Principal Component Analysis

or

Dimensionality Reduction

or

Feature Reduction/Column Reduction

n_components = 6

Step by Step Breakdown of PCA

1. Scale the data using standardization technique.
2. Calculates the Variance.
3. Calculates the Co Variance
4. Reduces the data in a Covariance Matrix.
5. Calculates the Eigen Values of the reduced/Cov. Matrix

Step 1. Calculation of Z Scores

$$Z = \frac{X - \overline{X}}{\sigma}$$

X → Discrete values of X

$\overline{X}$ → Mean of X

σ → Standard Deviation → $\sqrt{\frac{\sum(x-\overline{x})^2}{N}}$

Step 2.

$$\text{Variance}(x) = (\sigma)^2 = \frac{\sum (x - \bar{x})^2}{N}$$

$N \rightarrow$ No. of Data Points

Step 3

Formula for Co variance

Joint variance measure of two different features

Population Covariance Formula

$$\mathrm{Cov}(x,y) = \frac{\sum (x-\bar{x}) \cdot (y-\bar{y})}{N}$$

Cov(y,x) & Cov(x,y) remain the same

Population Covariance means calculating co variance of the entire dataset.

Sample Covariance Formula

$$\boxed{\frac{\sum (x-\bar{x}) \cdot (y-\bar{y})}{\underbrace{N-1}_{\text{sample size}}}}$$

Covariance Matrix

Diff b/w Corelation and Covariance

Correlation ranges from -1 to +1

Covariance ranges from $-\infty$ to $+\infty$

Covariance Matrix =

$$\begin{bmatrix} Var(x) & Cov(x,y) \\ Cov(y,x) & Var(y) \end{bmatrix}$$

Step 5

$$Cov(x,y) + \lambda I = 0$$

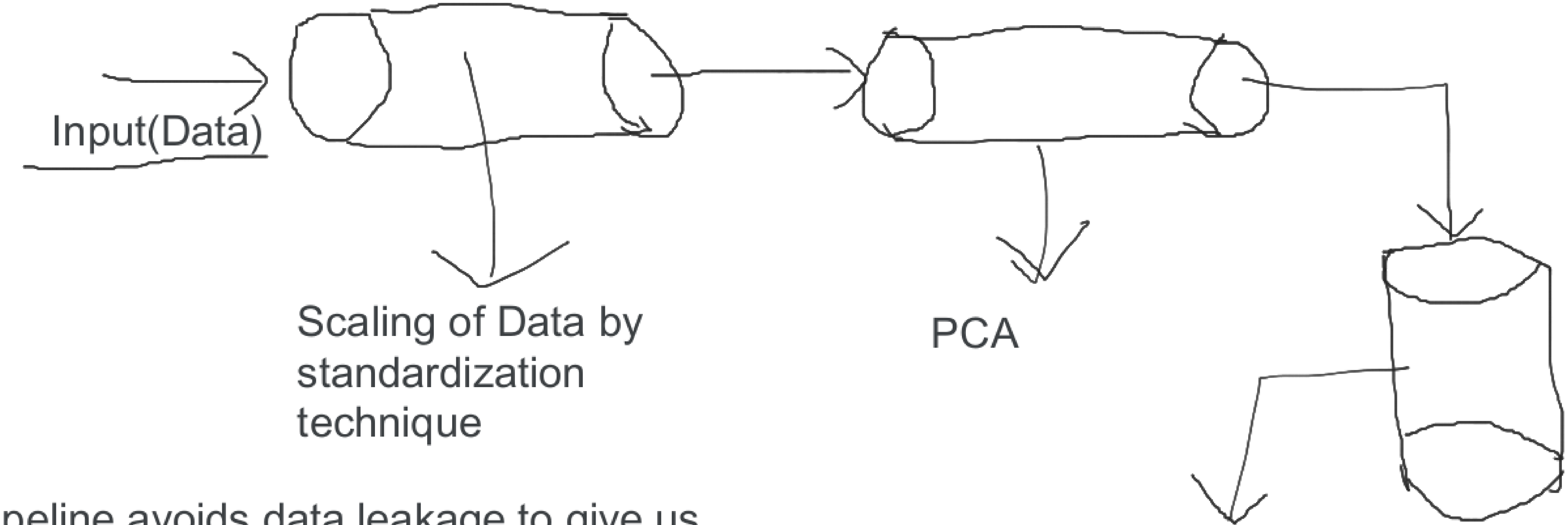
Eigen Vector

Reduced Data/
PCA-ed Data

```
a = [1,1,1]
b = np.diag(a)
```

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}_{3\times 3}$$

Pipeline



Pipeline avoids data leakage to give us right/accurate predictions.