Data Bangalore

Machine Learning Preparation :

How to Get Best Data?

Start Now

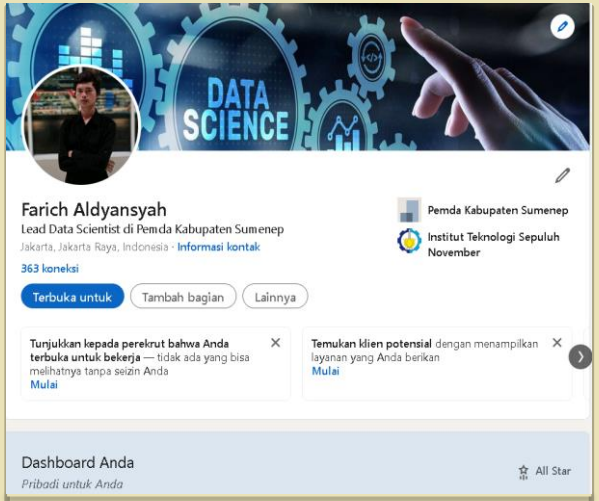
My Journey

amartha

airy

KOMINFO

Amploy

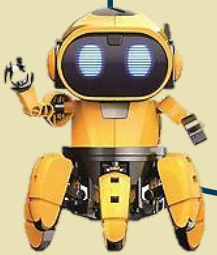


Farich aldyansyah

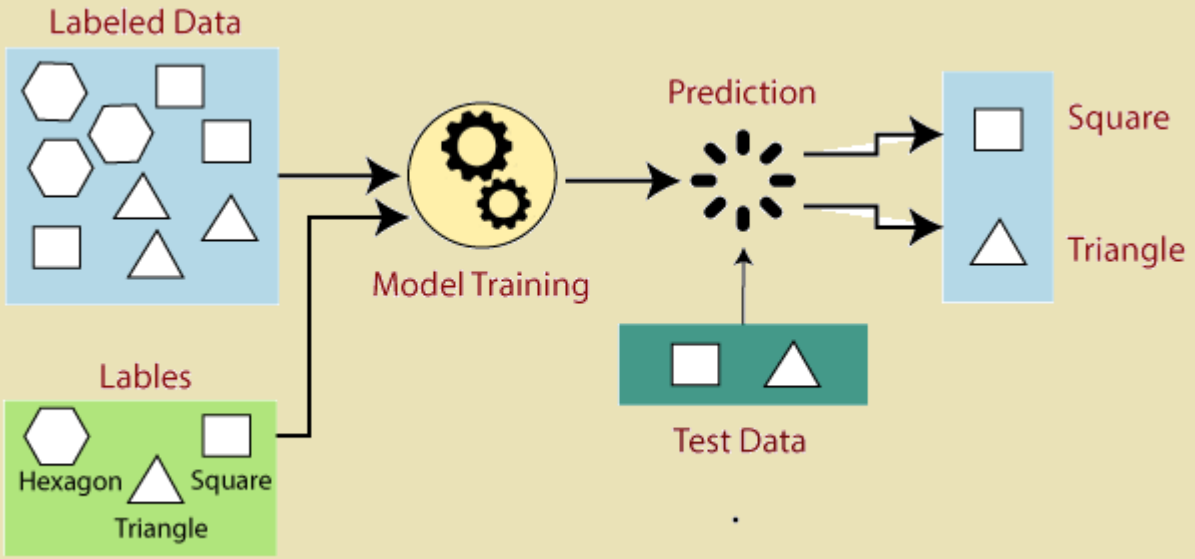
Contents of This Session

- Review Machine Learning dan Data Science
- Cara Mengelolah Feature dengan baik dan benar

