**Nama:** Limas Jaya Akeh **| Kelas:** LB-01 | **NIM:** 2201763240 | Machine Learning

**Session 2**

**Introduction to Machine Learning – 2**

**Review Question**

1. Can you name four of the main challenges in Machine Learning?

2. What is a test set and why would you want to use it?

3. What is the purpose of a validation set?

4. What is cross-validation and why would you prefer it to a validation set?

**Answers**

**1. Can you name four of the main challenges in Machine Learning?**

Tantangan yang ada dalam Machine Learning adalah:

* Data yang Buruk

Data yang buruk seperti nonrepresentative (tidak mewakilkan model atau dataset) dan adanya NULL dalam salah satu atribut dapat mengakibatkan kualitas data menjadi kurang baik, dan mempengaruhi training yang dilakukan terhadap mesin. Noise (karena pengukuran atau pengambilan fitur yang salah) mengakibatkan mesin sulit mengidentifikasi pattern yang ada.

* Kurangnya Data

Kuantitas data yang diperlukan dalam Machine Learning sangat banyak, bahkan untuk masalah yang simple memerlukan contoh data hingga ribuan, masalah-masalah kompleks seperti image recognition bisa menggunakan hingga milyaran data. Mencari data memerlukan resource seperti uang, tenaga, dan waktu yang tidak sedikit, banyak perusahaan mencari cara agar mendapatkan data yang relevan terhadap mesin yang akan diberikan dataset.

* Pengambilan Fitur yang Salah

Pengambilan fitur-fitur yang ada dalam dataset dapat berpengaruh langsung terhadap kualitas data, sehingga diperlukan perhatian khusus agar feature extraction yang dilakukan dengan baik dan benar.

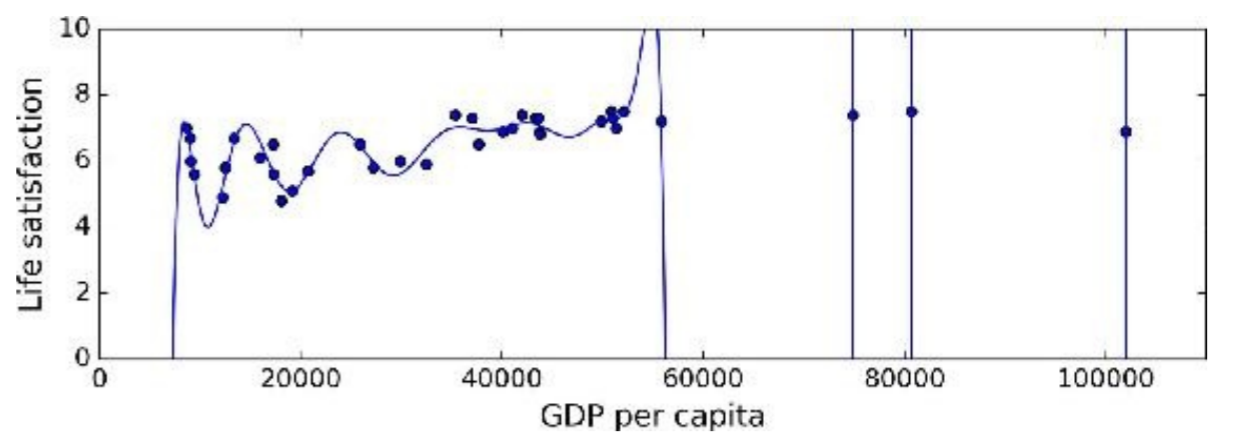
* Algoritma yang Buruk (Overfitting / Underfitting)

Overfitting terjadi saat model hanya menghasilkan result yang baik terhadap training data, tetapi akan salah ketika dicoba dengan test data.

A close up of a logo

Description automatically generatedHal ini terjadi karena algoritma yang dibuat terlalu terpengaruh oleh training set sehingga model terlalu ribet dan tidak dapat menangkap fitur secara umum.

Dalam contoh di samping, mesin mencoba membagi dua kelompok antara titik biru dan merah, model yang baik dapat meng-generalisasi (Digambar dengan garis hitam), tetapi karena mesin belajar terlalu spesifik sehingga ia mencoba mendapatkan hasil terbaik (Digambar dengan garis hijau) sehingga menjadi tidak solusi general.



A kite flying in the sky

Description automatically generated Underfitting terjadi ketika model yang dibuat terlalu simple sehingga model tidak dapat mengenali model baru secara baik, biasanya terjadi karena terdapat fitur atau parameter yang kurang yang harusnya ada dalam model yang baik.

Dalam contoh disamping, dataset yang ada dapat digambarkan dalam bentuk parabola (garis hitam) karena data terdiri dari dynamic data, akan tetapi model yang dibuat berupa linear (garis biru), hal ini akan mengakibatkan hasil prediksi yang buruk.

**2. What is a test set and why would you want to use it?**

Test set adalah sample data yang didapat dari dataset utama yang biasanya proporsinya lebih sedikit dibandingkan train set, yang biasanya digunakan untuk meng-evaluasi final model dari hasil latihan training set.

Test set biasa digunakan apabila model sudah dilatih secara baik, dan menilai apakah model sudah dapat memprediksi dengan baik atau tidak, test set dinilai baik apabila sangat bervariasi dan mencangkup seluruh class yang ada dalam suatu dataset.

