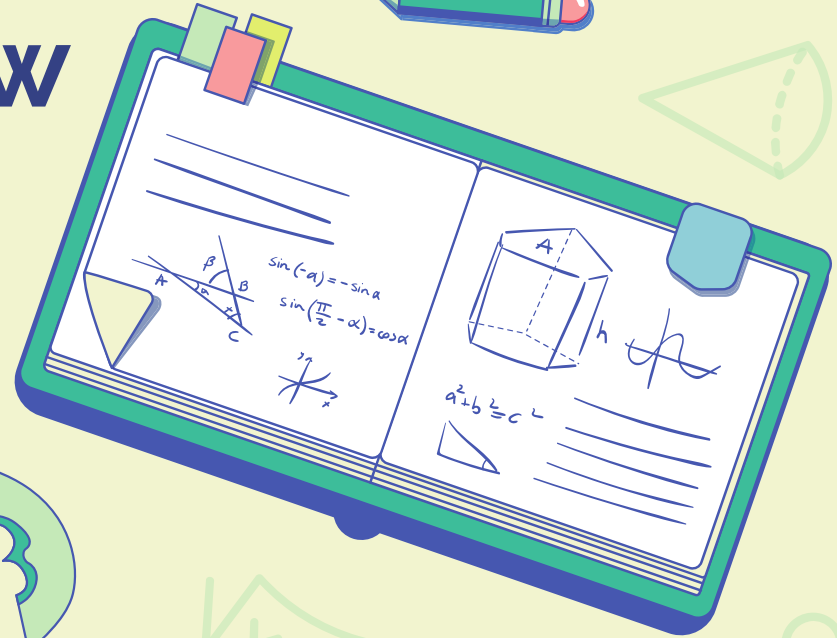


Learning Progress Review Week #10

By Optimistic team



Session 28

INTRODUCTION TO DATA MINING



DATA MINING



Apa Itu Data Mining?

Data mining adalah suatu proses penambangan informasi penting dari suatu data. Informasi penting ini didapat dari suatu proses yang amat rumit seperti menggunakan artificial intelligence, teknik statistik, ilmu matematika, machine learning, dan lain sebagainya. Teknik-teknik rumit tersebut nantinya akan mengidentifikasi dan mengekstraksi informasi yang bermanfaat dari suatu database besar.



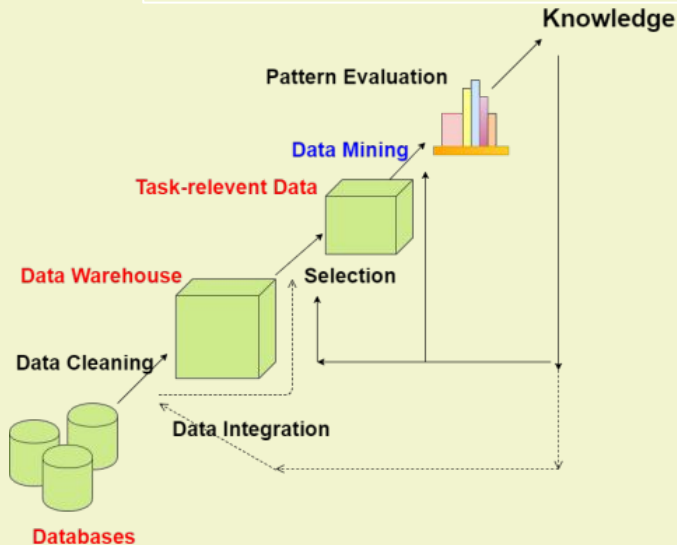
Contoh Data Mining

– Mencari nama tertentu yang lazim di wilayah/daerah tertentu.