Yuk! Bersihkan Data

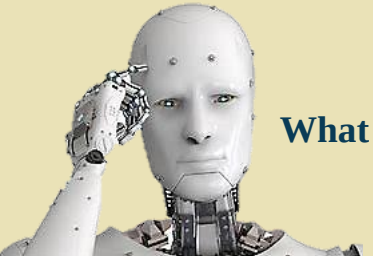


Review Machine Learning

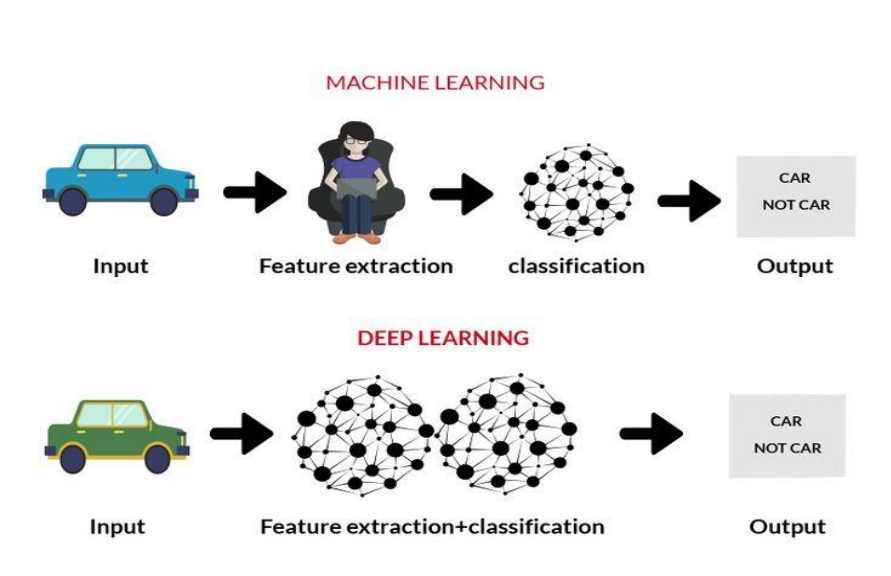


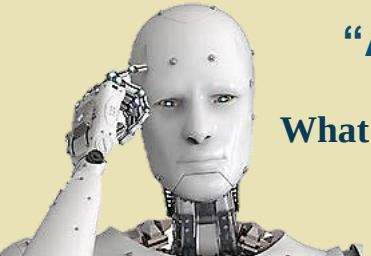
Machine Learning ialah bagaimana proses computer belajar terhadap data yang sudah diberi label/ tanda

What do you think?



Machine Learning vs Deep Learning





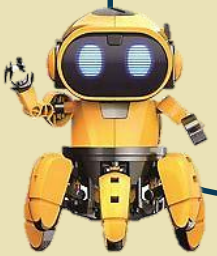
“Apakah semua kasus data harus menggunakan deep learning?”

What do you think?

Bagaimana cara mengelolah feature ?

Dengan **preprocessing data**,
Preprocessing data ialah langkah yang dilakukan oleh data science dengan mentransformasikan data kesuatu format yang dapat diterima baik oleh model.

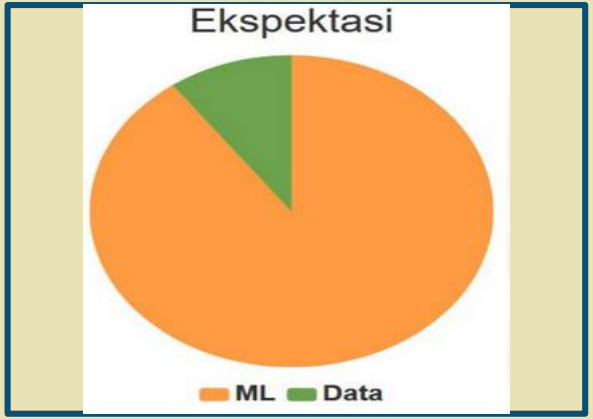
Garbage in Garbage out!



Realita bekerja dibidang data

Apakah data science harus berkutat didata atau modeling?

Ekspektasi



Realita

Data Kotor

Penyebab data kotor?

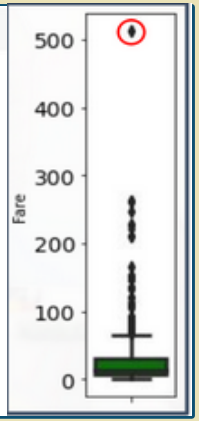
Asal Data Kotor

- Nilai data kosong (missing value)
- Perbedaan waktu pembaruan sistem
- Scammer, abuser
- Salah join tabel
- Data engineering mistakes

