Session 30

Advanced Data Preprocessing for ML

Table of Content

What will We Learn Today?

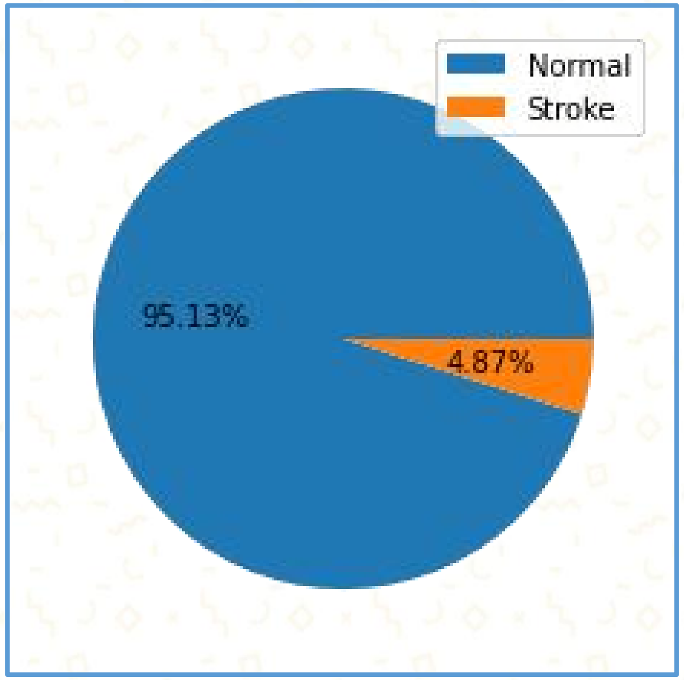
1. Imbalanced Dataset
2. Text Classification
3. Handling Text Data
4. Feature Extraction

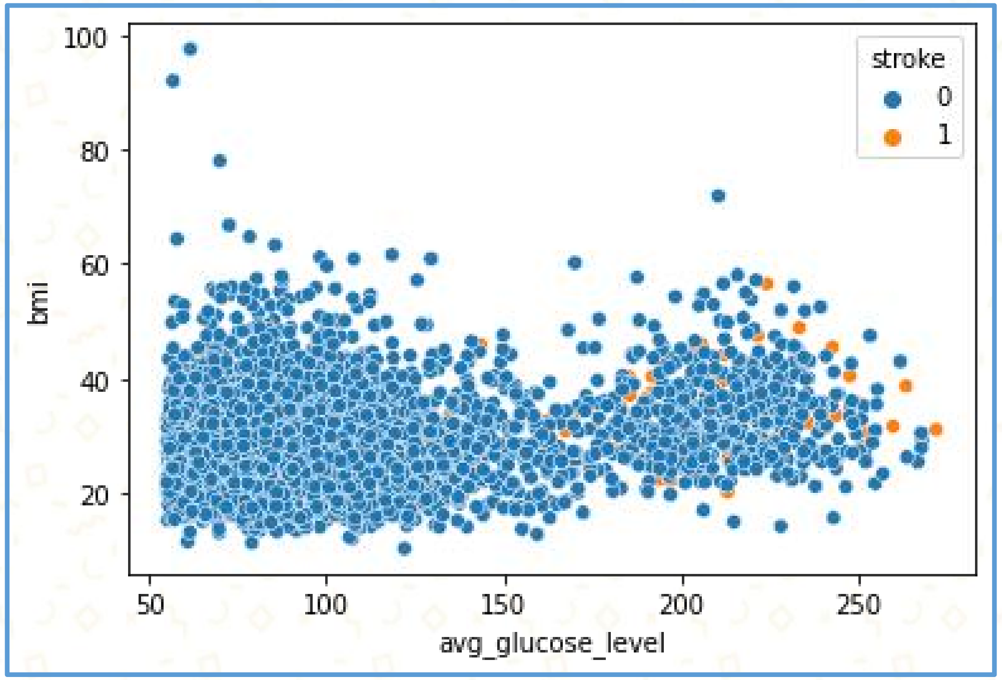
DigitalSkola
Uncover The World Of Digital Skills With Us

Imbalanced Dataset

Imbalanced dataset

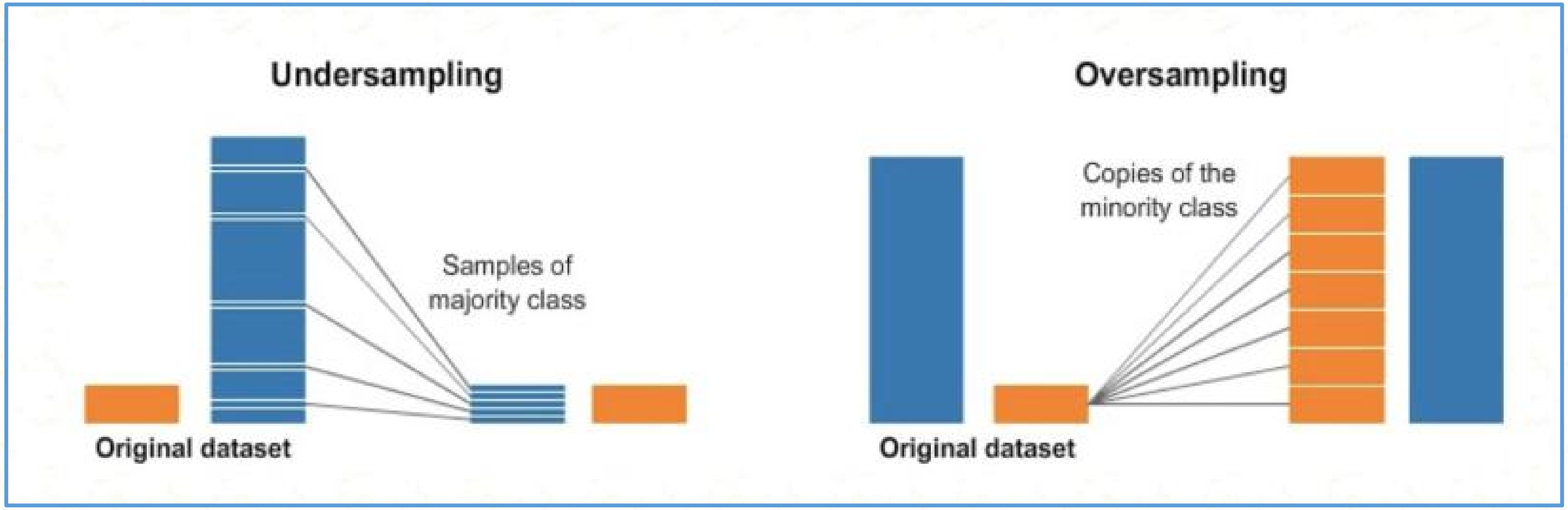
- *Imbalanced dataset* mengacu pada masalah klasifikasi di mana jumlah data per kelas tidak terdistribusi secara merata.
- *https://www.kaggle.com/fedesoriano/stroke-prediction-dataset?select=healthcare-dataset-stroke-data.csv*