(Contoh: O'Brien, O'Rourke, O'Reilly... nama yang lazim di wilayah Boston)

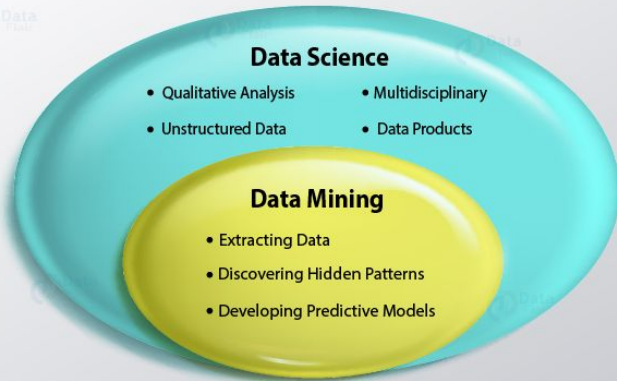
– Mengelompokkan dokumendokumen yang mirip yang dikembalikan oleh search engine berdasarkan konteksnya (misalkan Amazon rainforest, Amazon.com)

Proses Data Mining



1. Data Cleaning: menghilangkan noise dan data yang tidak konsisten. Serta mengatasi missing value.
2. Data Integration: data digabungkan dari berbagai sumber.
3. Data Selection: data yang relevan dengan proses analisis diambil dari basis data.
4. Data Transformation: data ditransformasikan dengan cara dilakukan peringkasan atau operasi agregasi.
5. Data mining: beberapa macam metode diaplikasikan untuk mengekstrak pola-pola data.
6. Pattern Evaluation: melakukan evaluasi serta interpretasi atas pola-pola menarik yang ditemukan.
7. Knowledge Presentation: mempresentasikan pengetahuan yang telah digali kepada user.

Data Science & Data Mining

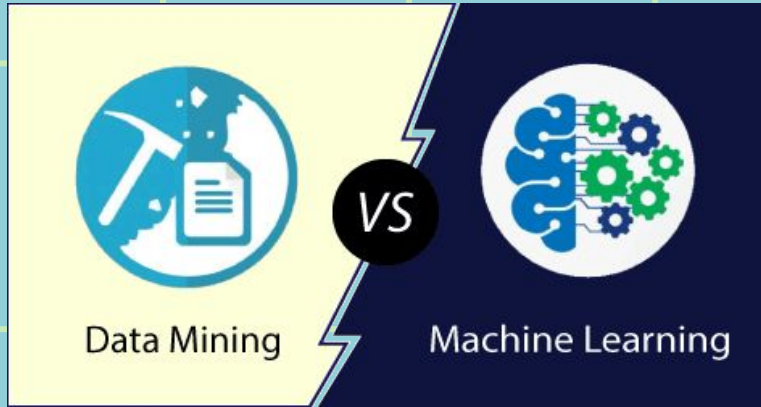


Data Science & Data Mining

Dibawah ini merupakan perbedaan utama antara data science dan data mining:

- Data mining merupakan bagian dari kegiatan dalam rangka mencari pola dibalik sekumpulan data dengan menggunakan kaidah KDD (Knowledge Discovery in Databases) sedangkan data science dengan memadukan ilmu matematika terapan dan ilmu komputer.
- Data science lebih luas dibandingkan data mining
- Data mining adalah subset dari data science

Machine learning dalam data mining lebih banyak digunakan untuk pengenalan pola, sedangkan data science memiliki penggunaan yang lebih umum.



Data mining	Machine learning
Extracting useful data from a huge quantity of statistics.	Introduce a set of rules from data in addition to from beyond experience.
Large databases with unstructured data.	Current information in addition to algorithms.
Models can be evolved for using a data mining.	System getting to know the set of rules can be used in the decision tree, neural networks and some other location of artificial intelligence.
Human interference is greater in it.	No human effort required after the layout.
Its miles used in cluster analysis.	Its miles utilized in web search, spam filter out, fraud detection, and laptop design.
facts mining summary from the records warehouse	system mastering reads machine

