Salah satu hal yang paling penting untuk diperhatikan saat mengembangkan model ML adalah mengecek apakah model tersebut **underfit** atau **overfit** (arti dari masing-masing model akan kami jelaskan di bawah ini).

Sebuah model ML yang layak untuk di-*deploy* di tahap produksi adalah model yang **tidak**underfit atau overfit. Sebuah model yang tidak overfit dan underfit juga disebut dengan model **good fit**. Model yang bukan good fit cenderung akan memiliki performa sangat buruk di tahap produksi.

Bagaimana untuk melihat apakah sebuah model overfit atau underfit?

Salah satu caranya telah kita pelajari sebelumnya, yakni dengan membagi dataset menjadi **train set** dan **test set**. Setelah data tersebut dibagi, kita akan melakukan pengembangan model hanya dengan data training. Hasil pengujian model terhadap data testing dapat memberitahu kita apakah model kita underfit atau overfit.

**Overfitting**

Bayangkan suatu saat Anda berkunjung ke suatu kota, kemudian Anda mengalami kejadian tidak menyenangkan seperti dicopet di dalam angkutan umum. Anda kemudian mungkin akan berpikir bahwa semua angkutan umum di kota tersebut tidak aman. Kadang-kadang, kita melakukan generalisasi yang berlebihan terhadap sesuatu.  
  
Demikian halnya pada mesin yang juga bisa terjebak pada persepsi generalisasi yang sama. Dalam machine learning, kondisi ini disebut sebagai overfitting.

Overfitting terjadi ketika model memiliki prediksi yang terlalu baik pada data training, namun prediksinya buruk pada data testing. Ketika sebuah model overfit, model tidak dapat melakukan generalisasi dengan baik sehingga akan membuat banyak kesalahan dalam memprediksi data-data baru yang ditemui pada tahap produksi.

Contoh kasus adalah sebuah model machine learning untuk mengenali gambar anjing. Sebuah model yang overfit akan sangat menyesuaikan dengan dataset. Nah, di dataset mayoritas dari gambar anjing adalah anjing berwarna hitam. Maka model akan berpikir bahwa setiap hewan yang berwarna hitam adalah anjing. Ketika model tersebut dipakai untuk memprediksi sebuah gambar kucing dan kuda berwarna hitam, maka prediksinya adalah anjing.

Sebelum men-*deploy*model ML ke tahap produksi, ada teknik sederhana untuk mengecek apakah model overfit atau tidak. Pada model klasifikasi jika akurasi model pada data training tinggi dan data testing rendah, maka model yang Anda kembangkan overfitting. Pada model jenis regresi, jika model membuat kesalahan yang tinggi pada data testing maka model tersebut overfitting.

Beberapa cara untuk menghindari overfitting yaitu:

* Memilih model yang lebih sederhana, contohnya pada data yang memiliki pola linier menggunakan model regresi linear daripada model decision tree.
* Mengurangi dimensi data contohnya dengan metode PCA yang telah Anda pelajari.
* Menambahkan data untuk pelatihan model jika memungkinkan.