How to handle imbalance dataset

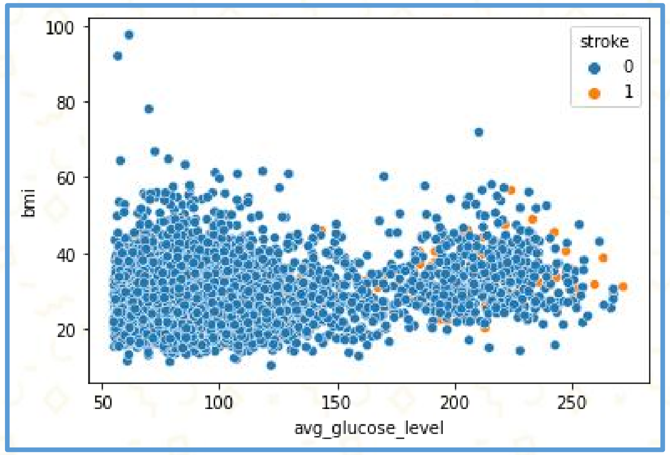
- *Under sampling* = Menyeimbangkan distribusi kelas dengan menghilangkan data dari kelas mayoritas secara acak.
- *Oversampling* = Meningkatkan jumlah instance di kelas minoritas dengan mereplikasinya secara acak.

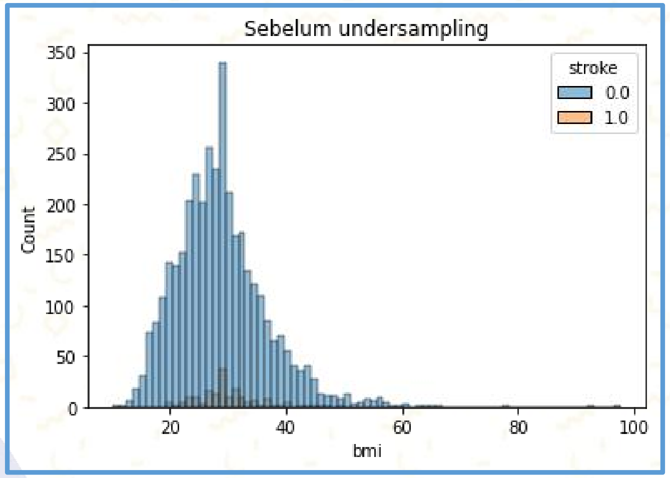


Random Under sampling

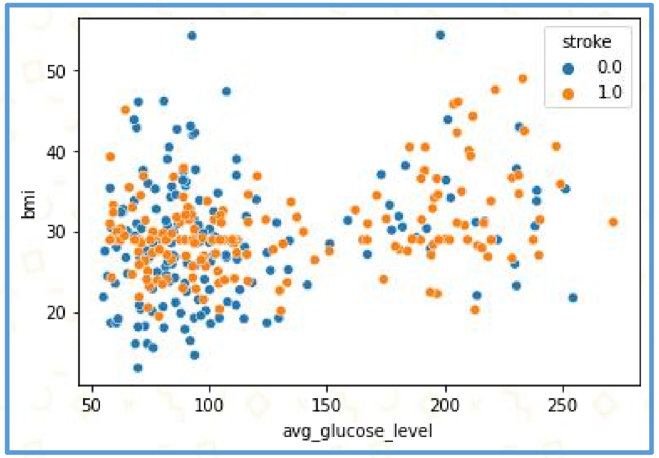
- Membadingkan *training set,* sebelum dan sesudah *undersampling*

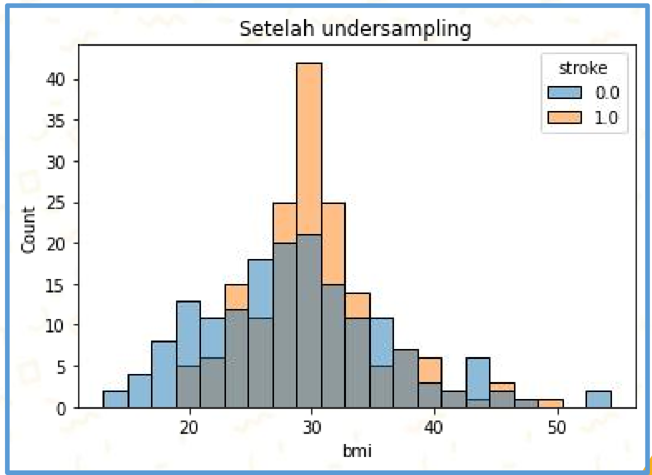
Sebelum





Sesudah



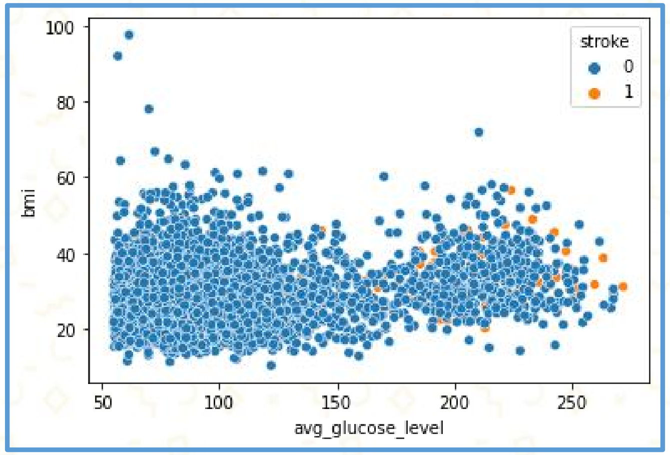



```
Sebelum undersampling
0.0     3663
1.0      169
dtype: int64
Setelah undersampling
1.0     169
0.0     169
dtype: int64
```

