DigitalSkola
Uncover The World of Digital Skills with Us

Day 3: Regression Models

Hands-on Building Machine Learning Regression Model

Table of Content

What will We Learn Today?

1. Pengertian Regresi dan *use-case Machine Learning*
2. Pemahaman Simple Linear Regression dan Multiple Linear Regression
3. Metrics evaluasi pada regresi (MAE, MSE, R2 score)
4. Diagnostik Regresi *Machine Learning Model*

Definisi *Regression Model*

Regresi → Metode prediktif yang digunakan untuk memodelkan hubungan antara variabel input (independen) dan output (dependen) dalam bentuk nilai kontinu (numerik)

Games Regression Model

Contoh Regression Machine Learning Model

No	Jenis *Real-Case* di Industri	Hasil output
1	Prediksi harga rumah berdasarkan lokasi, luas bangunan, dan jumlah kamar.	Harga Rumah
2	Prediksi harga biaya service kendaraan berdasarkan kondisi kendaraan	Harga biaya service
3	Prediksi harga kendaraan yang dijual berdasarkan kondisi kendaraan	Harga jual kendaraaan

Hasil output dari analisis yang dilakukan, pola tersebut dapat dibuat menjadi model *machine learning* untuk memprediksi data baru yang masuk.

#BertalentaDigital

Simple dan Multiple Linear Regression

Games: Tebak Harga

Jika anda ingin menjual rumah dan ada beberapa variasi rumah di sekeliling rumah anda dengan beda ukuran (sq.ft). Maka bagaimana anda menentukan harga rumah anda?

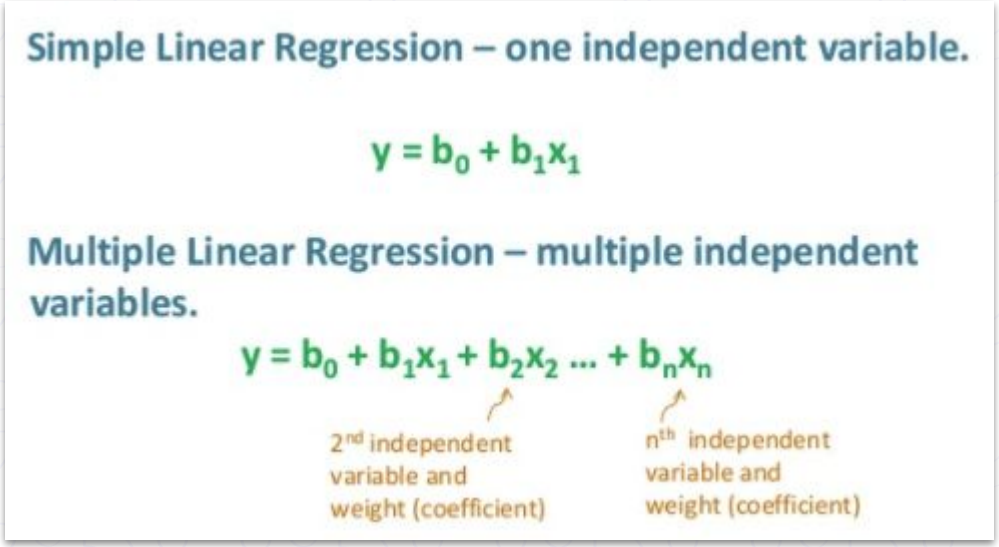
Rumah anda ukuran 3000 sq.ft, berapa harga rumah anda?

Perbedaan

Simple Linear Regression dan Multiple Linear Regression

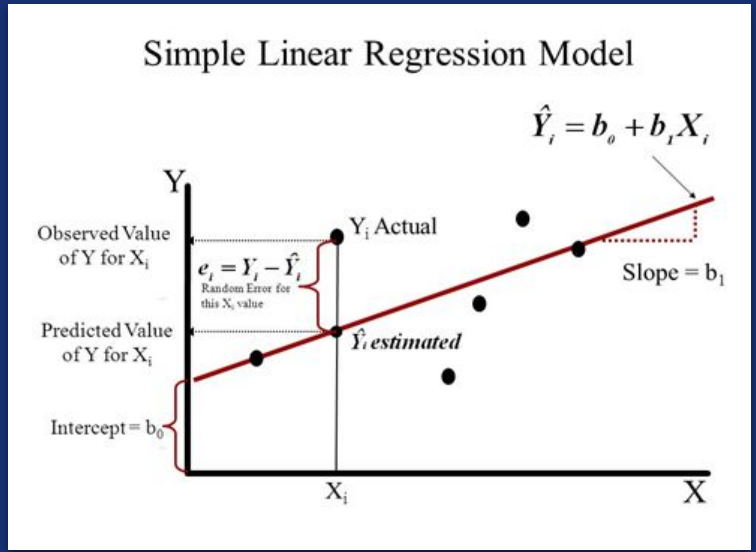
Simple linear regression adalah ketika terdapat satu input variabel dan satu output variable

