# Assignment Number 01

**Name: Mihir Unmesh Patil**

**Roll No: TYCOC213**
**Batch: C/C-3**

---

# Aim

The objective of this assignment is to analyze the Breast Cancer dataset using feature selection techniques and Principal Component Analysis (PCA). The study aims to identify the most important features, reduce dimensionality, and retain essential information to improve model performance and computational efficiency.

# Objectives

- Load the Breast Cancer dataset and explore its features.
- Apply feature selection techniques to identify the most significant features.
- Utilize **SelectKBest** and **SelectPercentile** for feature selection based on statistical significance.
- Standardize the dataset to ensure uniform feature scaling using **StandardScaler**.
- Apply **Principal Component Analysis (PCA)** to reduce dimensionality while preserving variance.
- Compare and analyze the selected features and principal components.
- Visualize the results to gain insights into feature importance and variance distribution.

# Theory

### Feature Selection

Feature selection is a technique used in machine learning to improve model accuracy and efficiency by eliminating irrelevant or redundant features. It helps prevent overfitting, reduces computation time, and enhances interpretability. In this study, we use two feature selection techniques:

1. **SelectKBest**
   - This method selects the top `k` features based on their statistical importance.
   - It uses the **ANOVA F-value** (`f_classif`) to measure the relevance of each feature to the target variable.

   F-value:
   $$F = \frac{\text{Between-group variability}}{\text{Within-group variability}} = \frac{\text{MSB}}{\text{MSW}}$$

   - Higher F-values indicate more important features.
2. **SelectPercentile**
   - Instead of selecting a fixed number of features like SelectKBest, this method retains a percentage of the most significant features.
   - The top `p%` of features are selected based on statistical scores.
   - Helps in dynamically selecting the most relevant subset of features.

## Standardization

Since different features in the dataset have varying ranges and scales, standardization is required before applying PCA. **StandardScaler** is used to transform the dataset so that each feature has a mean of 0 and a standard deviation of 1.

$$X_{\text{standardized}} = \frac{X - \mu}{\sigma}$$

## Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a widely used dimensionality reduction technique that transforms correlated features into a smaller set of uncorrelated variables called **principal components (PCs)**.

- PCA finds new feature representations by maximizing variance while minimizing information loss.
- Each principal component is a linear combination of the original features.
- The **explained variance ratio** determines how much information each principal component retains.
- PCA helps in visualizing high-dimensional data in a lower-dimensional space.

## Mathematical Representation of PCA

Given an input dataset **X** with n features, PCA involves the following steps:

1. Compute the covariance matrix of **X**.

$$\text{Cov}(X_i, X_j) = \frac{1}{n-1} \sum_{k=1}^{n} (X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)$$

2. Find the eigenvalues and eigenvectors of the covariance matrix.
3. Select the top k eigenvectors corresponding to the largest eigenvalues.
4. $$\text{Explained Variance Ratio} = \frac{\lambda_i}{\sum \lambda_j}$$
5. Project the original data onto the new k-dimensional space.

The resulting transformed dataset captures the most variance using fewer features, leading to better performance in machine learning models.

---

# Results

CODE: co Assignment_01_ML.ipynb

## Feature Selection Results

Using **SelectKBest**, we tested different values of k (30, 25, and 10). The results showed that a smaller subset of features still retained essential information:

- **K=30**: All features were selected.
- **K=25**: Most significant features were retained, with minor reductions.
- **K=10**: A small subset of highly relevant features was selected, focusing on key attributes.

Using **SelectPercentile**, we tested different selection percentages (25%, 15%, and 5%):

- **Top 25%**: Selected features retained most of the information.
- **Top 15%**: Fewer features, but essential patterns were still present.
- **Top 5%**: Only the most statistically significant features were retained.

Both methods highlighted the most critical features such as **mean radius, mean perimeter, mean concavity, worst radius, and worst concavity**, which are highly indicative of cancer classification.

### PCA Results

Applying PCA with **10 principal components**, we observed the following explained variance distribution:

| Principal Component | Explained Variance (%) |
|---|---|
| PC1 | 43.17% |
| PC2 | 19.85% |
| PC3 | 9.73% |
| PC4 | 6.53% |
| PC5 | 5.21% |
| PC6 | 4.19% |
| PC7 | 2.26% |
| PC8 | 1.68% |
| PC9 | 1.29% |
| PC10 | 1.20% |

The first **two components (PC1 and PC2)** together captured more than **60% of the variance**, indicating that a large portion of the dataset's information could be represented in just two dimensions. The cumulative variance plot confirmed that **most of the meaningful information was captured in the first few principal components**.

---

# Conclusion

This study demonstrated the effectiveness of **feature selection** and **dimensionality reduction** techniques in machine learning. The key findings include:

1. **Feature Selection**
   - **SelectKBest** and **SelectPercentile** successfully identified the most relevant features, reducing the dataset's complexity while preserving crucial information.
   - The top-selected features were strongly correlated with the target class, showing their importance in breast cancer classification.
2. **Principal Component Analysis (PCA)**

- ○ PCA significantly reduced the number of dimensions while retaining most of the dataset's variance.
  - ○ The first few principal components carried the majority of the dataset's variance, making it possible to work with fewer features while maintaining high predictive accuracy.
  - ○ PCA provided a clear visualization of the data distribution, aiding in better understanding and classification.
3. **Practical Implications**
  - ○ Feature selection improves machine learning models by reducing overfitting and improving computational efficiency.
  - ○ PCA helps in exploratory data analysis and visualization by representing high-dimensional data in a lower-dimensional space.
  - ○ These techniques are particularly useful in medical diagnostics, where large datasets with numerous features need to be analyzed efficiently.

# References:

**https://arxiv.org/pdf/1306.1326**
**https://jios.foi.hr/index.php/jios/article/download/1967/1021**