Ukuran Pemusatan Data

MEAN

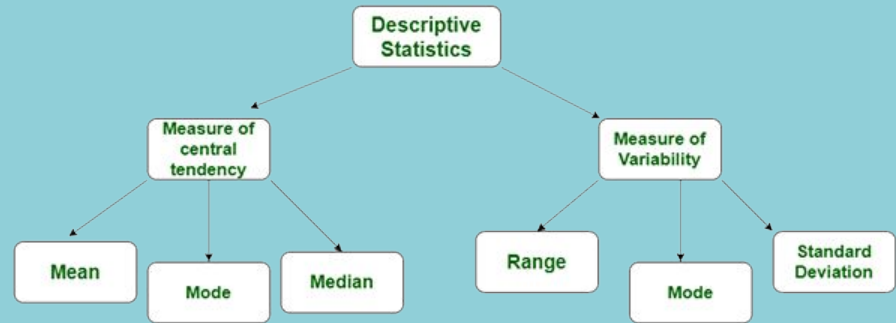
Mean atau istilah lainnya nilai rata-rata adalah jumlah keseluruhan data dibagi banyaknya data.

MODUS

Nilai yang paling banyak muncul.

MEDIAN

Median atau nilai tengah adalah pemusatan data yang membagi suatu data menjadi setengah (50%) data terkecil dan terbesarnya.



1. Descriptive

Descriptive merupakan suatu fungsi yang bertujuan memahami lebih jauh mengenai data yang diamati sehingga dapat diketahui perilaku dari sebuah data.

2. Predictive

Fungsi ini adalah sebuah fungsi yang menjelaskan suatu proses dalam menemukan pola tertentu dari sebuah data. Pola-pola yang digunakan diketahui dari beragam variabel yang terdapat pada data.

3. Classification

Fungsi ini bertujuan untuk menyimpulkan beberapa definisi karakteristik dari sebuah grup. Misalnya, pelanggan perusahaan yang sudah berpindah karena tersaingin oleh perusahaan lain.

4. Clustering

Clustering adalah identifikasi kelompok dari produk-produk atau barang-barang yang memiliki karakteristik khusus.

5. Association

Association merupakan identifikasi hubungan dari kejadian-kejadian yang sudah terjadi di suatu waktu.

6. Sequencing

Sequencing sebetulnya hampir sama dengan association tetapi untuk sequencing berfungsi untuk identifikasi hubungan-hubungan berbeda di sebuah periode waktu tertentu. Contohnya, para pelanggan yang berkunjung di supermarket secara berulang.

7. Forecasting

Fungsi ini bertujuan untuk memperkirakan nilai di suatu masa di masa mendatang sesuai dengan pola-pola dengan kumpulan data dalam jumlah besar. Contohnya, peralihan permintaan pasar.

Session 29

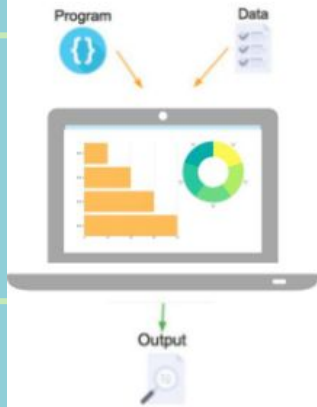
INTRODUCTION TO MACHINE LEARNING





Machine Learning adalah bagian dari kecerdasan buatan yang berkaitan dengan desain pengembangan algoritma/model yang memungkinkan komputer mengembangkan perilaku berdasarkan data empiris/data-data yang sudah ada. Machine Learning digunakan untuk mengoptimalkan kinerja menggunakan contoh data atau pengalaman masa lalu.