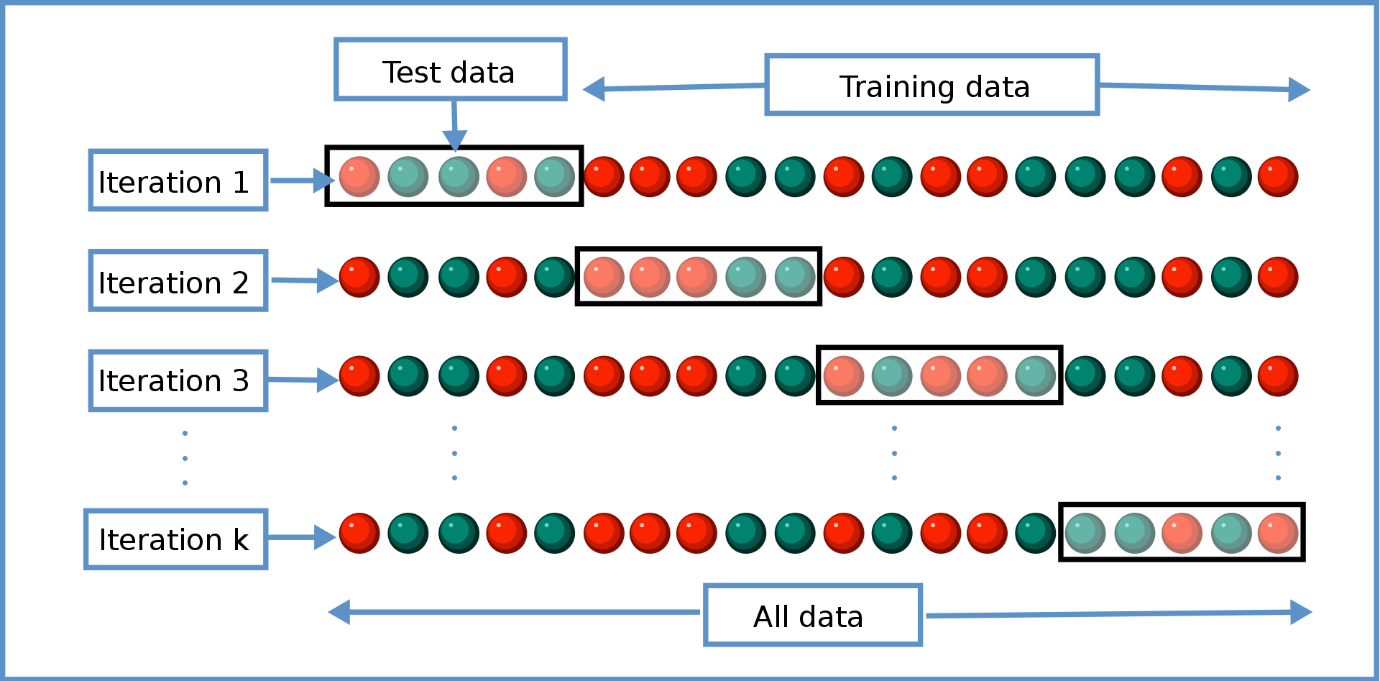
Test set biasa digunakan apabila Model sudah dilatih dengan Training set, sudah divalidasi dengan Validation set, dan biasanya rasio berupa 70:20:10 (70% Training set, 20% Validation set, 10% Test set); rasio ini dapat berubah sesuai kebutuhan, tetapi diusahakan test set merupakan bagian dengan jumlah paling minimum (tetapi tidak mendekati 0%) karena terlalu banyak testing dapat berdampak pada performa model yang kurang baik (karena kurang training dan validation).

**3. What is the purpose of a validation set?**

Validation set adalah sample yang didapat dari dataset utama yang digunakan untuk membuktikan bahwa model yang dibuat tidak overfitting atau underfitting. Akurasi yang didapat dapat membuktikan bahwa model akurat atau tidak dari dataset yang ada, dan apabila diperlukan perubahan atau modifikasi, evaluasi dari validation set dapat menjadi patokan untuk menghentikan training dan melakukan analisis ulang terhadap hyperparameter dan atribut yang ada.

Dalam Validation set, model tidak dimodifikasi secara weight tetapi secara arsitektur dari klasifikasi yang ada, dan mencari generalization error yang ada dalam model yang sudah dibuat.

**4. What is cross-validation and why would you prefer it to a validation set?**



Cross-validation adalah Teknik untuk memecah validation set menjadi subset-subset kecil yang kemudian akan dites oleh model dalam beberapa iterasi, dan dimana tiap iterasi akan dicoba data yang berbeda hingga semua data sudah selesai di test.

Dalam Cross-validation, tiap model dalam iterasi menggunakan sebagian besar validation set sebagai training data, dan menggunakan test data yang sudah ditandai sebagai test data, dan mencari tipe model dan hyperparameter sudah dipilih, final model kemudian dilatih menggunakan hyperparameter yang ada dalam validation set dan generalized error dihitung melalui test set.

Hal ini digunakan untuk meminimalisir penggunaan dataset yang ada, dan mencegah terjadinya overfitting atau selection bias, dan dapat mengambil kesimpulan apabila model dapat bekerja dalam berbagai macam scenario di dunia nyata. Salah satu metode Cross-Validation yang terkenal adalah K-Fold Cross-Validation.

Link Github: <https://github.com/limasakeh/ML-Pertemuan2>