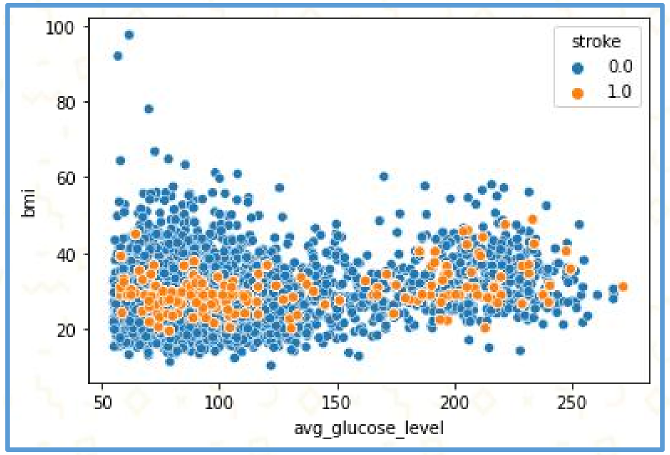
Random Over sampling

- Membadingkan *training set,* sebelum dan sesudah *oversampling*

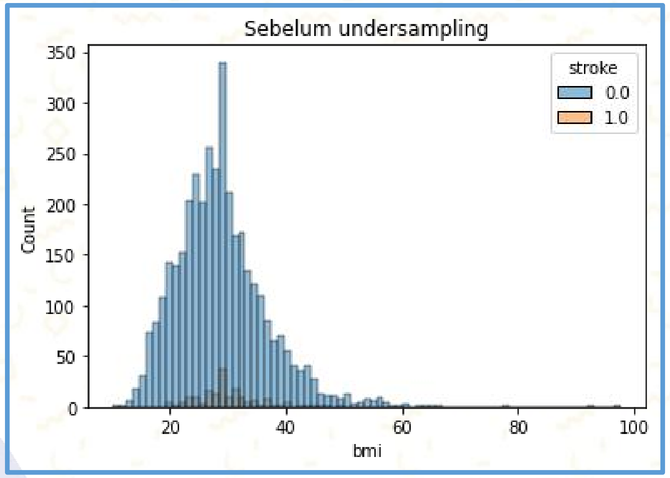
Sebelum

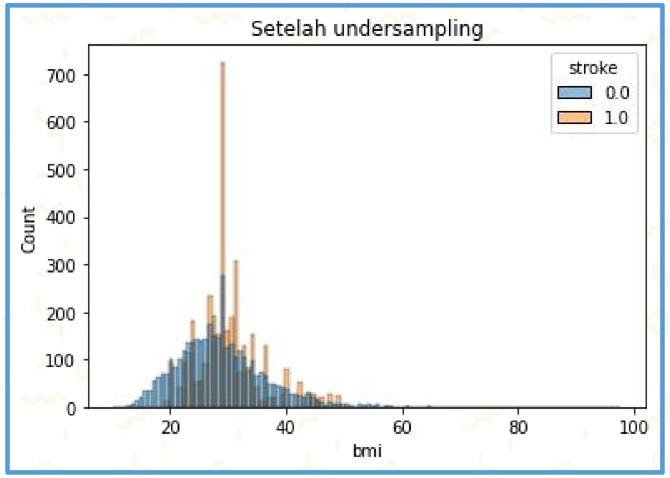


Sesudah



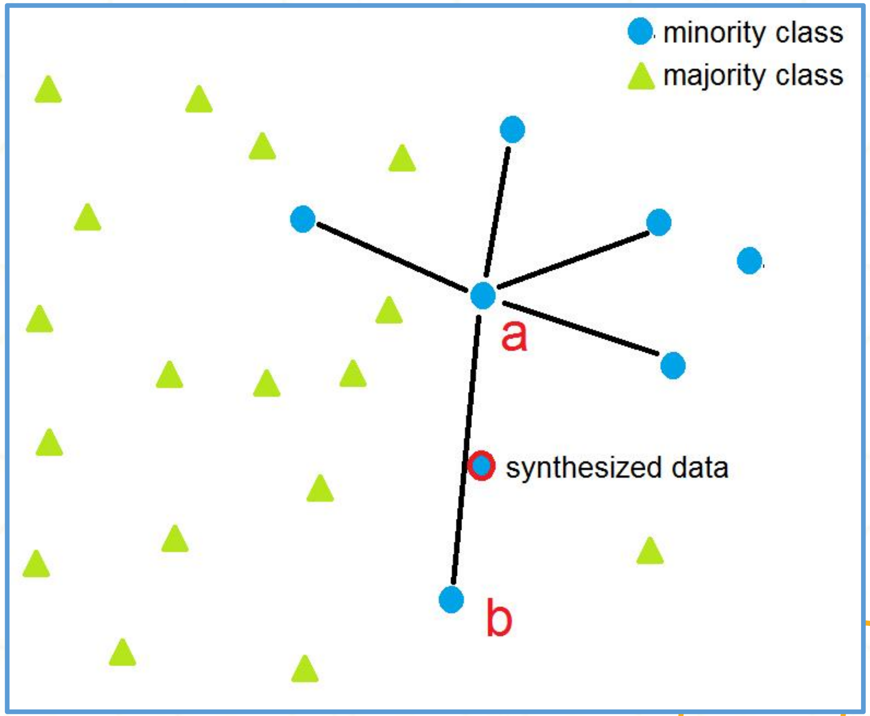

```
Sebelum oversampling
0.0    3663
1.0     169
dtype: int64
Setelah oversampling
1.0    3663
0.0    3663
dtype: int64
```





SMOTE

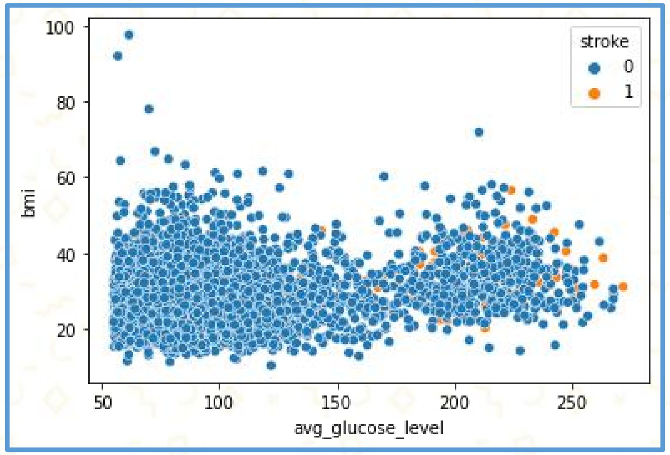
- *SMOTE = Synthetic Minority Oversampling Technique*
- Pendekatan *oversampling* yang menciptakan sampel kelas minoritas secara sintetis.
- Cara kerja:
  - Contoh acak dari kelas minoritas **a** dipilih terlebih dahulu.
  - Kemudian k dari tetangga terdekat (nearest neighbour) untuk contoh tersebut ditemukan (biasanya k=5).
  - Tetangga **b** yang dipilih secara acak.
  - Data sintetik **c** dibuat pada titik yang dipilih secara acak diantara dua data tersebut (**a** dan **b**)

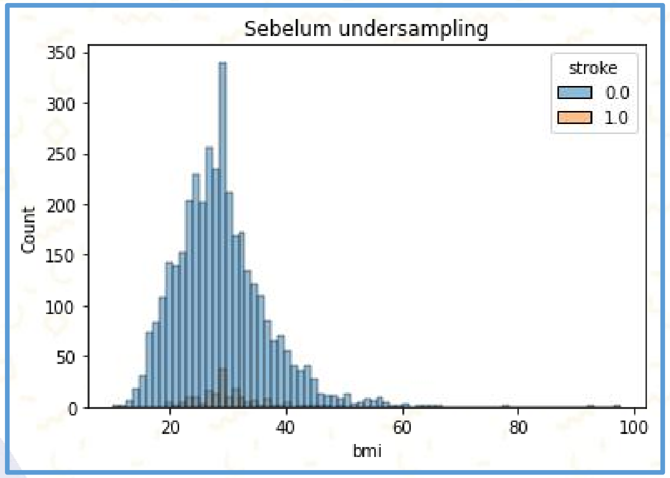


SMOTE

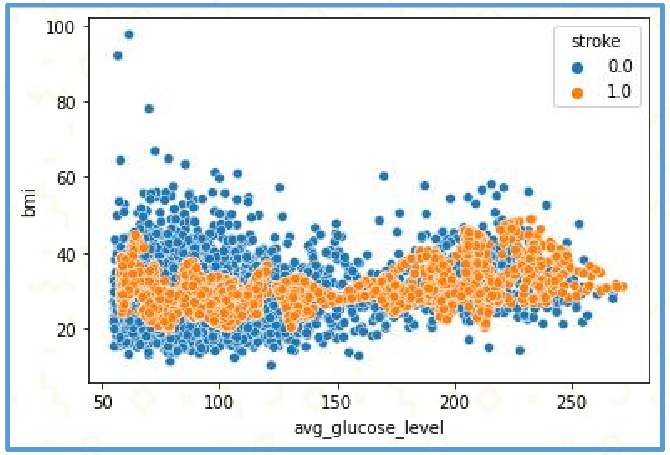
- Membadingkan *training set,* sebelum dan sesudah *SMOTE*

