Overfitting dan Underfitting dalam Machine Learning

Overfitting dan Underfitting adalah dua masalah umum yang dapat terjadi saat melatih model machine learning. Kedua masalah ini dapat menyebabkan model berkinerja buruk pada data baru yang belum pernah dilihatnya.

# Overfitting

Overfitting terjadi ketika model terlalu kompleks atau terlalu banyak “belajar” dari data latih, sehingga tidak dapat melakukan generalisasi dengan baik pada data baru. Model yang overfitting memiliki kinerja yang sangat baik pada data lati tetapi buruk pada data uji.

## Ciri-ciri Overfitting:

* Akurasi tinggi pada data latih: Model hampir sempurna dalam memprediksi data latih.
* Akurasi rendah pada data uji: Model tidak dapat membuat prediksi yang akurat pada data baru.
* Kompleksitas model tinggi: Model menggunakan banyak parameter atau filter, atau memiliki struktur yang rumit.

## Cara Mengatasi Overfitting:

* Regularisasi: Menambahkan penalti pada model untuk mengurangi kompleksitasnya.
* Cross-Validation: Membagi data latih menjadi beberapa bagian untuk mengestimasi kinerja model secara lebih akurat.
* Pruning (untuk Decision Tree): Mengurangi ukuran pohon keputusan dengan memangkas cabang yang tidak penting.
* Early Stopping: Menghentikan pelatihan model sebelum overfitting terjadi.

# Underfitting

Underfitting terjadi ketika model terlalu sederhana dan tidak dapat menangkap pola kompleks dalam data. Model yang underfitting memiliki kinerja yang buruk pada data latih maupun data uji.

## Ciri-ciri Underfitting:

* Akurasi rendah pada data latih: Model tidak dapat belajar dengan baik dari data latih.
* Akurasi rendah pada data uji: Model juga tidak dapat membuat prediksi yang akurat pada data baru.
* Kompleksitas model rendah: Model menggunakan sedikit parameter atau fitur, atau memiliki struktur yang sangat sederhana.

## Cara Mengatasi Underfitting:

* Menambahkan Fitur: Menambahkan fitur baru yang mungkin relevan untuk model.
* Menggunakan Model yang Lebih Kompleks: Menggantikan model dengan model yang memiliki kapasitas pembelajaran yang lebih tinggi.
* Mengurangi Regularisasi: Jika Anda menggunakan regularisasi, coba kurangi kekuatannya.

# Visualisasi Overfitting dan Underfitting

Anda dapat menggunakan grafik learning curve untuk memvisualisasikan overfitting dan underfitting. Grafik ini menunjukan bagaimana kinerja model (akurasi atau loss) berubah seiring bertambahnya jumlah data latih.

## **Tugas Hari 11:**

1. Pemahaman Konsep: Jelaskan dengan kata-kata Anda sendiri apa itu overfitting dan underfitting. Berikan contoh masing-masing.
2. Identifikasi Masalah: Amati kembali model Logistic Regression yang Anda buat sebelumnya. Apakah Anda melihat tanda-tanda overfitting atau underfitting? Jelaskan alasan Anda.
3. Solusi (Opsional): Jika Anda mengidentifikasi masalah overfitting atau underfitting, coba terapkan salah satu teknik yang telah dijelaskan di atas untuk mengatasinya.

**Catatan:**

* Tugas nomor 3 bersifat opsional karena membutuhkan pemahaman yang lebih mendalam tentang machine learning. Jika Anda belum siap, tidak masalah untuk melewatinya.

**Selamat Mengerjakan Tugas! 😣**

**Tugas:**

1. Pemahaman Konsep:

**Overfitting:** Pada kasus ini mungkin terjadi banyaknya melakukan pembelajaran, sehingga model tersebut melakukan atau mungkin mencoba melakukan yang paling baik. Akan tetapi pada data uji model ini sangat lemah dikarenakan terlalu banyak melakukan pembelajaran sehingga terjadinya Overfitting.

**Underfitting:** Pada kasus ini mungkin karena pemodelan terlalu lebih sederhana bisa juga dimaksud modelnya terlalu simpel atau mungkin melakukan pembelajaran yang sedikit sehingga hasilnya kurang maksimal.

1. Identifikasi Masalah:

Tidak ada tanda-tanda Overfitting dan Underfitting. Alasan, karena kemarin membuat dengan ketentuan random\_state=42, jadi tidak ada yang menjadi masalah.   
Catatan: Tolong berikan penjelasannya

1. Solusi (Opsional):

**Penjelasan Lebih Lanjut:**

**Overfitting:** Overfitting terjadi ketika model terlalu kompleks dan "menghafal" data latih, bukannya belajar pola umum. Ini menyebabkan model bekerja sangat baik pada data latih tetapi buruk pada data baru.

**Underfitting:** Underfitting terjadi ketika model terlalu sederhana dan tidak dapat menangkap kompleksitas data. Model ini tidak akan bekerja baik pada data latih maupun data baru.

**Penjelasan tentang random\_state=42:**

Anda benar bahwa penggunaan random\_state=42 dalam train\_test\_split membantu memastikan konsistensi dalam pembagian data. Dengan mengatur nilai seed ini, Anda akan mendapatkan pembagian data yang sama setiap kali Anda menjalankan kode. Ini berguna untuk reprodusibilitas eksperimen, tetapi tidak secara langsung mencegah overfitting atau underfitting.

**Mengapa Model Anda Tidak Mengalami Overfitting atau Underfitting (Kemungkinan):**

**Dataset Sederhana:** Dataset Titanic relatif kecil dan sederhana, sehingga model Logistic Regression yang Anda gunakan mungkin tidak cukup kompleks untuk overfitting.

**Fitur yang Relevan:** Anda telah melakukan feature engineering dengan baik, sehingga fitur-fitur yang digunakan dalam model cukup relevan untuk memprediksi kelangsungan hidup.

**Regularisasi Default: Logistic Regression di scikit-learn memiliki regularisasi L2 secara default, yang membantu mencegah overfitting.**

**Penting untuk Dicatat:**

**Meskipun model Anda saat ini tidak menunjukkan tanda-tanda overfitting atau underfitting, penting untuk tetap waspada terhadap kedua masalah ini saat Anda bekerja dengan dataset dan model yang lebih kompleks.**