Classification 2

- ***Target variable***

  `Variable` yang ingin diprediksi/dimodelkan, sering disebut sebagai `respon/dependent variable`.

- ***Predictor***

  `Variable` yang digunakan untuk memprediksi `target variable`, sering disebut sebagi `independent variable`.

- ***Classification***

  Metode yang digunakan untuk memprediksi `target variable` bertipe kategorik (factor).

- ***Missing value***

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

- ***Deletion***

  Membuang `variable`/kolom pada data yang memiliki jumlah `missing value` melebihi 50% dari jumlah observasi

- ***Observation***

  Data yang dikumpulkan sebagai informasi, secara umum mengacu pada 1 baris data yang terdiri dari beberapa variabel.

- ***Full analysis***

  Membuang `observation`/baris yang mengandung `missing value`. Cara ini dilakukan jika jumlah `observation` yang mengandung `missing value` tidak melebihi 5% dari total `observation`.

- ***Imputation***

  Mengisi `missing value` dengan suatu nilai tertentu.

- ***Feature engineering***

  Tahapan untuk menambah jumlah `variable`/kolom berdasarkan informasi dari `variable` lain yag sudah ada.

- ***Feature selection***

  Feature selection merupakan tahapan dalam memilih variabel yang akan digunakan.

- ***Independent Events***

  Kejadian yang masing-masing tidak saling berkaitan satu sama lain.

- ***Dependent Events***

  Kejadian yang masing-masing saling berkaitan satu sama lain (kemungkinan terjadinya suatu kejadian akan memepengaruhi kemungkinan terjadinya kejadian lain.

- ***Laplace smoothing***

  Menambahkan nilai pada setiap prediktor, biasanya adalah 1 pada masing-masing jumlah kejadian.

- ***Text mining***

  Analisis yang dilakukan dengan memanfaatkan teks sebagai prediktor.

- ***Corpus***

  Objek yang digunakan pada `text mining`, berisi kumpulan teks yang disebut sebagai dokumen.

- ***Punctuation***

  Tanda baca pada teks.

- ***Stopwords***

  Kumpulan kata yang tidak memiliki makna jika diikutsertakan dalam pemodelan klasifikasi.

- ***Stemming***

  Proses untuk mengambil kata dasar dari sebuah kata berimbuhan.

- ***DocumentTermMatrix***

  Proses untuk membagi setiap konten teks menjadi kata-kata yang mewakili `predictor`.

- ***Confusion Matrix***

  Metrics yang digunakan untuk mengukur kebaikan model classification, terdiri dari `accuracy`, `recall`, `specificity`, dan `precision`.

- ***ROC (Receiver Operating Characteristic)***

  Kurva yang menggambarkan performa model klasifikasi untuk seluruh `threshold`.

- ***AUC***

  Luas area di bawah kurva `ROC`, menggambarkan keberhasilan model klasifikasi dalam memprediksi/membedakan kedua kelas dari `target variable`.

- ***Overfit***

  Keadaan dimana model yang dibuat hanya dapat memprediksi dengan baik `data train`. Namun, ketika melakukan prediksi pada `data test`, model tersebut tidak dapat memprediksi dengan baik.

- ***Entropy***

  Derajat kehomogenan.

- ***Information gain***

  Penurunan `entropy` setelah terjadi pembagian/spliting data.

- ***Pruning***

  Membatasi pembentukan cabang pada pohon (menyederhanakan pohon yang dibentuk).

- ***Terminal node***

  Bagian data yang sudah tidak dapat terbagi lagi.

- ***Data train***

  Bagian data yang digunakan untuk membuat model.

- ***Data test***

  Bagian data yang digunakan untuk mengevaluasi kebaikan model.

- ***Cross Validation***

  Proses untuk membagi data menjadi dua bagian, yaitu `data train` dan `data test`.

- ***K-fold cross validation***

  Membagi data sebanyak k bagian, setiap bagian akan digunakan menjadi train dan test secara bergantian.

- ***Class imbalance***

  Keadaan dimana suatu kategori/level lebih mendominasi keseluruhan `target variable` (kelas mayoritas) dibandingkan kategori/level lainnya (kelas minoritas).

- ***Sampling***

  Megambil sebanyak n bagian data secara acak.

- ***Down-sample***

  Proses `sampling` pada observasi kelas mayoritas, sebanyak jumlah observasi pada kelas minoritas. Tujuannya untuk menyamakan jumlah observasi pada kelas mayoritas dan minoritas.

- ***Up-sample***

  Proses `sampling` pada observasi kelas minoritas, sebanyak jumlah observasi pada kelas mayoritas. Tujuannya untuk menyamakan jumlah observasi pada kelas mayoritas dan minoritas.

- ***Ensamble method***

  Kumpulan beberapa algoritma prediktif untuk memperoleh performa yang lebih baik.

- ***OOB (Out of Bag) error rate***

  Besar error dari hasil prediksi pada data aktual yang belum dilihat oleh model.

- ***Variable importance***

  `Predictor` yang dianggap penting dalam model.