

# Learning Progres Review Week #11

*By Optimistic team*



## Session 31

# Advanced Data Preprocessing for Machine Learning





## Feature Engineering

---

Merupakan Proses rekayasa data menggunakan teknik tertentu yang disesuaikan dengan kondisi dan permasalahan yang ada sehingga dapat menambah informasi yang tadinya tidak tersedia menjadi tersedia tanpa adanya data tambahan selain data yang digunakan. Tujuan Adanya feature engineering ialah memberi kemungkinan pada model machine learning yang dikembangkan untuk dapat belajar lebih banyak sehingga akurasi yang dimilikinya dapat meningkat.





## Handling Text Data

---

Proses mengeksplorasi dan menganalisis sejumlah besar data berbentuk teks yang tidak terstruktur dengan bantuan perangkat lunak yang dapat mengidentifikasi konsep, pola, topik, kata kunci dan atribut lainnya pada data teks.






## Data Cleaning (Pre-Processing Text)

---

Data Cleaning bertujuan untuk membersihkan data teks yang kotor agar dapat lebih mudah diproses pada tahapan selanjutnya.

Contoh:

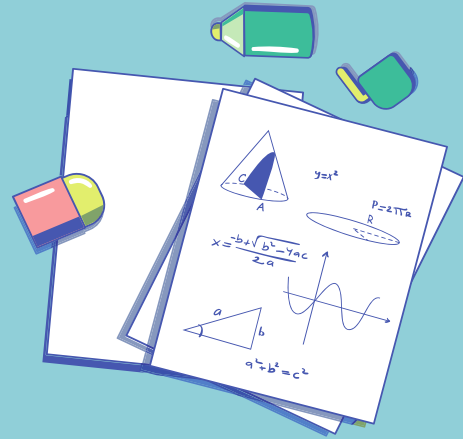
- Mengubah ke huruf kecil
  - Menghapus simbol
  - Menghapus atau menerjemahkan angka ke dalam teks
  - Menghapus mentions, hashtags, dan url
  - Menghapus whitespace
- 

- ***Tokenization***

Tokenization merupakan proses memisahkan atau membagi kalimat yang terdapat pada data teks menjadi per kata.

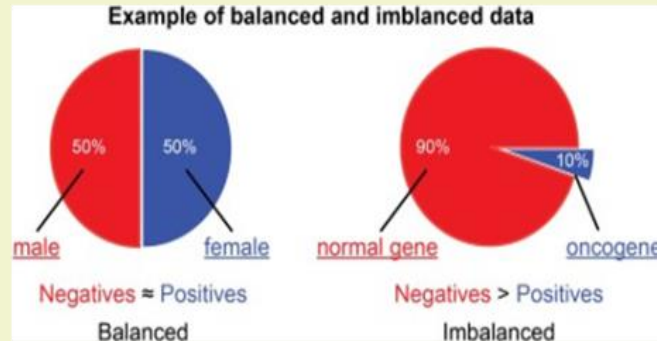
- ***Removing Stop-word***

Removing stop-word berfungsi untuk menghapus kata-kata umum yang muncul dalam jumlah besar namun dianggap tidak memiliki makna dengan tujuan menghilangkan kata yang memiliki informasi rendah pada sebuah teks sehingga dapat fokus pada kata yang mempunyai informasi penting.



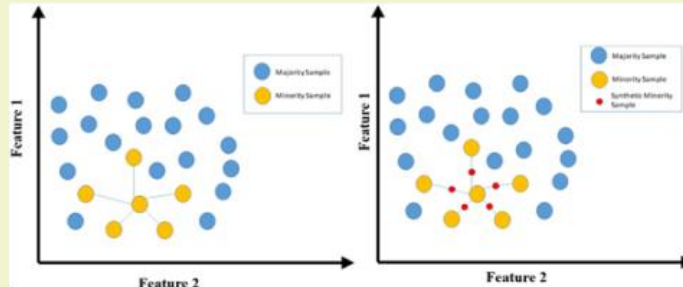
## Imbalance Dataset

Imbalance data merupakan keadaan dimana variabel target yang akan diprediksi atau diklasifikasikan jumlahnya tidak seimbang. Ketidak seimbangan pada *dataset* ini dapat mempengaruhi proses belajar *machine learning* sehingga nantinya model yang dikembangkan akan cenderung mengeluarkan hasil sesuai dengan data yang paling banyak dia pelajari dan mengakibatkan keakuratan yang kurang baik dari model tersebut.



