

DAFTAR ISTILAH

<i>Hyperparameter tuning</i>	: Proses memilih nilai terbaik untuk <i>hyperparameter</i> dalam model <i>machine learning</i> agar performa model menjadi optimal
<i>Overfitting</i>	: Model <i>machine learning</i> terlalu fokus pada data latih sehingga gagal dalam memahami data baru (model terlalu kompleks dan belajar terlalu detail)
<i>Underfitting</i>	: Model <i>machine learning</i> terlalu sederhana sehingga model gagal memahami data (tidak mampu menangkap pola penting dalam data)
<i>Train_test_split</i>	: Sebuah fungsi dari <i>scikit-learn</i> yang digunakan dalam proses membagi data
<i>Stratify</i>	: Parameter opsional di <i>train_test_split</i> yang menjamin proporsi kelas atau label di <i>training set</i> dan <i>test set</i> tetap sama seperti keseluruhan dataset (menjaga proporsi kelas)
<i>Random_state</i>	: Sebuah parameter <i>integer</i> untuk mengontrol <i>seed generator</i> pada proses pengacakan (mengunci proses acak agar hasil pembagian bisa replikasi)
<i>Max_df</i>	: Batas tertinggi seberapa sering suatu kata boleh muncul di semua teks. Kalau terlalu sering, dianggap tidak penting dan dibuang.
<i>Min_df</i>	: Batas terendah seberapa sedikit suatu kata boleh muncul. Kalau terlalu jarang, dianggap tidak relevan dan dibuang.
<i>N_estimator</i>	: Total jumlah pohon yang akan dibangun. Semakin banyak pohon, model bisa belajar banyak pola, tapi resiko <i>overfitting</i>
<i>Max_depth</i>	: Kedalaman maksimum pohon. Untuk mengontrol kompleksitas, dimana pohon yang lebih dalam bisa menangkap pola rumit, namun mudah <i>overfitting</i>
<i>Min_child_weight</i>	: Jumlah bobot minimal yang dibutuhkan agar sebuah node di <i>split</i> (apabila ada sebuah kumpulan data yang sampai dititik tertentu karena sudah memenuhi syarat pembagian sebelumnya)
<i>Learning_rate</i>	: Skala kontribusi setiap pohon baru terhadap prediksi akhir,
<i>Subsample</i>	: Proporsi baris (sampel) data yang digunakan untuk membangun tiap pohon
<i>Colsample_bylevel</i>	: Proporsi fitur (kolom) yang diambil secara acak di setiap level <i>split</i> pohon. Misal 0,5 berarti setiap

		level hanya melihat 50% fitur, cara lain untuk menambah keragaman pohon dan mencegah <i>overfitting</i> .
<i>Confusion matrix</i>	:	Tabel yang menampilkan jumlah prediksi benar dan salah model klasifikasi
<i>Receiver Operating Characteristic (ROC) Curve</i>	:	Grafik yang menunjukkan hubungan antara <i>True Positive Rate</i> (TPR) dan <i>False Positive Rate</i> (FPR) pada berbagai nilai ambang (<i>threshold</i>)
<i>Area Under the Curve (AUC)</i> .	:	Ukuran luas area di bawah <i>ROC Curve</i> . Nilai antara 0–1; semakin mendekati 1, semakin baik kemampuan model dalam membedakan kelas
<i>Scraping</i>	:	Teknik otomatisasi untuk mengambil (“mengorek”) data dari halaman web
<i>Stem_and_check_affix</i>	:	Proses <i>text preprocessing</i> untuk memotong kata (<i>stemming</i>) menjadi bentuk akar (<i>root</i>) dan memeriksa imbuhan (<i>affix</i>) seperti awalan atau akhiran.
<i>Grid search</i> ,	:	Metode pencarian otomatis untuk menemukan kombinasi hyperparameter terbaik. Mencoba semua kombinasi nilai dalam “grid” parameter, lalu memilih yang memberikan performa model tertinggi.