Multiple linear regression adalah ketika terdapat beberapa input variabel dan menghasilkan satu output variabel



Linear Regression in Detail

- **Yactual** → nilai sebenarnya dari data yang ada
- **Yestimated** → nilai prediksi dari model machine learning
- **Slope** menunjukkan kecuraman suatu garis Intercept menunjukkan lokasi dimana garis itu memotong sumbu.
- **Slope** dan **intercept** menentukan hubungan linear antara dua variable dan digunakan untuk memperkirakan tingkat perubahan rata-rata.



Metrics Evaluasi Regression Model

Jenis Metrics Evaluasi pada Regresi

MAE

Mean Absolute Error

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

MSE

Mean Squared Error

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

R2

R-squared

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

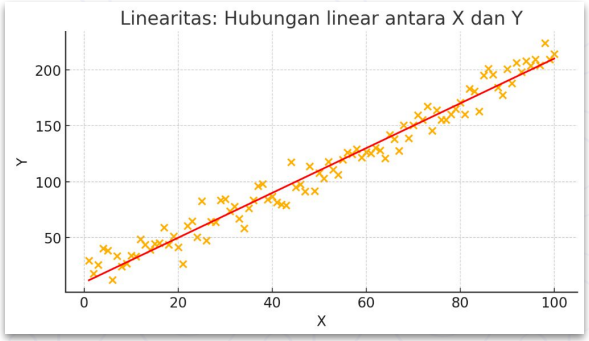
Detail penjelasan →

Jenis Metrics Evaluasi pada Regresi

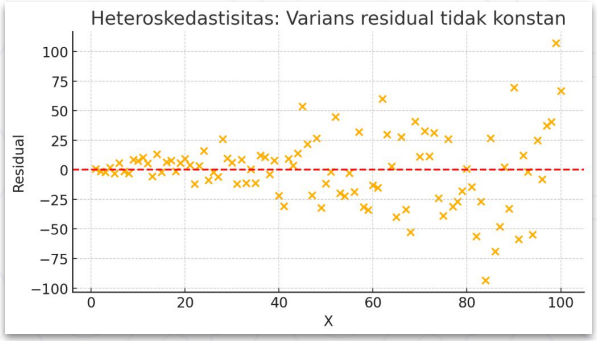
Metrik	Arti	Karakteristik	Ilustrasi
MAE	Rata-rata **selisih absolut** antara nilai prediksi dan aktual	Mudah dipahami, tidak terlalu penalti untuk outlier	Jika prediksi = 102 dan nilai asli = 100, error = 2
MSE	Rata-rata dari **kuadrat selisih** prediksi dan aktual	Memberi penalti lebih besar untuk error besar (karena dikuadratkan)	Jika error 2, MSE = 4; jika error 10, MSE = 100
R^2 Score	Seberapa besar **variasi data** yang bisa dijelaskan oleh model (0–1)	Semakin mendekati 1, model semakin baik	R^2 = 0.9 berarti 90% variasi target bisa dijelaskan model

Asumsi & Diagnostik

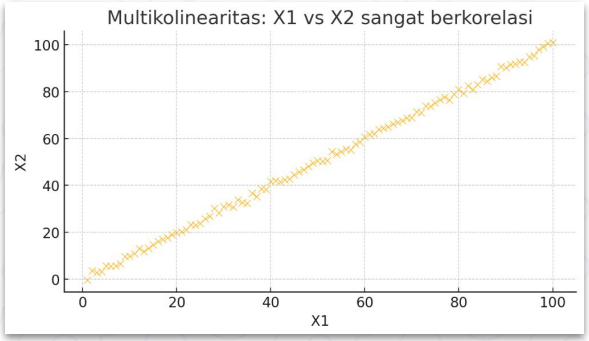
Model Linear Regresi



Linearitas → Menunjukkan hubungan satu variabel data dengan variable data lainnya



Heteroskedastisitas → pelanggaran terhadap asumsi homoskedastisitas (harusnya residual menyebar merata).