## SMOTE (Synthetic Minority Oversampling Technique)

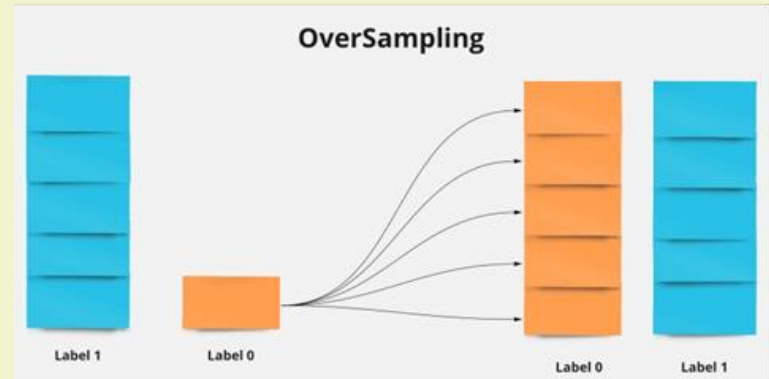
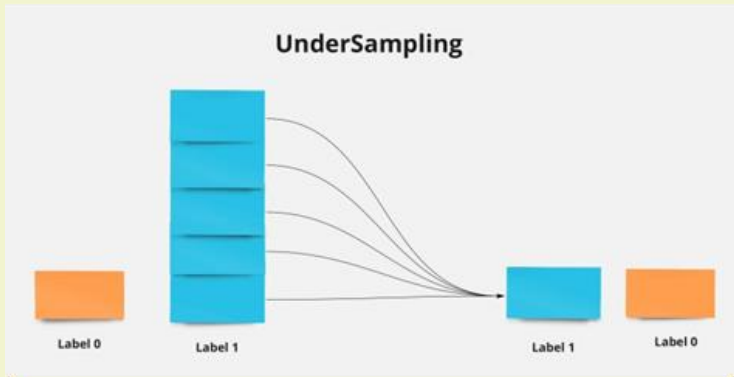
Teknik sampling dengan membuat data sintetis dari data minoritas sehingga jumlahnya akan sama dengan data mayoritas. Data sintetis atau buatan ini akan berbeda dengan data aslinya namun secara ilmiah masih memiliki karakteristik yang hampir serupa dengan data asli sehingga keabsahan data akan sedikit lebih terjaga dibanding jika menggunakan teknik oversampling biasa.





## Handling Imbalance Dataset

Undersampling merupakan teknik sampling secara acak dengan tujuan mengurangi jumlah data yang lebih banyak sehingga perbandingan datanya sama. Sedangkan, Oversampling adalah teknik sampling secara acak dengan tujuan menduplikasi data yang lebih sedikit jumlahnya hingga perbandingannya sama dengan data yang lebih banyak.



# Session 32

# Classification I



# *Klasifikasi*

Klasifikasi adalah model dari supervised learning yang digunakan untuk menampilkan prediksi berdasarkan dari class atau labelnya.



# Aplikasi dari Klasifikasi

Model *supervised learning* klasifikasi biasanya di aplikasikan pada kasus-kasus seperti berikut:



Fraud Detection



Cancer Cell Classification



Credit Scoring

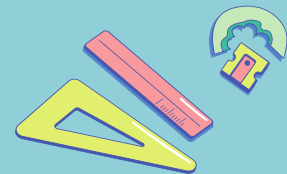


Spam Detector



Churn Analysis

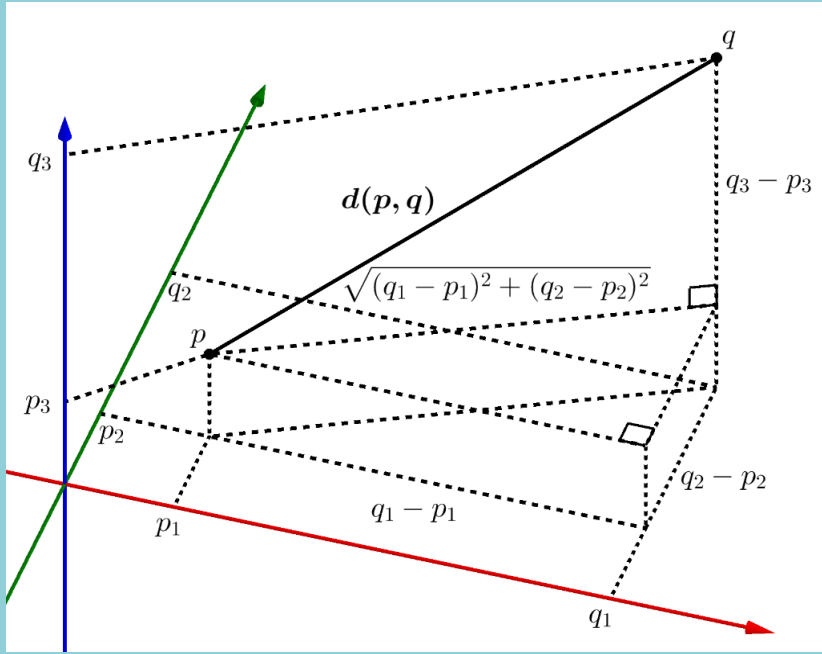
1. **Fraud Detection**, memprediksi/ menyelidiki/ memberikan ciri-ciri yang termasuk kategori Fraud (penipuan atau kecurangan). Fraud Detection & Prevention membantu perusahaan dalam menghentikan berbagai jenis penipuan internet.
2. **Cancer Cell Classification**, memprediksi pasien apakah terdeteksi terkena kanker atau tidak berdasarkan data dengan kelas positif kanker dan negative kanker.
3. **Credit Scoring**, alat dan teknik prediksi yang membantu lembaga keuangan untuk meminjamkan. Tujuan penilaian kredit adalah untuk menetapkan calon pelanggan atau pelanggan ke satu kelompok "pelanggan baik" atau "pelanggan buruk".
4. **Spam Detector**, memprediksi apakah email yang masuk adalah spam atau bukan berdasarkan data yang di inputkan ke dalam model klasifikasi.
5. **Churn Analysis**, Memprediksi pelanggan apakah akan berpindah ke perusahaan lain atau tidak.



## *K-Nearest Neighbor*

K Nearest Neighbour atau KNN adalah teknik mengklasifikasikan data berdasarkan seberapa banyak data terdekat. Menggunakan distance algorithm untuk mengklasifikasikan data

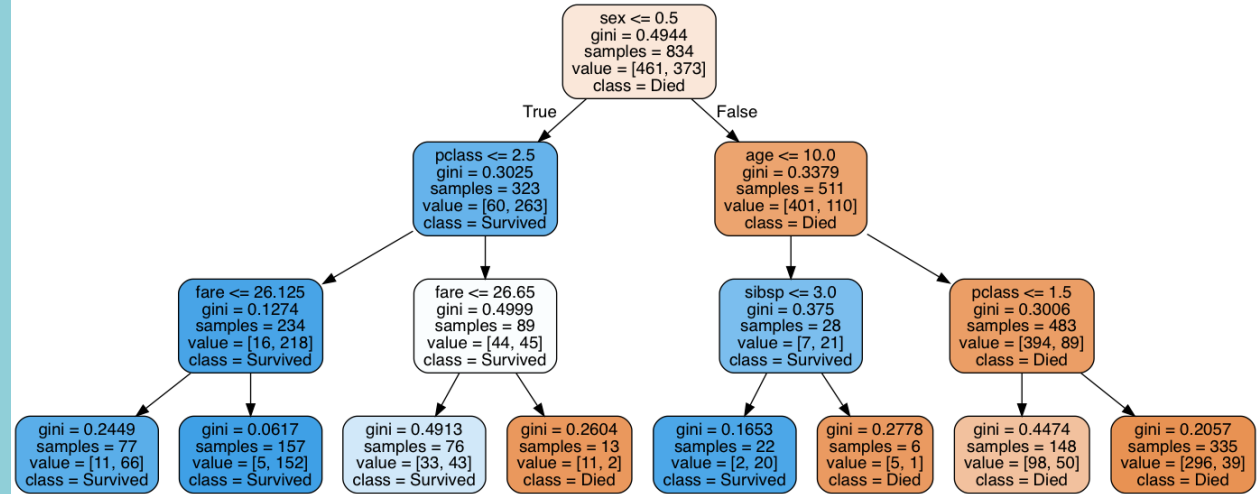




Data terdekat diidentifikasi dengan jarak. Ada beberapa metode untuk mengukur jarak. Salah satu metode yang populer adalah Euclidean.