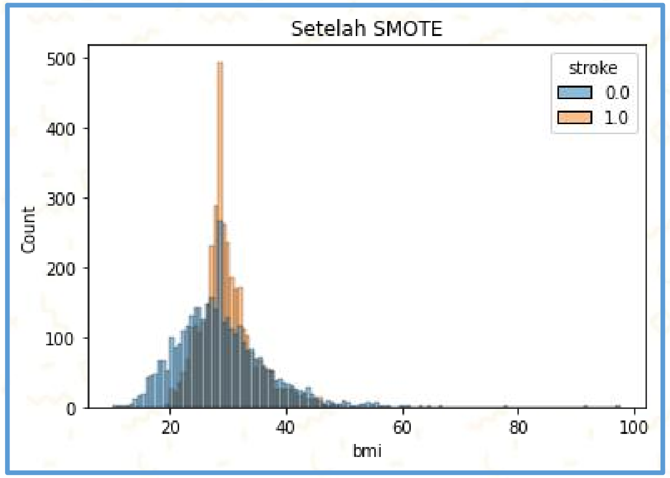
Sebelum





Sesudah

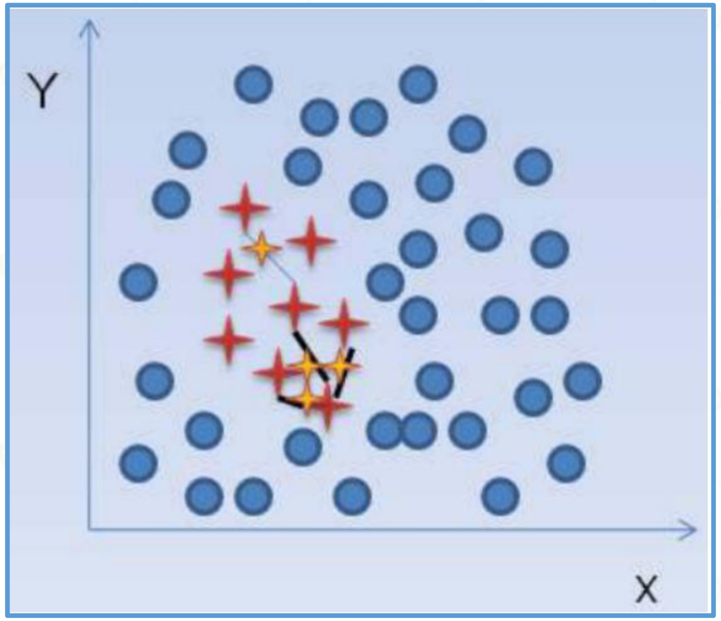


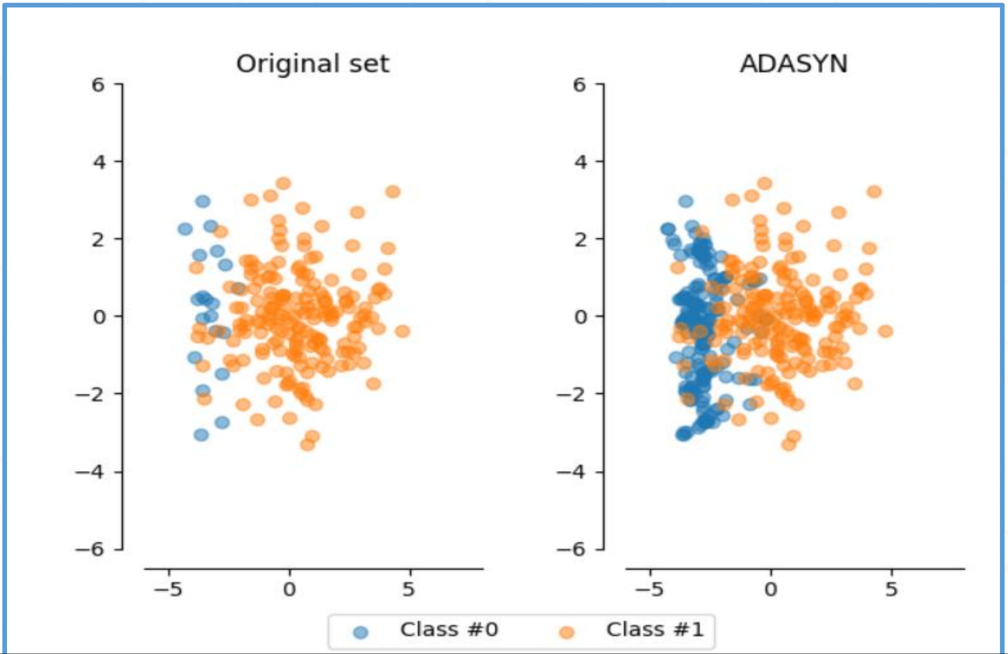



```
Sebelum SMOTE
0.0    3663
1.0     169
dtype: int64
Setelah SMOTE
1.0    3663
0.0    3663
dtype: int64
```

ADASYN

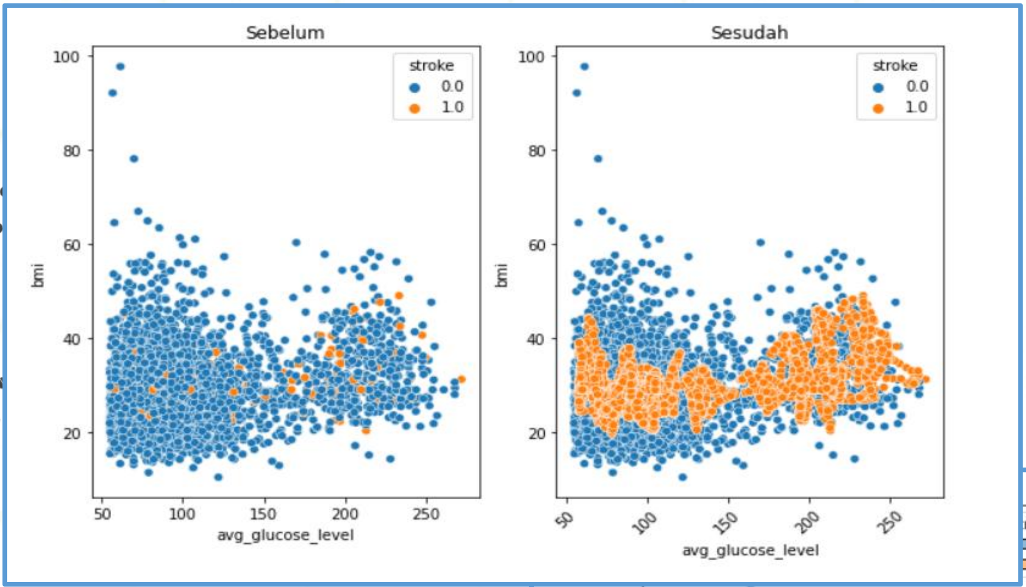
- *ADASYN = Adaptive Synthetic Sampling Approach for Imbalanced Learning*
- Ide penting dari ADASYN adalah menggunakan pembobotan untuk contoh kelas minoritas
- ADASYN akan fokus pada sampel yang sulit untuk diklasifikasikan sementara SMOTE tidak akan membuat perbedaan apa pun.