Multikolinearitas → kedua input variabel memiliki pola yang sama. Dapat dipilih salah satu saja.

DigitalSkola
Uncover The World of Digital Skills with Us

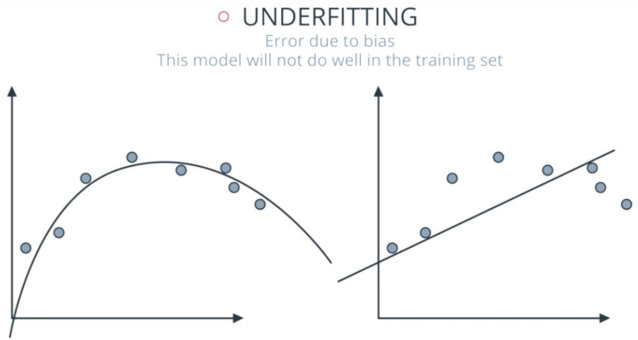
Thank You.

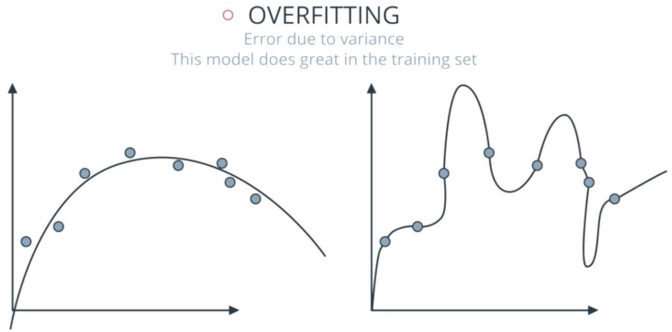
DigitalSkola
Uncover The World of Digital Skills with Us

Day 3: Model Tuning

Tuning up your machine learning model

Masih ingat soal ini?





source: udacity machine learning nanodegree

Table of Content

What will We Learn Today?

1. Apa itu *Hyperparameter Tuning*
2. *Cross Validation* pada *model tuning*
3. Mencari Parameter Optimal untuk Model
4. Praktek Optimasi Machine Learning Model

Hyperparameter Tuning

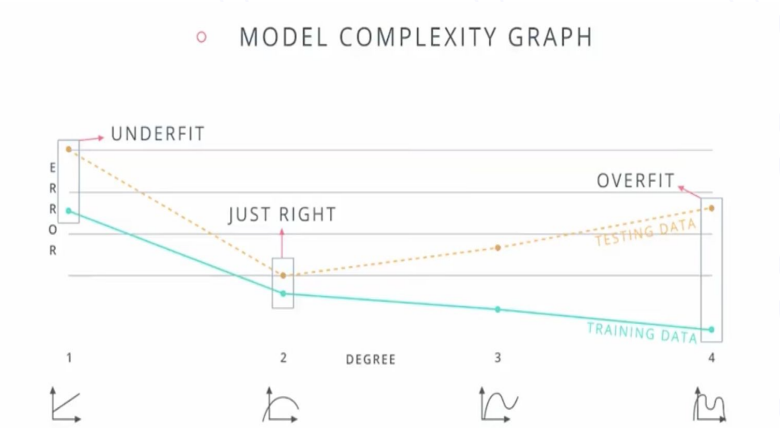
hyperparameter adalah parameter yang ditentukan sebelum model dilatih. Seperti:

- Jumlah pohon dalam Random Forest (n_estimators)
- Nilai C dalam SVM
- Learning rate dalam Gradient Boosting