# Desicion Tree

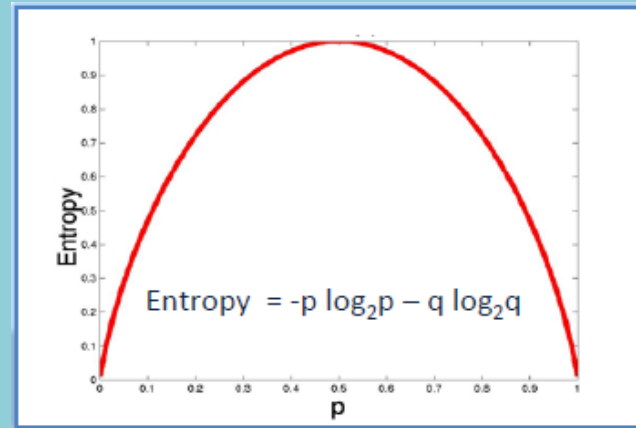
Decision tree merupakan metode pengklasifikasian yang paling umum digunakan oleh machine learning. Metode ini merupakan metode sederhana untuk mengklasifikasikan data. Pada dasarnya, algoritma akan membagi data menjadi dua kondisi, dan terus diturunkan (dibagi dua kondisi) sampai kondisinya tidak bisa dibagi lagi.





## Entropy decision tree:

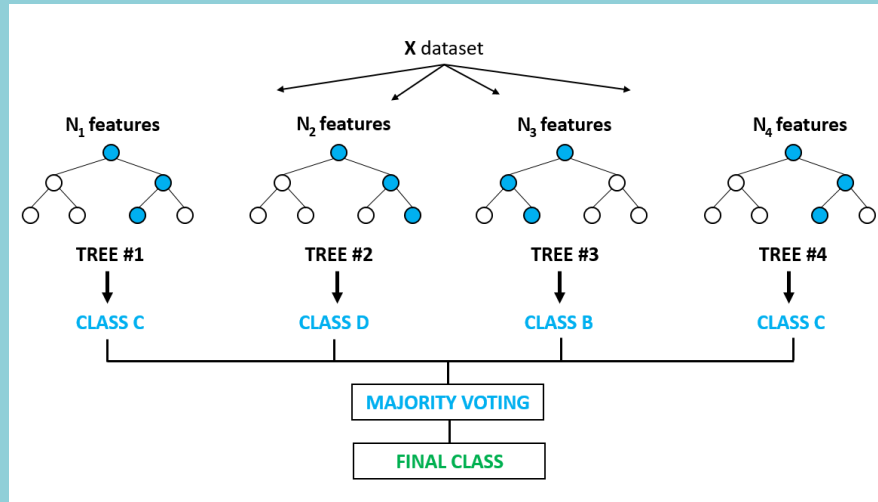
- Ukuran ketidakpastian terkait dengan variable acak.
- Semakin tinggi Entrophy semakin tinggi ketidakpastian, semakin kecil entrophy semakin rendah ketidakpastian.
- Entrophy bersyarat.



$$\text{Entropy} = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$$

# Random Forest

Random forest merupakan turunan atau metode yang berasal dari pengembangan decision tree. Decision tree yang ada di dalam random forest terdiri dari berbagai macam bootstrap yang berbeda. Random forest mengkombinasikan decision tree untuk menghasilkan akurasi yang lebih tinggi.



Evaluation metrics atau evaluasi pengukuran prediksi dapat digunakan dengan :

- Accuracy, Dari semua data yang terprediksi, berapa benar. Untuk melihat performance secara keseluruhan.
- Recall, Untuk menilai dari sudut pandang aktual dari satu kelompok data.
- Precision, Untuk menilai dari sudut pandang prediksi dari satu kelompok data.