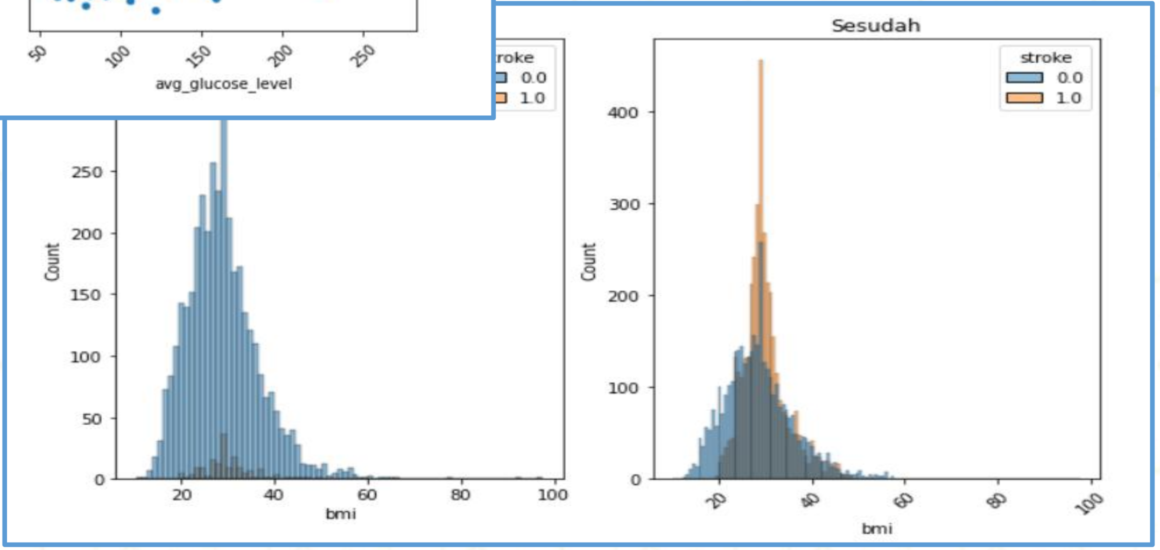


ADASYN

- Membadingkan *training set,* sebelum dan sesudah *ADASYN*




```
Sebelum ADASYN
0.0    3663
1.0     169
dtype: int64
Setelah ADASYIN
1.0    3676
0.0    3663
dtype: int64
```



Text Classification

Dataset types

age	anaemia	creatinine_p	diabetes	ejection_f	high_bloo	platelets	serum_cre	serum_so	sex	smoking	time	DEATH_EVENT
75	0	582	0	20	1	265000	1.9	130	1	0	4	1
55	0	7861	0	38	0	263358	1.1	136	1	0	6	1
65	0	146	0	20	0	162000	1.3	129	1	1	7	1
50	1	111	0	20	0	210000	1.9	137	1	0	7	1
65	1	160	1	20	0	327000	2.7	116	0	0	8	1
90	1	47	0	40	1	204000	2.1	132	1	1	8	1
75	1	246	0	15	0	127000	1.2	137	1	0	10	1
60	1	315	1	60	0	454000	1.1	131	1	1	10	1

Heart failure clinical records

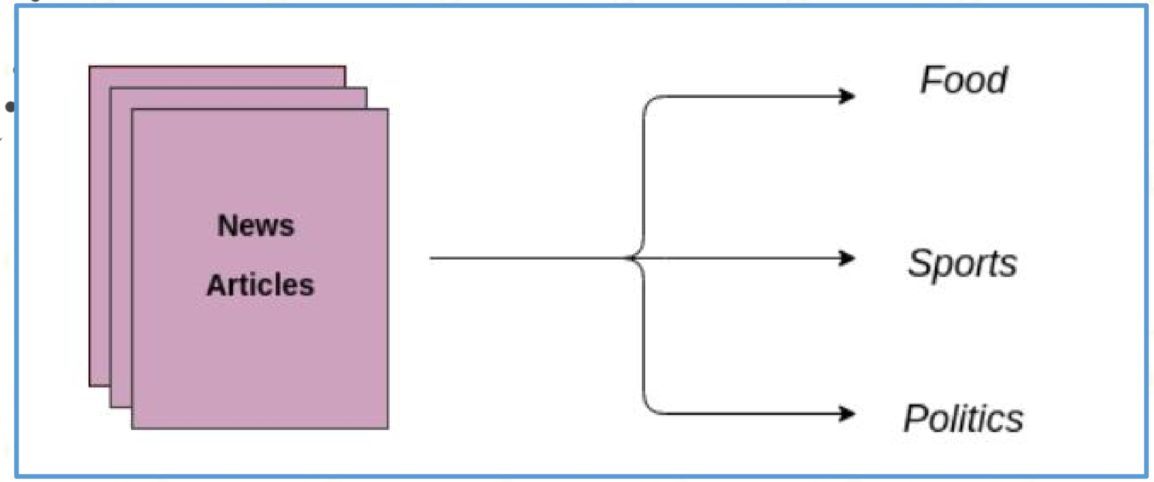
airline_sent	airline_sent	negativerea	negativerea	airline	airline_sent	name	negativerea	retweet_co	text	t
neutral	1			Virgin America		cairdin		0	@VirginAmerica What @dhepburn said.	
positive	0.3486		0	Virgin America		jnardino		0	@VirginAmerica plus you've added commercials to the	
neutral	0.6837			Virgin America		yvonnalynn		0	@VirginAmerica I didn't today... Must mean I need to t	
negative	1	Bad Flight	0.7033	Virgin America		jnardino		0	@VirginAmerica it's really aggressive to blast obnoxiou	
negative	1	Can't Tell	1	Virgin America		jnardino		0	@VirginAmerica and it's a really big bad thing about it	
negative	1	Can't Tell	0.6842	Virgin America		jnardino		0	@VirginAmerica seriously would pay $30 a flight for seats that didn't have this playing. it's really the only bad thing about flying VA	
positive	0.6745		0	Virgin America		cjmcginnis		0	@VirginAmerica yes, nearly every time I fly VX this â€œ	
neutral	0.634			Virgin America		pilot		0	@VirginAmerica Really missed a prime opportunity for	
positive	0.6559			Virgin America		dhepburn		0	@virginamerica Well, I didn'tâ€¦but NOW I DO! :-D	
positive	1			Virgin America		YupitsTate		0	@VirginAmerica it was amazing, and arrived an hour ea	
neutral	0.6769		0	Virgin America		idk_but_youtube		0	@VirginAmerica did you know that suicide is the secon	
positive	1			Virgin America		HyperCamiLax		0	@VirginAmerica I <3 pretty graphics. so much better	
positive	1			Virgin America		HyperCamiLax		0	@VirginAmerica This is such a great deal! Already think	
positive	0.6451			Virgin America		mollanderson		0	@VirginAmerica @virginmedia I'm flying your #fabulou	
positive	1			Virgin America		sjespers		0	@VirginAmerica Thanks!	
negative	0.6842	Late Flight	0.3684	Virgin America		smartwatermelon		0	@VirginAmerica SFO-PDX schedule is still MIA.	
positive	1			Virgin America		ItzBrianHunty		0	@VirginAmerica So excited for my first cross country f	

