Unsupervised Learning

- ***Missing value***

  Keadaan data memiliki nilai yang hilang (tidak diketahui nilainya).

- ***Standarization***

  Proses untuk menyeragamkan skala data yang berbeda, umumnya dikenal sebagai `scaling`.

- ***Covariance***

  Nilai yang menggambarkan hubungan (positif/negatif/tidak ada hubungan) antara dua variabel numerik. Namun, `covariance` tidak dapat menggambarkan seberapa erat/kuat hubungan tersebut karena nilai `covariance` tidak memilki batasan yang mutlak (- inf, + inf).

- ***Correlation***

  Nilai yang menggambarkan keeratan hubungan (positif/negatif/tidak ada hubungan) antara dua variabel numerik.

  - Nilai `correlation` mendekati 1 artinya kedua variabel berhubungan erat dan hubungannya positif
  - Nilai `correlation` mendekati -1 artinya kedua variabel berhubungan erat dan hubungannya negatif
  - Nilai `correlation` mendekati 0 artinya kedua variabel tidak saling berhubungan

- ***Principal Component***

  Dimensi/variabel baru yang berisi rangkuman informasi dari keseluruhan variabel awal (data awal).

- ***Principal Components Analysis***

  Proses interpretasi model principal component. Analisis yang dilakukan pada umumnya terkait hubungan antarvariabel, sebaran data berdasarkan principal component yang terbentuk, dan total informasi yang berhasil dipertahankan.

- ***Eigen values dan Eigen vector***

  Pasangan value dan vector yang merepresentasikan informasi (variansi) dari sebuah matriks. Pasangan nilai ini digunakan untuk membangun nilai principal component.

- ***Biplot***

  Plot yang menggambarkan posisi data berdasarkan nilai principal component dan pengaruh tiap variabel ke dua dimensi `principal component`.

- ***Outlier***

  Data yang nilainya sangat ekstrim dibandingkan dengan data lainnya, pada konteks tertentu dapat diidentifikasi sebagai data anomali.

- ***Reconstruction***

  Proses transformasi nilai principal component ke variabel awal.

- ***Clustering***

  Proses pengelompokkan data berdasarkan jarak antarobservasi (kemiripan).

- ***Centroid***

  Pusat cluster.

- ***Euclidean distance***

  Salah satu ukuran jarak, digunakan pada algoritma K-means Clustering.

- ***Between sum of square***

  Jarak tiap pusat cluster (`centroid`) ke pusat data secara keseluruhan.

- ***Within sum of square***

  Jarak tiap observasi ke `centroid` (pusat cluster) tiap cluster.

- ***Total sum of square***

  Jarak tiap observasi ke `centroid` global (titik tengah apabila hanya terdapat satu cluster).

- ***Elbow method***

  Salah satu metode yang dapat digunakan untuk menentukan jumlah cluster yang optimum dilihat dari titik potong nilai K yang tidak lagi secara signifikan meningkatkan homogenitas cluster.