Unsupervised Learning

- ***Missing value***

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

- ***Standarization***

  Proses untuk menyeragamkan skala data yang berbeda, sering disebut sebagai `scalling`.

- ***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 untuk membuat `principal comonent`.

- ***Eigen values***

  Nilai yang merepresentasikan jumlah/besar informasi (variansi) yang dimiliki oleh setiap PC.

- ***Eigen vector***

  Kumpulan nilai yang memproyeksikan data awal ke setiap `principal component`

- ***Biplot***

  Plot yang menggambarkan posisi data berdasarkan hasil `principal component analysis` dan besarnya pengaruh setiap variabel ke `principal component` 1 dan `principal component` 2.

- ***Outlier***

  Data yang nilainya sangat ekstrim, sering disebut sebagai data yang anomali.

- ***Reconstruct***

  Proses transformasi hasil `principal component analysis` ke data awal.

- ***Clustering***

  Proses mengelompokkan data berdasarkan jarak terdekat (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***

  Jumlah nilai `Between sum of square` dan nilai `Within sum of square`

- ***Elbow method***

  Salah satu metode yang digunakan untuk menentukan jumlah cluster yang optimum.