		Nilai sebenarnya	
		TRUE	FALSE
Nilai prediksi	TRUE	TP (True Positive) <i>Correct result</i>	FP (False Positive) <i>Unexpected result</i>
	FALSE	FN (False Negative) <i>Missing result</i>	TN (True Negative) <i>Correct absence of result</i>

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

## Evaluation Metrics

# Session 33

## Classification II



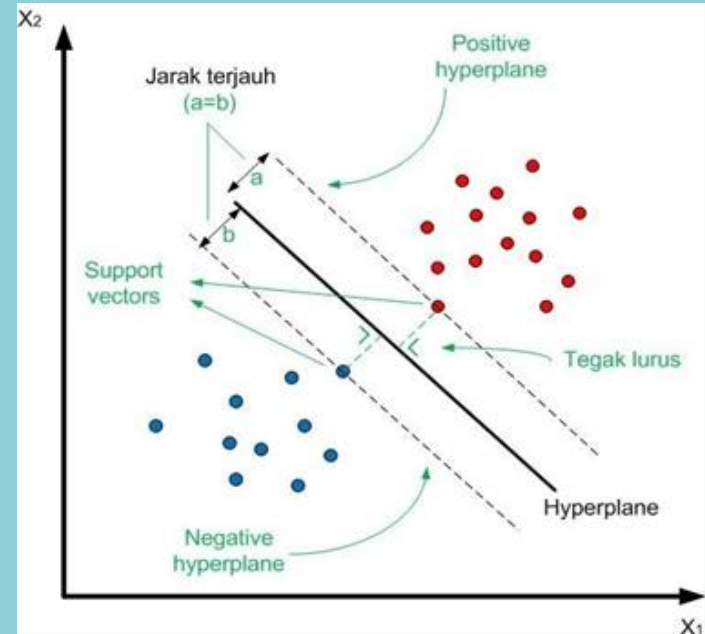
## *Support Vector Machine (SVM)*

**SVM merupakan suatu teknik untuk melakukan prediksi, baik prediksi dalam kasus regresi maupun klasifikasi.**

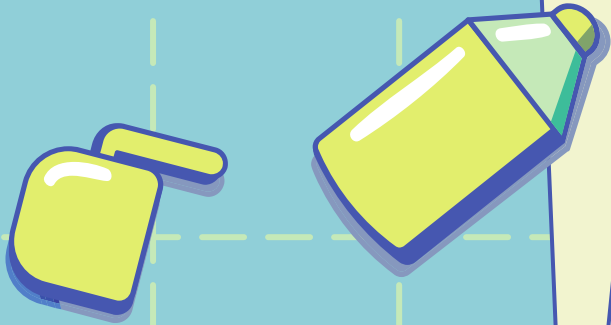


Teknik SVM digunakan untuk mendapatkan fungsi pemisah (*hyperplane*) yang optimal untuk memisahkan observasi yang memiliki nilai variabel target yang berbeda.

Penentuan kelas ditentukan oleh garis *hyperplane*. Kita dapat membuat banyak *hyperplane*, lalu kita menentukan mana *hyperplane* yang cocok.

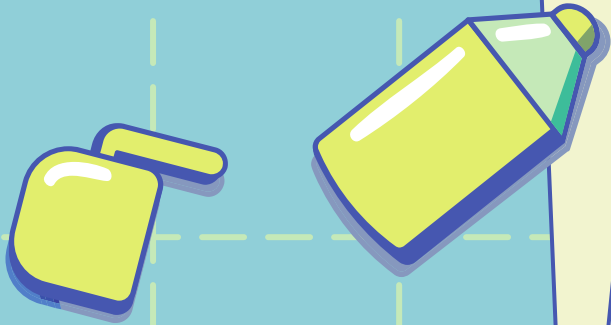


## Kelebihan SVM



- Bekerja dengan baik dengan pembagian margin yang jelas
- Efektif pada data dengan dimensi/fitur yang banyak
- Efektif pada data yang memiliki dimensi yang lebih besar daripada jumlah sampel
- Lebih hemat memory