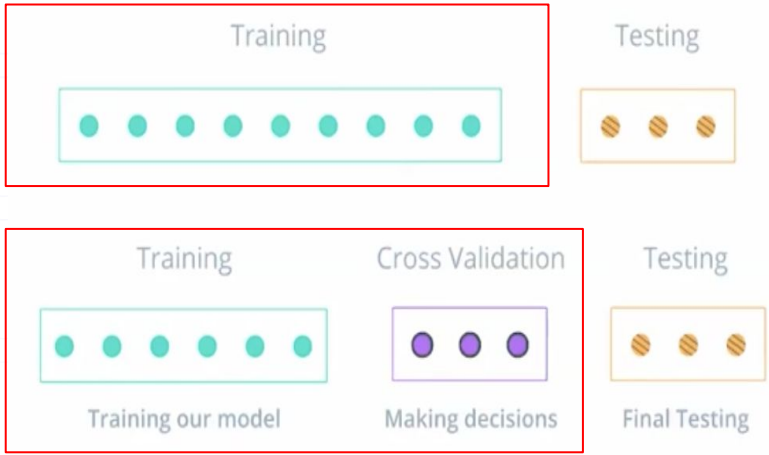
Hyperparameter tuning adalah proses mencari kombinasi hyperparameter terbaik untuk menghasilkan performa model yang optimal.

Cross Validation

Cross-validation adalah teknik untuk mengevaluasi model agar tidak overfitting dan underfitting.



source: udacity machine learning nanodegree

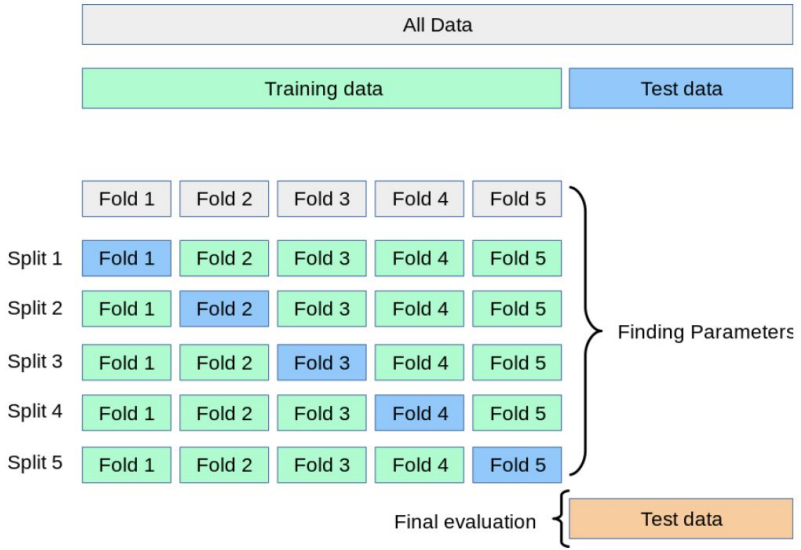


Beberapa Teknik

Pada Cross Validation

Ada beberapa teknik:

- Holdout Validation (train/test split)
- K-fold Cross Validation ✅
- Leave-One-Out (LOOCV)
- Stratified K-fold



K-Fold Cross Validation

Penjelasan detail

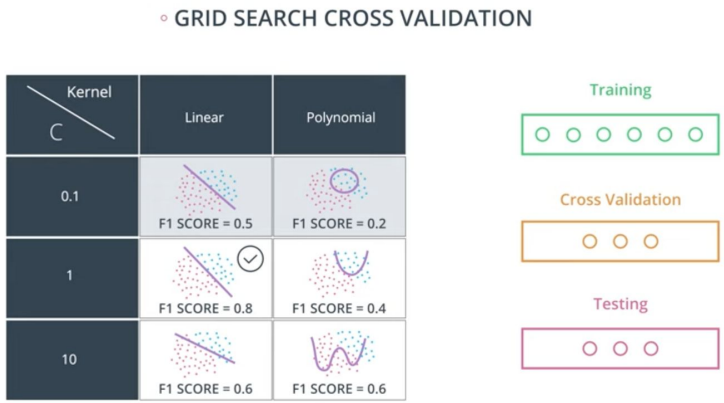
Fold Ke-	Data Latih (Train)	Data Validation	Akurasi
1	D2, D3, D4, D5, D6, D7, D8	D1	0.80
2	D1, D3, D4, D5, D6, D7, D8	D2	0.85
3	D1, D3, D4, D5, D6, D7, D8	D3	0.83
4	D1, D2, D3, D5, D6, D7, D8	D4	0.87
5	D1, D2, D3, D4, D6, D7, D8	D5	0.82
Rata-rata			0.834

