dataset into training and testing subsets

Cell 9: Apply PCA to the training and test data

Cell 10: Visualise the PCA

This visualises the explained variance of each principal component and the cumulative explained variance across all components after applying PCA. In essence, this code provides a visual representation to understand the variance explained by the principal components.



Observing the plots can help one make decisions, such as determining the optimal number of components to retain for a certain level of explained variance.

The graph plotted is commonly used in Principal Component Analysis (PCA) to visualise how much of the total variance in the dataset is explained by each principal component. The cumulative variance is explained by adding more principal components. Let's break down the graph:

1. X-axis (Principal Component Index): This represents the index or order of the principal components. Principal components are sorted in decreasing order of explained variance, so the first component (index 0) explains the most variance, the second component (index 1) explains the second most, and so on.
2. Y-axis (Explained Variance Ratio): This axis represents the ratio of the variance explained by each individual principal component to the total variance in the original dataset. It quantifies how much information each principal component captures.

a. Individual Explained Variance (bar chart): Each vertical bar in the bar chart shows the proportion of variance explained by a single principal component. The first bar (at index 0) explains the most variance, and subsequent bars explain less and less variance.
b. Cumulative Explained Variance (step plot): The step plot shows the cumulative explained variance as you add more principal components. It starts at 0 and increases with each additional component. The cumulative explained variance gives you an idea of how much total variance in the dataset is retained as you consider more and more components.

3. Now, let's interpret what the graph is telling us:

a. Individual Explained Variance (bar chart): It shows that the first principal component (at index 0) explains the highest variance in the data, followed by the second, third, and so on. This is typical in PCA, where components are ordered by their ability to explain variance.
b. Cumulative Explained Variance (step plot): This curve shows that if you use only the first few principal components (up to around index 17.5 in your case), you can capture a substantial amount of the total variance in the original dataset. As you move along the x-axis to the right, adding more components, the cumulative explained variance increases. When it reaches 1.0 (100%), it means that all the variance in the dataset has been explained.
c. In summary, the graph illustrates how much of the dataset's variance can be retained by considering different numbers of principal components. It shows that with a relatively small number of components, you can capture a large portion of the dataset's variance, which is a key insight in dimensionality reduction techniques like PCA.

How to make decisions, such as determining an optimal number of components to retain for a certain level of explained variance.

- Determining the optimal number of principal components to retain in PCA is a crucial decision and can impact the performance of your machine learning model. There are several common methods for making this decision:

a. Explained Variance Threshold: One way to decide the number of components to retain is by setting a threshold for the cumulative explained variance. You can choose a threshold based on how much variance you want to retain. For example, if you want to retain 95% of the variance, you can look for the point on the cumulative explained variance curve where it crosses or exceeds 95%.
b. Elbow Method: The elbow method involves looking for an "elbow" point in the cumulative explained variance curve. This is where adding more components yields diminishing returns regarding explained variance. You can visually inspect the graph and choose the number of components where the curve begins to level off.
c. Cross-Validation: If you plan to use the PCA-transformed data for supervised learning, you can perform cross-validation with different numbers of retained components and choose the number that results in the best performance on your specific task (e.g., classification or regression
d. Domain Knowledge: Sometimes, domain knowledge can guide your decision. If you have prior knowledge about the dataset or the problem you're trying to solve, you may have insights into the importance of certain features or components, which can help you decide how many to retain.

Determining the optimal level of Principal components.

Based on the plot, where the cumulative explained variance increases step-wise from 0.3 to 1.0, and the individual explained variance decreases, most of the variance in our dataset is captured within the first few principal components. The optimal number of principal components to retain depends on specific requirements. However, based on the cumulative explained variance plot, We consider retaining enough components to capture a substantial portion of the total variance while keeping it close to 1.0 (100%). So we chose 0.99(99%) In this case, we achieved this by retaining approximately 19 principal components.

Output after Cell 4 is shown below.

	CACUL1	XRCC1	VPS4B	LOC79160	ZFAS1	CNTN2	TTTY11	ZC3HC1	PTGIS	KCNA2	...	LRRC37A6P	SCAP	STX10	ANAPC13	GRWD1	C
0	10.10	9.09	9.68	5.99	8.68	6.35	3.91	7.19	5.66	5.29	...	5.55	8.54	9.69	8.20	7.36	
1	10.45	9.15	9.51	6.18	8.46	6.60	3.92	7.05	5.54	5.36	...	5.52	8.81	10.42	7.23	7.45	
2	9.89	9.61	9.24	6.25	8.24	6.26	3.74	7.24	5.44	5.69	...	5.50	8.95	8.76	7.58	7.66	
3	10.36	9.32	9.32	6.25	8.44	6.32	4.08	7.14	5.41	5.28	...	5.64	8.91	9.32	7.63	7.32	
4	9.79	9.04	9.02	6.26	8.25	6.33	3.88	7.24	5.68	5.23	...	5.51	8.86	9.09	7.77	7.68	

5 rows × 20207 columns

Definition: Specificity-->True Positive Rate

Sensitivity-->True Negative Rate

Cell-9

This code applies PCA to reduce the dimensionality of both the training and testing datasets while retaining 99% of the data's variance. It then prints out the shapes of the transformed datasets to provide insight into the number of retained principal components

Cell-10

This code snippet visualises the explained variance of each principal component and the cumulative explained variance across all components after applying PCA. In essence, this code provides a visual way to understand the variance explained by the principal components. By observing the plots, one can make decisions, such as determining an optimal number of components to retain for a certain level of explained variance.