Contoh Data Kotor

```
1 df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```



Bagaimana cara preprocessing data?

- Membersihkan/menambal data-data yang kosong
- Menghilangkan data yang redundan
- Menghilangkan data duplikat yang tidak diinginkan
- Mengubah tipe/format data
- Mengubah skala/distribusi data untuk mempermudah machine learning
- Menambahkan data sintetis/duplikat

Garbage in Garbage out!

01

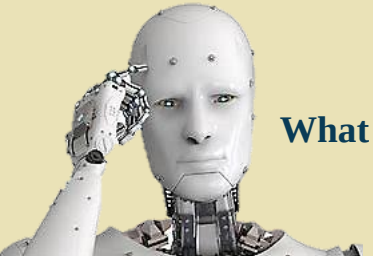
memeriksa dataset

Melihat bagaimana **kondisi dataset!**

Ada Missing Value?

Missing value adalah istilah untuk data yang hilang.
Perhatikan bahwa data ini memiliki berbagai macam cara untuk mengatakan bahwa data pada *cell* tertentu adalah *missing*, misalnya:

- *cell*nya dikosongkan
- ditulis dengan n/a, NA, na, ataupun NaN (biasanya ada dipython)
- ditulis dengan symbol –
- ataupun mempunyai nilai yang cukup aneh seperti nilai 100 pada variable sampo

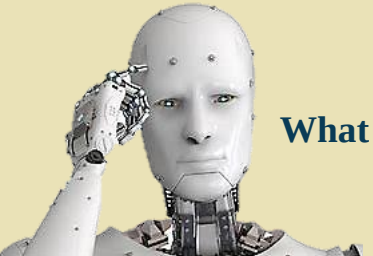


What do you think?

Tipe Missing Value?

- Missing completely at random (MCAR) : Data hilang secara acak, dan tidak berkaitan dengan variabel tertentu.
- Missing at random (MAR) : Data di suatu variabel hilang hanya berkaitan dengan variabel respon/pengamatan.
- Missing not at random (MNAR) : Data di suatu variabel y berkaitan dengan variabel itu sendiri, tidak terdistribusi secara acak.

Pada MCAR dan MAR, kita boleh menghilangkan data dengan *missing value* ataupun mengimputasinya. Namun pada kasus MNAR, menghapus *missing value* akan menghasilkan bias pada data. Menambahkan data tidak selalu bagus.

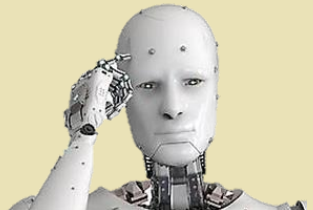


What do you think?

Antisipasi Dataset Kotor

1. Membersihkan Dataset (Destroy Dataset)

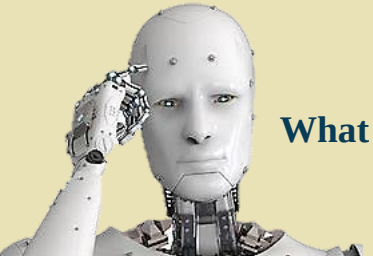
- Ketika mempunyai missing value <10% lebih baik melakukan penghapusan baris dataset yang kosong (***Listwise Deletion)***
- Ketika mempunyai missing value >30% lebih baik hapus variable/ feature datasetnya
- Ketika mempunyai variable/ feature yang nilainya memiliki unique yang banyak lebih baik hapus variable/ feature datasetnya
- Ketika mempunyai variable/ feature yang tidak memiliki hubungan dengan label boleh dihapus variable/ feature datasetnya



Mengabaikan Missing Value?

Beberapa algoritma machine learning atau metode analisis lainnya dapat dengan sendirinya menghandle missing value, contohnya adalah decision tree, k-Nearest Neighbors (kNN), Gradient Boosting Method (GBM) yang dapat mengabaikan missing value, ataupun XGBoost yang dapat mengimputasi sendiri missing value pada data.