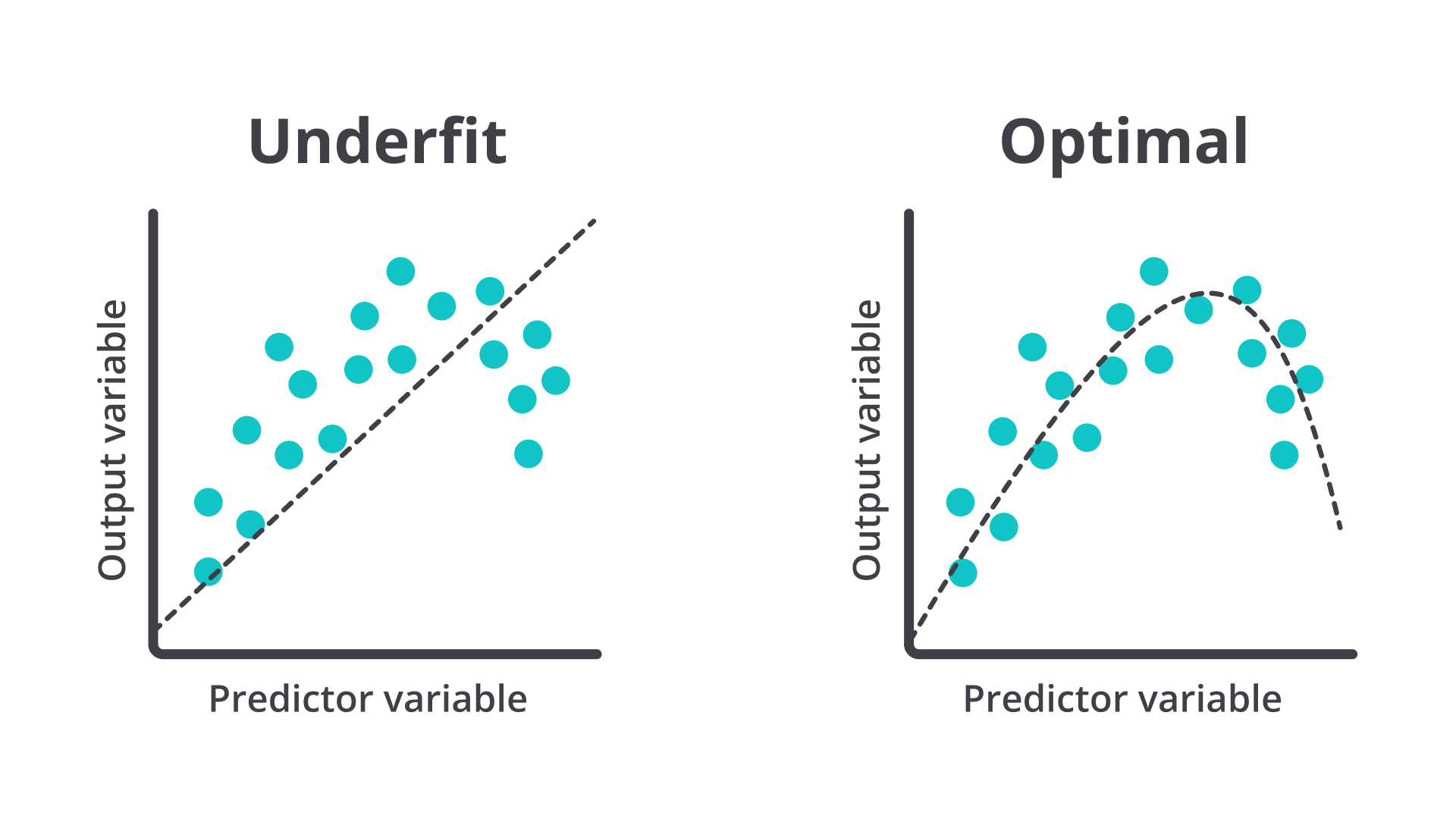
**Underfitting**

Underfit terjadi ketika model terlalu sederhana dan tidak mampu untuk menyesuaikan pola yang terdapat pada data latih. Bagaimana untuk mengenali apakah sebuah model underfit atau tidak?

Sebuah model dapat dikatakan underfit jika memiliki eror yang tinggi pada data training. Underfitting menandakan bahwa model tersebut belum cukup baik dalam mengenali pola yang terdapat pada data latih. Misalnya ketika sebuah model dilatih pada data latih yang memiliki 50 sampel coklat dan 50 sampel kacang. Setelah pembelajaran dengan data latih, model malah mengenali pada data latih terdapat 90 sampel coklat dan 10 sampel kacang.

Pada kasus klasifikasi, underfitting ditandai ketika model memiliki akurasi yang rendah pada data training. Pada kasus regresi, underfitting terjadi ketika model memiliki tingkat eror yang tinggi.

Pada ilustrasi regresi di bawah, model di sebelah kiri belum menyesuaikan dengan baik terhadap pola yang terdapat pada data. Bandingkan dengan model di sebelah kanan.

[](https://www.dicoding.com/academies/184/tutorials/8477?from=8537)

Cara menghindari underfitting adalah dengan menyeleksi model atau meningkatkan performa dengan tuning hyperparameter yang akan dibahas di submodul selanjutnya. Kualitas data juga sangat mempengaruhi dataset. Model machine learning yang sangat kompleks sekalipun tidak akan memiliki performa yang baik jika data yang digunakan memiliki kualitas yang buruk. Ingat prinsip: “Garbage in, garbage out” ya.

**Good Fit**

Model ML yang baik adalah model good fit atau model yang tidak underfit maupun overfit. Sebuah model good fit akan memprediksi lebih baik dan membuat lebih sedikit kesalahan di tahap produksi. Contoh dari model yang tidak good fit seperti di bawah.

[](https://www.dicoding.com/academies/184/tutorials/8477?from=8537)

Sebuah model membuat banyak kesalahan dalam memprediksi gambar di atas. Seperti sebuah kaos kaki dan anjing yang diprediksi sebagai gajah india, dan seekor paus yang diprediksi sebagai *african grey*. Tentunya Anda tidak ingin agar model ML membuat banyak kesalahan seperti di atas bukan. Berikut sebuah tabel yang membandingkan model yang underfitting, good fit, dan overfitting pada masalah regresi dan klasifikasi.