## Kekurangan SVM

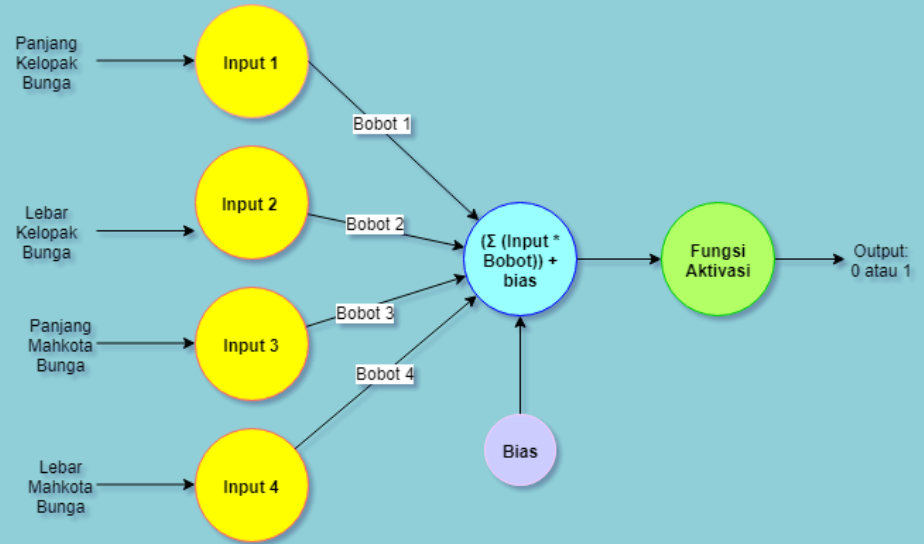


- Tidak cocok pada data yang besar karena membutuhkan waktu training yang lama
- Tidak cocok jika data target memiliki batasan yang saling overlapping
- Tidak secara langsung memberikan probabilitas



# Perceptron

Perceptron adalah salah satu metode Jaringan Syaraf Tiruan (JST) sederhana yang menggunakan algoritma training untuk melakukan klasifikasi secara linier. Perceptron digunakan untuk melakukan klasifikasi sederhana dan membagi data untuk menentukan data mana yang masuk dalam klasifikasi dan data mana yang diluar klasifikasi. Perceptron dapat kita gunakan untuk memisahkan data yang dapat kita bagi menjadi 2 kelas, misalnya kelas C1 dan kelas C2.



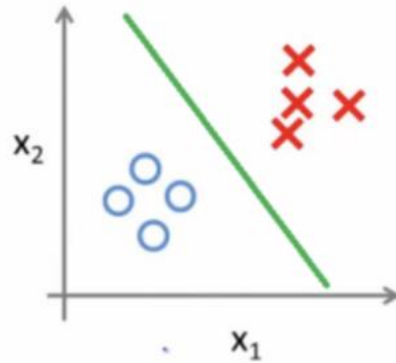
## Multiclass

Dalam machine learning, multi class atau multinomial classification adalah masalah dalam contoh klasifikasi ke dalam satu dari tiga atau banyak kelas. Walaupun beberapa algoritma klasifikasi biasanya mengizinkan penggunaan lebih dari dua kelas (ada yang menggunakan algoritma biner), namun hal ini dapat diubah menjadi classifier multinomial dengan berbagai strategi. Klasifikasi multi class tidak bisa disamakan dengan klasifikasi multi label, dimana beberapa label harus diprediksi untuk setiap instance.

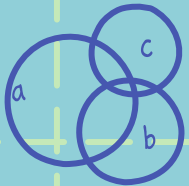
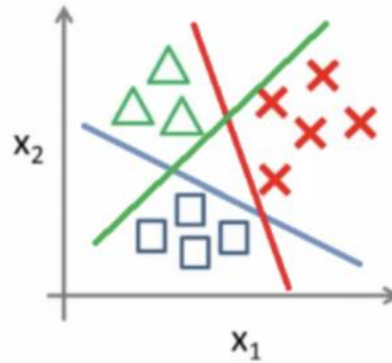


# Multiclass

Binary classification:



Multi-class classification:



$$\sqrt{x-y}$$

$$E=mc^2$$

$$(x-y)^2$$

*Thank you!*

## **Our Team**

- 1. Aldiva Wibowo**
- 2. Asprizal Rizky**
- 3. Gilang Rahmat R**
- 4. Lutfia Humairosi**
- 5. Millenia Winadya P**