Traditional Programming

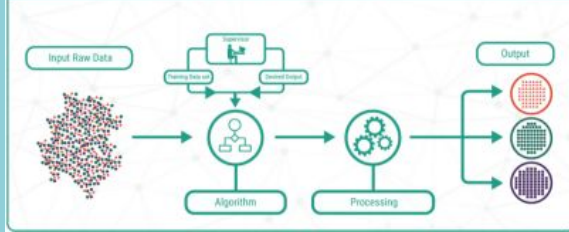


Machine Learning

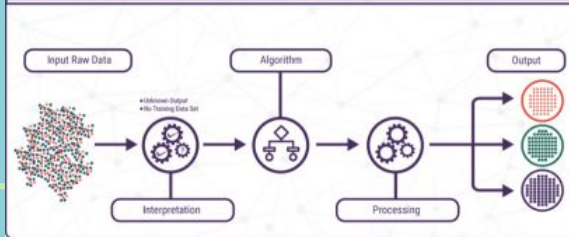


Machine Learning adalah bagian dari kecerdasan buatan yang berkaitan dengan desain pengembangan algoritma/model yang memungkinkan komputer mengembangkan perilaku berdasarkan data empiris/data-data yang sudah ada. Machine Learning digunakan untuk mengoptimalkan kinerja menggunakan contoh data atau pengalaman masa lalu.

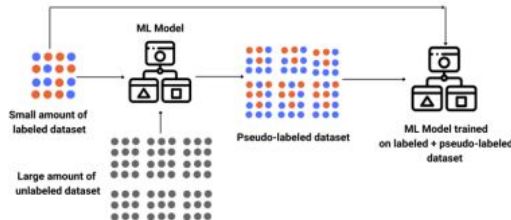
SUPERVISED LEARNING



UNSUPERVISED LEARNING



Semi-supervised learning use-case



Types of Learning

1. Supervised Learning

Algoritma pada supervised learning akan mempelajari paten dari training data/data sebelumnya, dimana data sebelumnya harus memiliki target class. Contoh kasus yang masuk dalam supervised learning ialah Classification, regression/prediction.

2. Unsupervised Learning

Unsupervised Learning adalah ketika training data yang dimiliki tidak mempunyai target class. Contoh kasus yang masuk dalam unsupervised learning ialah Clustering.

3. Semi-supervised Learning

Sebagian training data memiliki outputs.

4. Reinforcement Learning

Rewards diberikan ketika agent mengerjakan tugas tertentu.

Stage in Machine Learning

1. Data Preprocessing

Pada data preprocessing dataset akan dibersihkan karna data cenderung kotor, misalnya terdapat outliers maka pada data preprocessing akan melakukan remove outlier.

2. Train Models

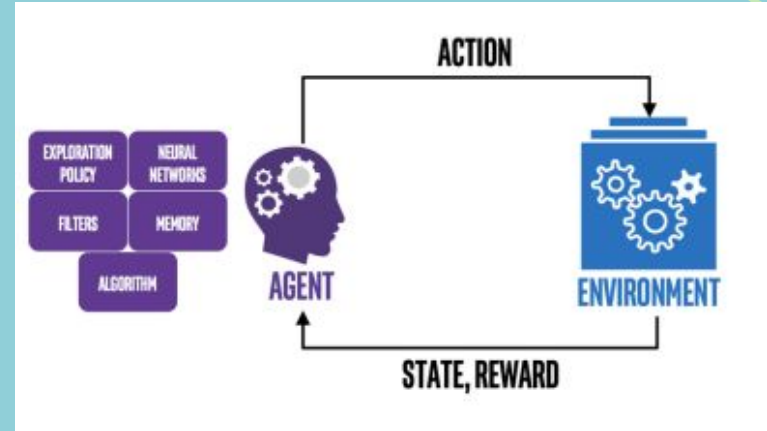
Setelah data tersebut bersih, maka selanjutnya data akan di train models. Pada stage ini akan dilakukan pemilihan algorithm yang tepat untuk digunakan.

3. Evaluate Model

Pada stage evaluate model dilakukan assess performance yaitu melakukan evaluasi terhadap algorithm yang kita gunakan.