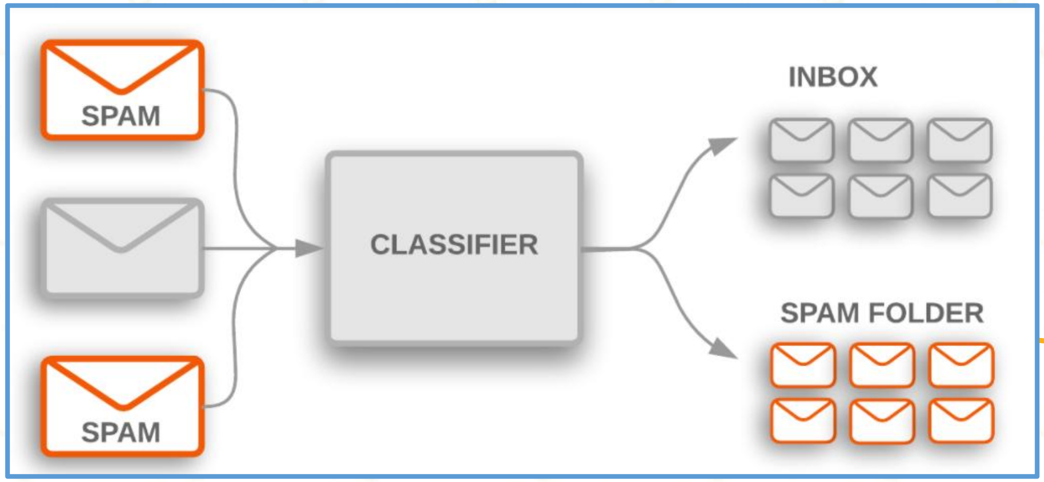
Twitter airlines sentiment

Cat and Dog dataset

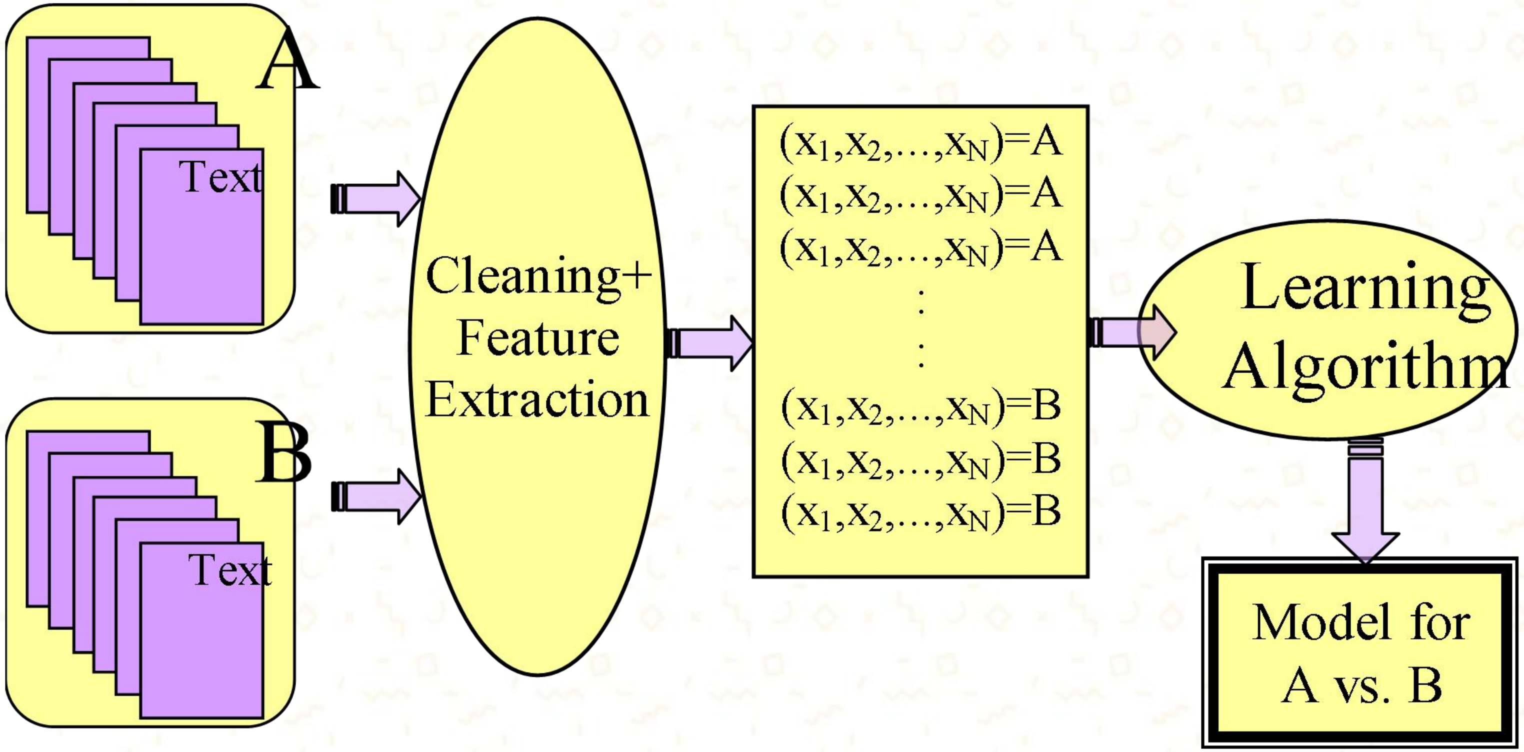
Text Classification

- *Text classification* juga dikenal sebagai *text tagging* atau *text categorization* adalah proses mengkategorikan teks ke dalam kelompok tertentu.
- *Text classification* adalah salah satu tugas dasar dalam *natural language processing (NLP)* dengan aplikasi yang luas contohnya *sentiment analysis, topic labeling, spam detection, dan intent detection*.