Jenis Data	Digunakan untuk	Dalam Cross-Validation
Train	Melatih model di tiap fold	✅ Ya
Validation	Menilai performa model tiap fold	✅ Ya
Test	Evaluasi akhir (1x, tidak ikut tuning)	❌ Tidak ikut CV

Mencari Parameter Optimal

GridSearchCV

GridSearchCV secara otomatis **mencoba semua kombinasi hyperparameter** yang kita tentukan dan mengevaluasinya dengan cross-validation (biasanya k-fold)




```python
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestClassifier

param_grid = {
    'n_estimators': [50, 100],
    'max_depth': [3, 5, None]
}

grid_search = GridSearchCV(RandomForestClassifier(), param_grid, cv=5)
grid_search.fit(X, y)

print("Best Parameters:", grid_search.best_params_)
```

GridSearchCV

Ilustrasi

Percobaan	n_estimators	max_depth	Fold 1	Fold 2	Fold 3	Rata-rata Akurasi (%)
1	50	3	78.5	77.2	79.0	78.2
2	50	5	79.1	80.0	80.5	79.9
3	100	3	80.0	81.0	81.5	80.8
4	100	5	83.0	83.5	84.2	**83.6** ✅

Practice Time!

Buka Notebook Day 3

Lakukan untuk 3 model yang berbeda

Jenis Metrics Evaluasi pada Regresi

MAE

Mean Absolute Error

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

MSE

Mean Squared Error

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

R2

R-squared

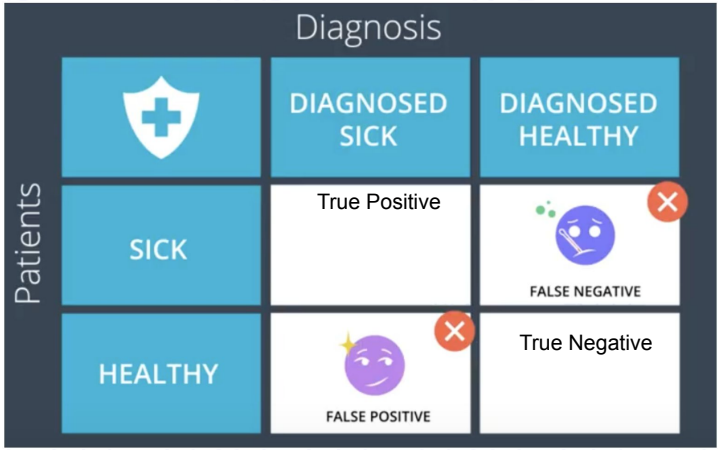
$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Detail penjelasan →

Precision, Recall & F1 Score for Classification

False positive: when the patient is healthy, and you diagnose them as sick

False negative: when the patient is sick, and you diagnose them as healthy



source: udacity machine learning nanodegree

Which one you choose?

$$Precision = \frac{(TruePositives)}{(TP+FP)}$$

Out of all the points predicted to be positive, how many of them were actually positive?

$$Recall = \frac{(TruePositive)}{(TP+FN)}$$

Out of the points that are labeled positive, how many of them were correctly predicted as positive?

F1 score: the average between precision and recall

DigitalSkola
Uncover The World of Digital Skills with Us

Thank You.

DigitalSkola
Uncover The World of Digital Skills with Us

Day 3: Introduction to NLP

Memprediksi kategori data dari input berupa text

Table of Content

What will We Learn Today?

1. Penggunaan *Natural Language Processing*
2. Langkah-langkah *machine learning NLP Classification*
3. Data Cleansing (Stopwords)
4. Text Processing (Stemming & Lemmatization)

Mengapa Natural Processing Language (NLP)?

- Memahami pola text sesuai tujuan
- Mendeteksi teks dari bahasa yang digunakan (positif/negatif)
- Membaca ribuan text dan menafsirkan text tersebut time-consuming dan expensive