Ataupun jika ada beberapa kolom yang tidak memberikan informasi apa apa, kita dapat membiarkan missing value ada di kolom tersebut karena kolom tersebut pun tidak memberikan informasi yang signifikan

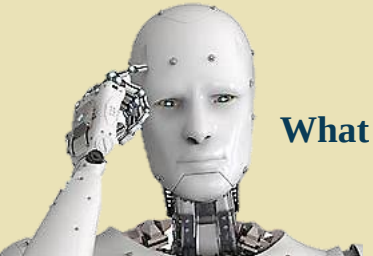


What do you think?

Mengimputasi Missing Value?

Kita dapat menggantikan missing value tersebut dengan suatu nilai, ada beberapa metode dalam mengimputasi missing value, sebagai berikut :

- Univariate Imputation : Imputasi dengan median / mean / modus
- Multivariate Imputation : Single Imputation, Metode metode yang dapat digunakan adalah memprediksi nilai *missing* dengan menggunakan metode metode *supervised learning* seperti kNN, regresi linear, regresi logistik (untuk data kategorik)
- Advance Imputation : *missing value* pada data *Time Series using* Interpolasi Linear dengan memperhitungkan tren seasonal

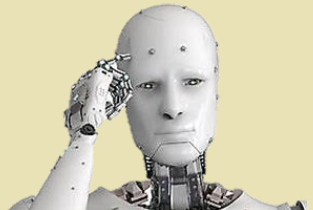


What do you think?

Antisipasi Dataset Kotor

2. imputation data-data yang kosong (numeric type)

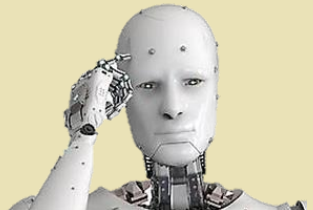
- Isi nilai yang kosong/ missing value dengan rata-rata/ mean pada kolom, jika bentuk distribusi data mendekati normal
- Isi nilai yang kosong/ missing value dengan median pada kolom, jika bentuk distribusi data skew “dikatakan skew apabila angka skew >|2.5|”
- Isi nilai yang kosong/ missing value dengan min/ max/ angka tertentu jika kita memiliki dasar/ referensi pada observasi data (baris dataset)
- Isi nilai yang kosong/ missing value dengan nilai regresi



Antisipasi Dataset Kotor

3. imputation data-data yang kosong (categorical type)

- Isi nilai yang kosong/ missing value dengan nilai tertentu dengan dasar kuat
- Isi nilai yang kosong/ missing value dengan modus pada kolom
- Isi nilai yang kosong/ missing value dengan nilai regresi logistik



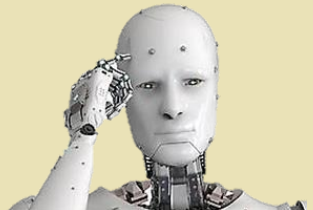
Antisipasi Dataset Kotor

4. duplicated data

Yaitu dataset dengan keadaan memiliki nilai observasi (baris data) yang kembar (terduplikasi), langkahnya :

- Cek apakah ada data yang terduplikasi?
- Hapus observasi data yang kembar/ terduplikasi

Terkadang kita sengaja membuat duplikasi untuk memenuhi asumsi imbalance learning



Outliers Data

Yaitu kondisi dimana data point (baris) yang nilainya ekstrim/jauh berbeda dari data-data lain pada umumnya. Dimana bisa muncul dari Kesalahan pada pengambilan data dan Keberadaan individu-individu yang ‘spesial’.
“Outlier yang perlu ditangani adalah outliers ekstrim, kenapa?”

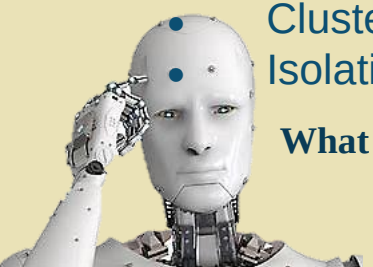
Univariate Outliers : penanganan outlier dengan referensi satu variable

- Quartiles (Boxplot)
- Asumsi Normal
- Asumsi distribusi lain

Multivariate Outliers : penanganan outlier dengan referensi semua variable