4. Deploy Model

Apabila model sudah bagus dan memiliki nilai akurasi yang tinggi maka selanjutnya akan dilakukan deploy model, dengan tujuan model yang kita buat dapat digunakan oleh user secara luas.



Bias and Variance

1. Bias

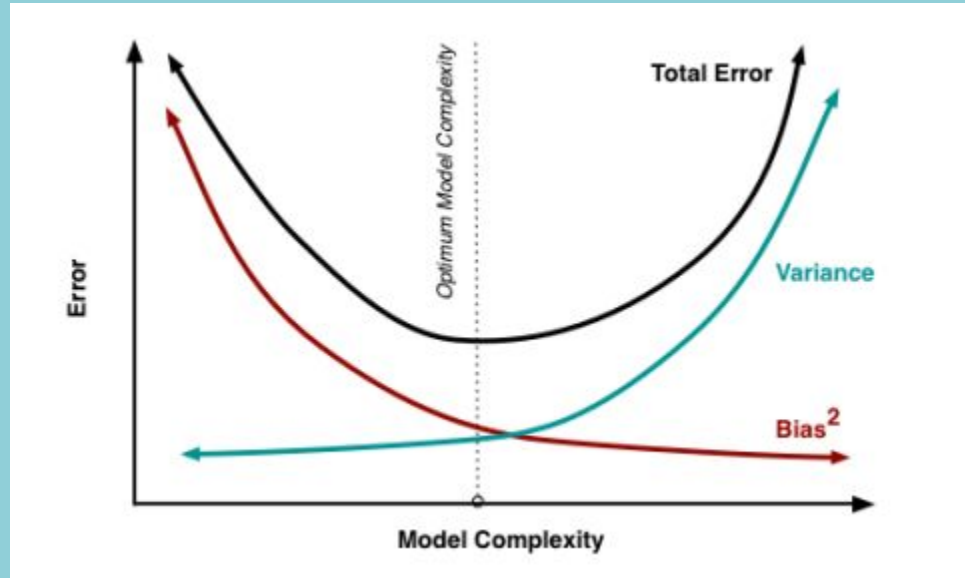
Bias adalah perbedaan antara rata-rata hasil prediksi dari model machine learning yang kita develop dengan nilai data yang sebenarnya. Bias juga dapat dikatakan sebagai error pada data training, error akan tinggi apabila model yang dilakukan terlalu sederhana.

2. Variance

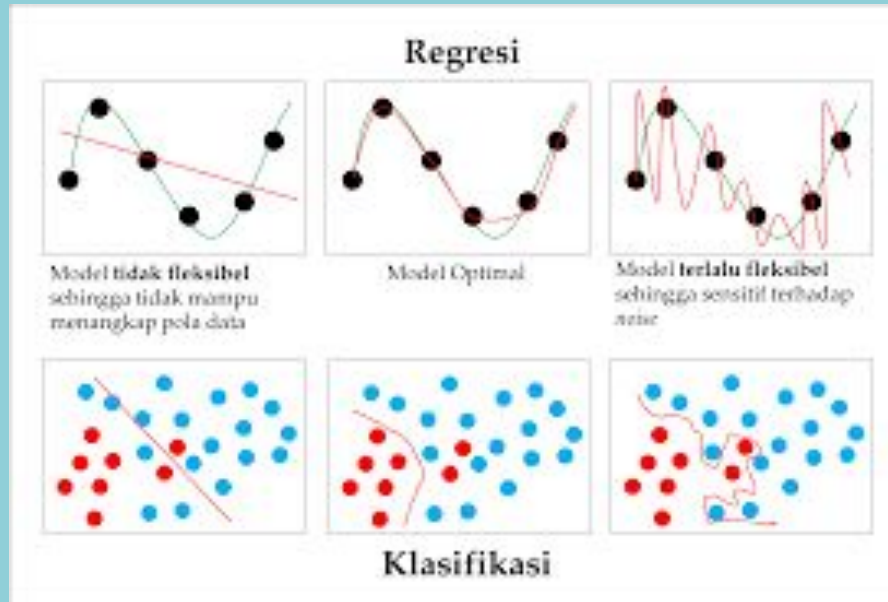
Variance adalah variable dan prediksi yang memberikan kita informasi persebaran data hasil prediksi. Bias juga dapat dikatakan sebagai error pada data testing, biasanya variance tinggi dikarenakan pada saat training set memiliki akurasi yang bagus, namun saat diaplikasikan pada testing set memiliki akurasi yang tidak bagus.