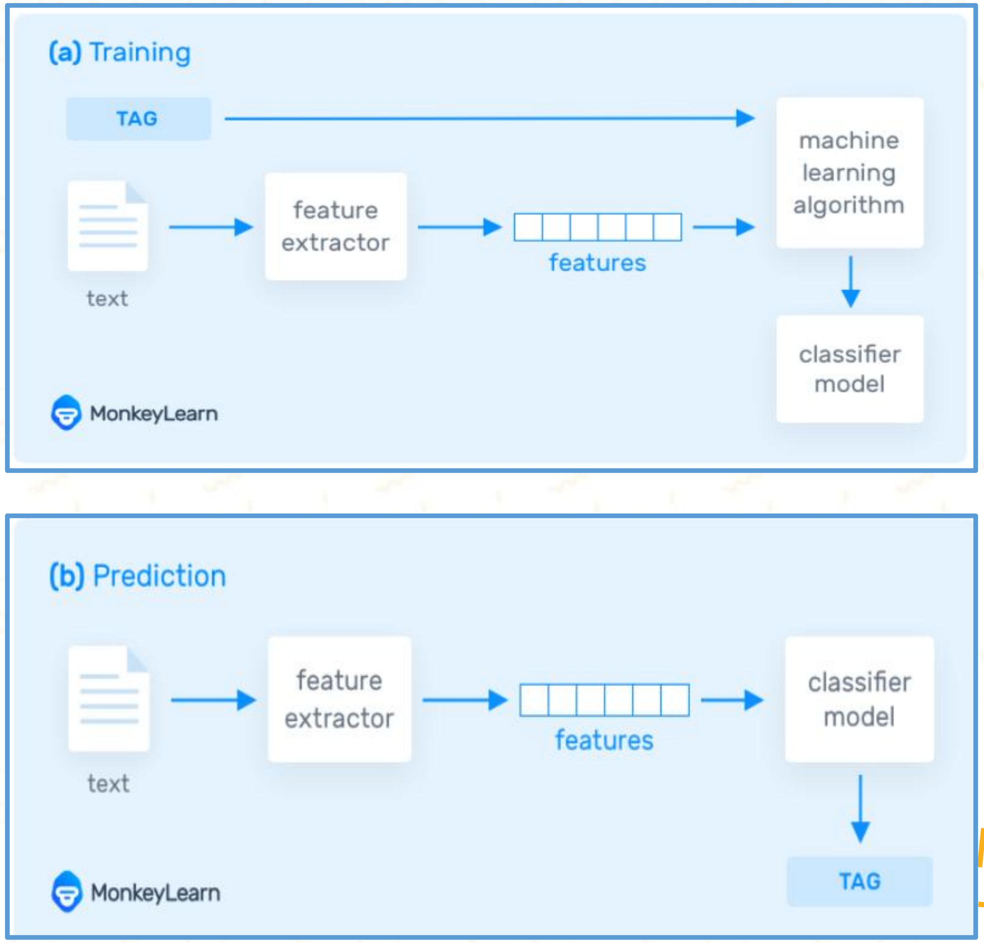


Text Classification



Text Classification

- Ada tiga pendekatan dalam *text classification*
- ***Rule-based System***
  - Teks dipisahkan ke dalam kelompok terorganisir menggunakan *handicraft linguistic rules.*
- ***Machine Learning-based System***
  - *ML-based classifier* membuat klasifikasi berdasarkan pengamatan sebelumnya dari kumpulan data
- ***Hybrid System***
  - Menggabungkan *machine learning classifier* dengan *rule-based system*, digunakan untuk meningkatkan performa.

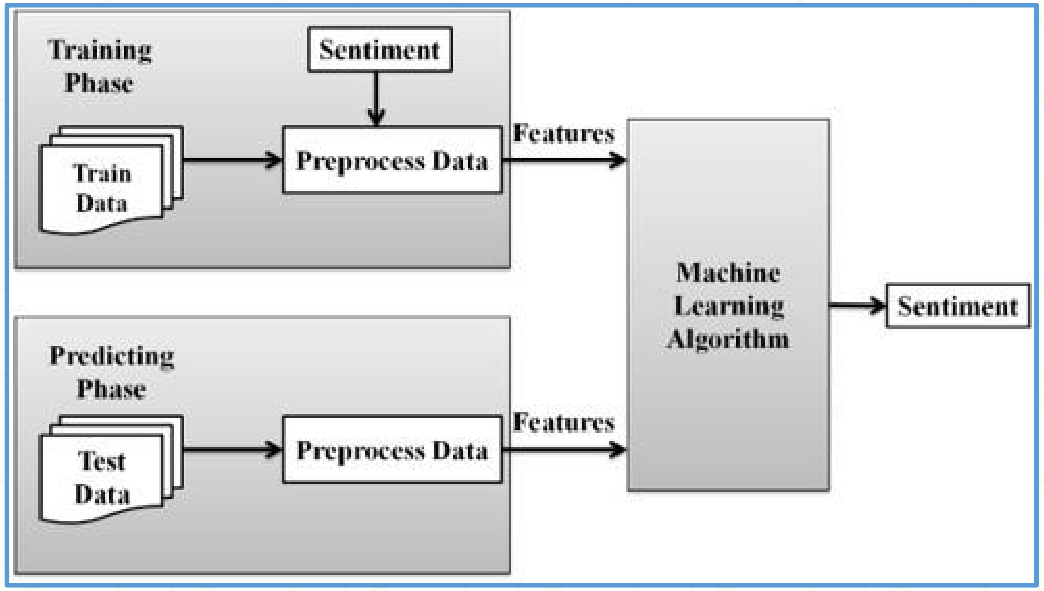


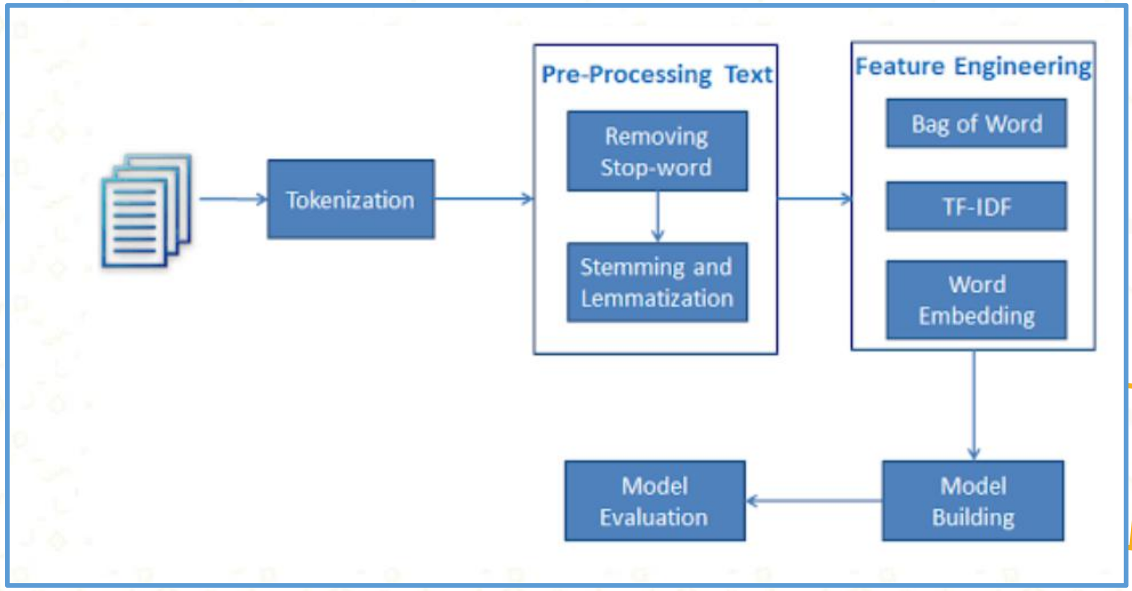
Machine Learning-based System

https://monkeylearn.com/text-classification/

Sentiment Analysis-Definition

- Salah satu contoh aplikasi dari *text classification* adalah *sentiment analysis.*
- Adalah metode yang secara otomatis memahami persepsi pelanggan terhadap suatu produk atau layanan berdasarkan komentar mereka.