- Clustering (DBSCAN)
- Isolation Forest

What do you think?

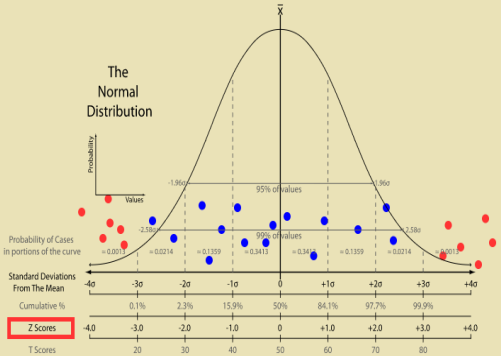


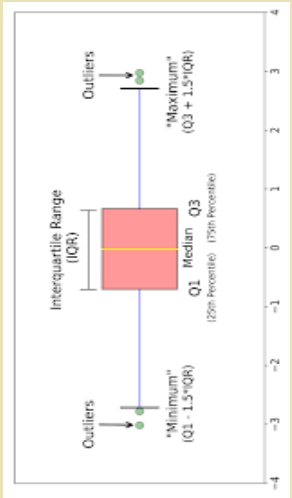
Antisipasi Dataset Kotor

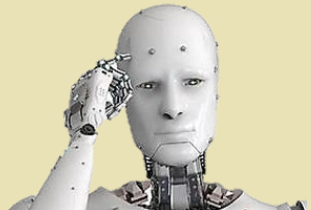
5. Outliers

Dan berikut cara menangani outlier data :

- Menghapus outlier berdasarkan z-score
- Menghapus outlier berdasarkan IQR





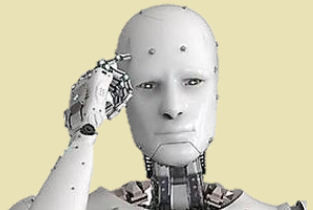


Antisipasi Dataset Kotor

6. Noisy Data

Yaitu kondisi dimana data point (baris) yang nilainya salah sehingga mengganggu insight dan machine learning ‘belajar’. Disebabkan karena :

- Kesalahan instrumen pengukuran: Misal di alat IoT pada saat cuaca buruk/baterai yang lemah.
- Kesalahan input/entry
- Transmisi yang tidak sempurna
- inkonsistensi penamaan

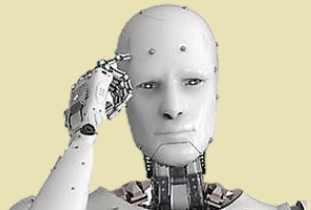


Antisipasi Dataset Kotor

7. Normalization dan Standardization

Normalization adalah proses mengubah nilai-nilai suatu feature menjadi skala tertentu. Sedangkan Standardization adalah proses mengubah nilai-nilai feature sehingga mean = 0 dan standard deviation = 1

- Gunakan standarisasi ketika variabel saya memiliki satuan yang berbeda (yaitu cm vs ft)
- Gunakan normalisasi ketika variabel yang berbeda sifatnya (yaitu usia dan pendapatan adalah dua variabel yang sama sekali berbeda), skala perbedaan 10x dan lebih banyak.



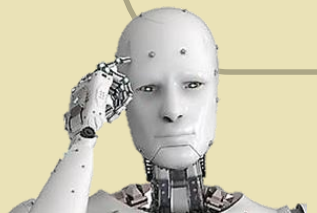
Antisipasi Dataset Kotor

8. feature encoding

- **One Hot Encoding :** We use this categorical data encoding technique when the features are nominal

- **Label encoding :** We use this categorical data encoding technique when the categorical feature is ordinal.

	Degree	Dog	Cat	Sheep	Lion
0	High school				
1	Masters	1	0	0	0
2	Diploma	0	1	0	0
3	Bachelors	0	0	1	0



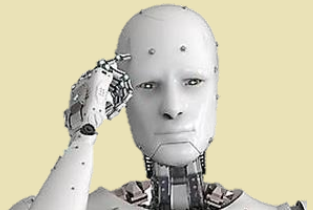
Antisipasi Dataset Kotor

9. class imbalance

Class imbalance adalah sebuah kondisi dalam masalah klasifikasi dimana distribusi nilai unik pada target/label sangat timpang.

Cara menangani class imbalance :

- Berikan ukuran akurasi yang lebih ‘pintar’ (sesi selanjutnya)
- Hilangkan class imbalance pada data dengan over/undersampling
- Gunakan K-fold Cross-Validation dengan cara yang benar



02 Modelling

Lakukan modelling pada data yang sudah dilakukan preprocessing data

Build Models

Buat model berdasarkan kebutuhan dataset dan goals business understanding

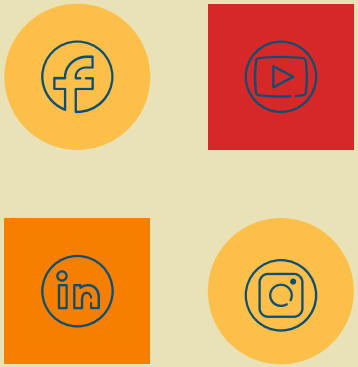
Supervised Learning

- Classification
- Regression

Unsupervised Learning

- Clustering
- Dimentional Reduction
- Association Rule

THANK YOU



Do you have any questions?
Get in databangalore.co.id

Linkdln : Farich Aldyansyah