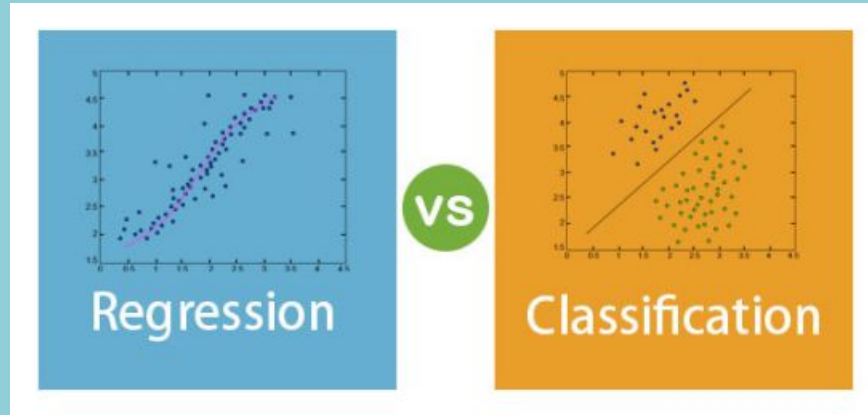
Bias and Variance tradeoff



Underfitting and overfitting



Regression vs Classification

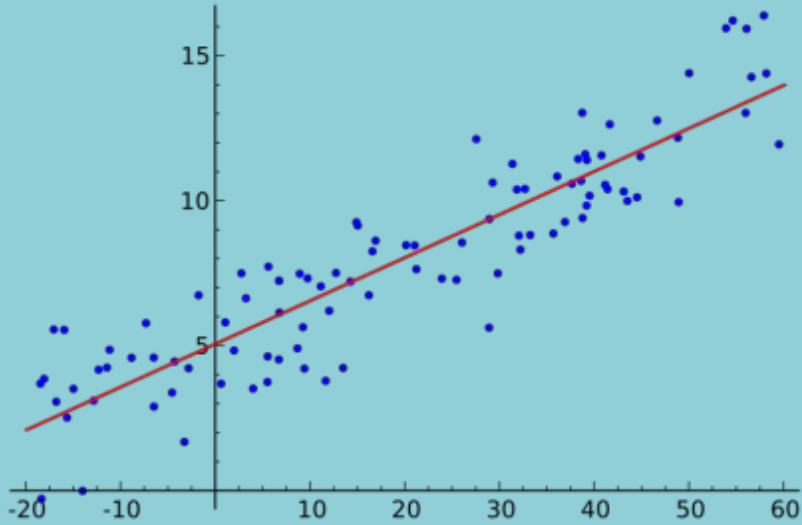


Regression

Regresi adalah metode yang mencoba untuk menentukan kekuatan dan karakter hubungan antara satu variabel dependen dan serangkaian variabel independen. Algoritma regresi ialah nilai kontinu.

Classification

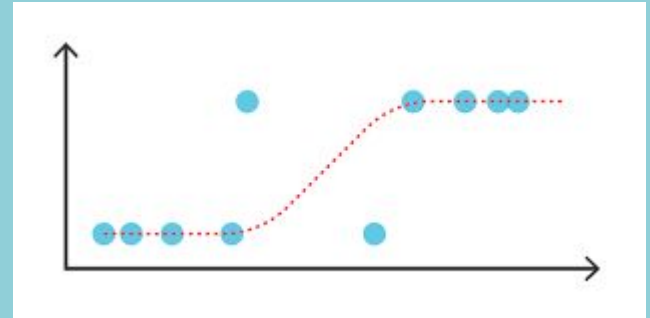
Klasifikasi adalah sebuah metode untuk mengklasifikasikan atau mengkategorikan beberapa item yang belum berlabel ke dalam sebuah set kelas diskrit.