Example

- *A sentiment analysis job about the problems of each major U.S. airline. Twitter data was scraped from February of 2015 and contributors were asked to first classify positive, negative, and neutral tweets,*
- Source : https://www.kaggle.com/crowdflower/twitter-airline-sentiment

airline_sent	airline_sent	negativerea	negativerea	airline	airline_sent	name	negativerea	retweet_co	text	t
neutral	1			Virgin America		cairdin		0	@VirginAmerica What @dhepburn said.	
positive	0.3486		0	Virgin America		jnardino		0	@VirginAmerica plus you've added commercials to the	
neutral	0.6837			Virgin America		yvonnalynn		0	@VirginAmerica I didn't today... Must mean I need to t	
negative	1	Bad Flight	0.7033	Virgin America		jnardino		0	@VirginAmerica it's really aggressive to blast obnoxiou	
negative	1	Can't Tell	1	Virgin America		jnardino		0	@VirginAmerica and it's a really big bad thing about it	
negative	1	Can't Tell	0.6842	Virgin America		jnardino		0	@VirginAmerica seriously would pay $30 a flight for seats that didn't have this playing. it's really the only bad thing about flying VA	
positive	0.6745		0	Virgin America		cjmcginnis		0	@VirginAmerica yes, nearly every time I fly VX this â€o	
neutral	0.634			Virgin America		pilot		0	@VirginAmerica Really missed a prime opportunity for	
positive	0.6559			Virgin America		dhepburn		0	@virginamerica Well, I didn'tâ€¦ but NOW I DO! :-D	
positive	1			Virgin America		YupitsTate		0	@VirginAmerica it was amazing, and arrived an hour ea	
neutral	0.6769		0	Virgin America		idk_but_youtube		0	@VirginAmerica did you know that suicide is the secon	
positive	1			Virgin America		HyperCamiLax		0	@VirginAmerica I <3 pretty graphics. so much bette	
positive	1			Virgin America		HyperCamiLax		0	@VirginAmerica This is such a great deal! Already think	
positive	0.6451			Virgin America		mollanderson		0	@VirginAmerica @virginmedia I'm flying your #fabulou	
positive	1			Virgin America		sjespers		0	@VirginAmerica Thanks!	
negative	0.6842	Late Flight	0.3684	Virgin America		smartwatermelon		0	@VirginAmerica SFO-PDX schedule is still MIA.	
positive	1			Virgin America		ItzBrianHunty		0	@VirginAmerica So excited for my first cross country f	

Handling text dataset

Tokenization

- *Tokenization* adalah memecah teks mentah menjadi potongan-potongan kecil *(chunks).*
- *Tokenization* memecah teks mentah menjadi kata-kata, yang disebut *tokens*.
- *Tokens* ini membantu dalam memahami konteks atau mengembangkan model untuk NLP.

```
Text
"The cat sat on the mat."
↓
Tokens
"the", "cat", "sat", "on", "the", "mat", "."
```

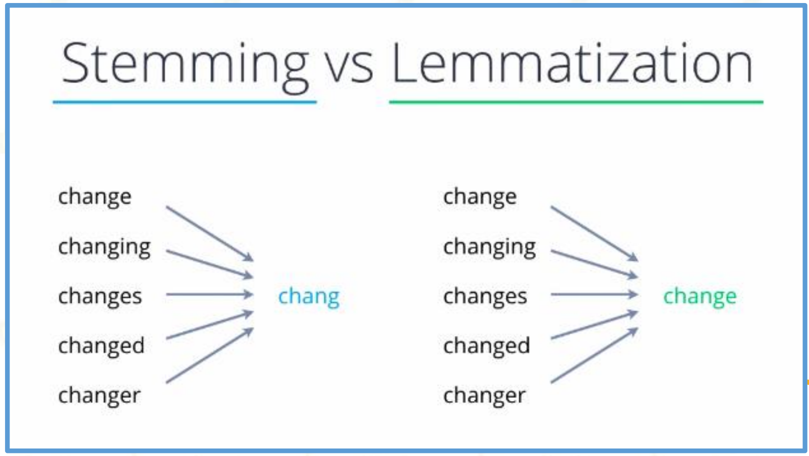
Pre-processing the Text

- *Removing stop words*
  - Tanda baca (*Punctuations)*
    - Example : .?””,’;:-[]()
  - Preposisi (*Prepositions)*
    - Example : "in," "at," "on," "of," and "to."
- *Stemming*
  - *Stemming* adalah proses mereduksi kata-kata menjadi bentuk kata dasar *(word stem),* atau akar kata *(root form)*
    - Example : walker, walked, walking => walk
- *Lemmatization*
  - *Lemmatization* adalah proses mengubah kata ke bentuk dasarnya *(base form).*
  - Mengubah kata menjadi bentuk dasar *(base form)* yang bermakna.

Stop Words

These words include:

- a
- I
- the
- in
- of
- for
- at
- to
- on
- with
- from



Feature Extraction

Bag of Words

- Frequensi istilah *(term)* dalam dokumen.
- Kita fokus pada skema pengkodean yang mewakili kata-kata, tanpa informasi tentang urutan

```
doc1 = "saya belajar pemrograman dan belajar melukis"
doc2 = "saya membantu adik saya belajar menulis"
doc3 = "ibu belajar menjahit"
```

adik	belajar	dan	ibu	melukis	membantu	menjahit	menulis	pemrograman	saya
0	2	1	0	1	0	0	0	1	1
1	1	0	0	0	1	0	1	0	2
0	1	0	1	0	0	1	0	0	0

Digital Skola
Uncover The World Of Digital Skills With Us

TF-IDF

- *TF-IDF = Term frequency–inverse document frequency,*
- *Numerical statistic* yang mencerminkan betapa pentingnya sebuah kata bagi dokumen dalam kumpulan atau *corpus*
- TF-IDF adalah skor frekuensi kata yang mencoba menonjolkan kata-kata yang lebih menarik, misalnya sering muncul dalam dokumen tetapi tidak di seluruh dokumen.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

For a term i in document j:

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

idf(t) = log [n / df(t)] + 1

sklearn formula

```
doc1 = "saya belajar pemrograman dan belajar melukis"
doc2 = "saya membantu adik saya belajar menulis"
doc3 = "ibu belajar menjahit"
```

adik	belajar	dan	ibu	melukis	membantu	menjahit	menulis	pemrograman	saya
0	2	2.0986123	0	2.0986	0	0	0	2.09861229	1.405465
2.098612	1	0	0	0	2.09861229	0	2.0986123	0	2.81093
0	1	0	2.0986	0	0	2.09861229	0	0	0

Let’s practice

DigitalSkola

Uncover The World Of Digital Skills With Us

Thank YOU