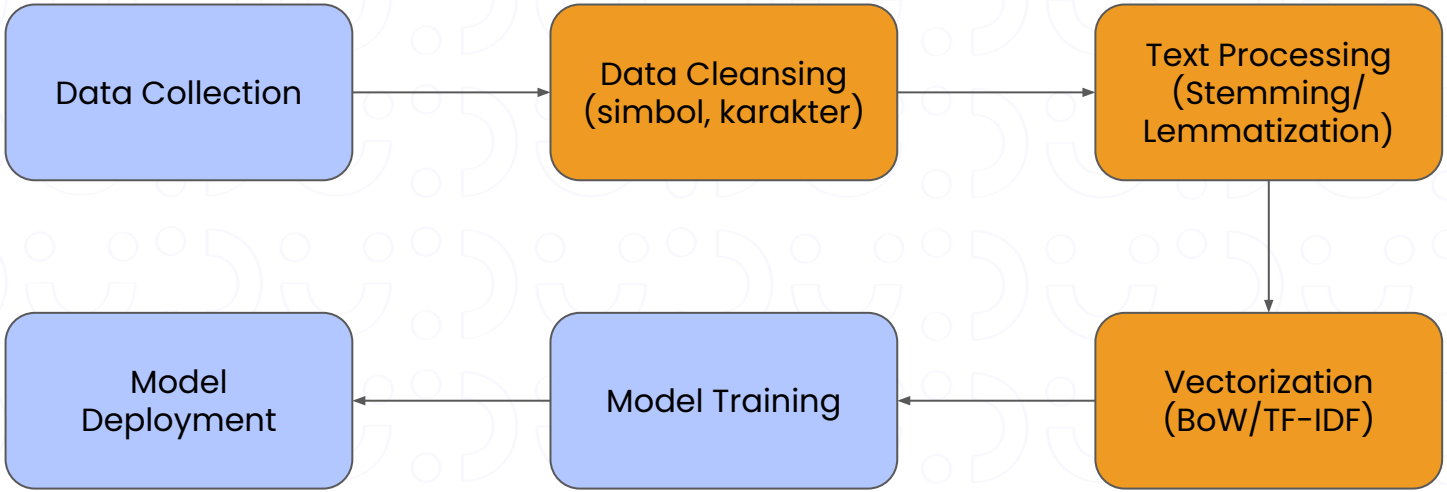
Penjelasan NLP

"Cabang dari AI yang memungkinkan komputer untuk memahami, menafsirkan, dan menghasilkan bahasa manusia."

💡Contoh Industri:

- **Finansial**: Analisis sentimen pelanggan terhadap layanan bank dari review Google atau komentar Twitter.
- **FMCG**: Klasifikasi opini konsumen terhadap produk (positif/negatif) dari e-commerce.
- **Otomotif**: Deteksi keluhan pelanggan terhadap model mobil tertentu dari survey online.

Machine Learning for NLP Classification



Tokenisasi

Tokenisasi adalah proses memecah text menjadi unit-unit kecil yang disebut token, biasanya kata, sub-kata, atau karakter.

💡 Contoh

- Mobil listrik adalah masa depan
- Saya menyukai makanan yang enak dan bergizi

💡 Setelah tokenisasi

- ['Mobil', 'listrik', 'adalah', 'masa', 'depan', '.']
- ['Saya', 'menyukai', 'makanan', 'yang', 'enak', 'dan', 'bergizi']

Data Cleansing (Stopwords)

Stopwords adalah kata-kata umum yang sering muncul tapi tidak menambah makna signifikan, seperti “yang”, “dan”, “adalah”, dll.

💡 Contoh

- Dari kalimat *“Saya merasa mobil ini sangat bagus dan nyaman”*, kata penting adalah “mobil”, “bagus”, “nyaman”.

💡 Stopwords Indonesia (contoh):

['saya', 'yang', 'dan', 'karena', 'ini', 'itu', 'di', 'ke', 'dengan', 'untuk', 'adalah', 'bahwa', 'seperti', 'dalam']

Text Processing

Stemming

Memotong kata ke bentuk dasarnya

Kata Asli	Hasil Stemming
mengantarkan	antar
menyampaikan	sampai
penggunaan	guna
memperbaiki	baik
terpenuhi	penuh
dilakukannya	laku

Lemmatization

Mengubah kata ke bentuk dasar

Kata Asli	Jenis Kata	Lemmatized
better	Adjective	good
children	Noun (plural)	child
went	Verb (past)	go
am/is/are	Verb (to be)	be
studies	Verb/Noun	study
feet	Noun	foot

#BertalentaDigital

Hands-on Python

Buka Notebook Classification ML Model

DigitalSkola
Uncover The World of Digital Skills with Us

Thank You.