Linear Regression

Logistic Regression

Logistic regression adalah algoritma klasifikasi machine learning yang digunakan untuk memprediksi ketika variabel dependen adalah katehoris.



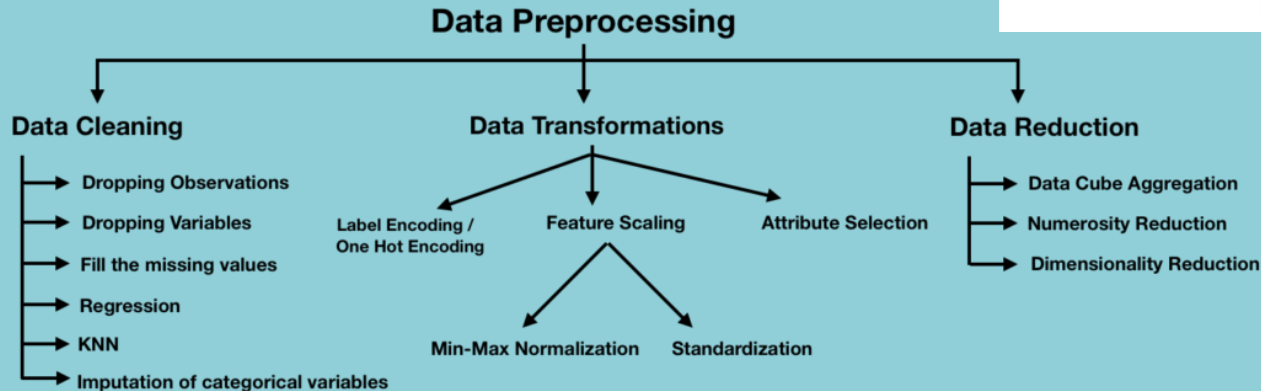
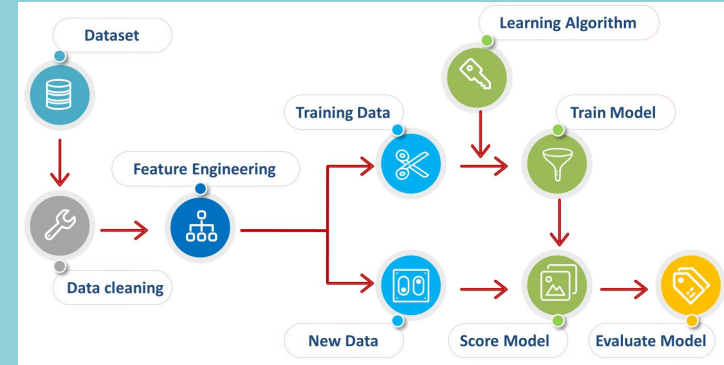
Session 30

DATA PREPROCESSING FOR MACHINE LEARNING



What is Data Preprocessing?

Data preprocessing adalah tahap menyiapkan/membersihkan data yang kotor untuk selanjutnya akan diproses menggunakan model machine learning. Dengan dilakukannya data preprocessing maka dapat meningkatkan efisiensi model dan meningkatkan performa model.



Data Preprocessing = Feature Engineering

What is Feature?

Feature adalah data dependen atau predictor yang digunakan untuk analysis.

	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr.	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, Mr.	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, M.	female	26	0	0	STON/O2. 31	7.925		S
5	4	1	1	Futrelle, Mrs.	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr. W.	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, Mr. J.	male		0	0	330877	8.4583		Q

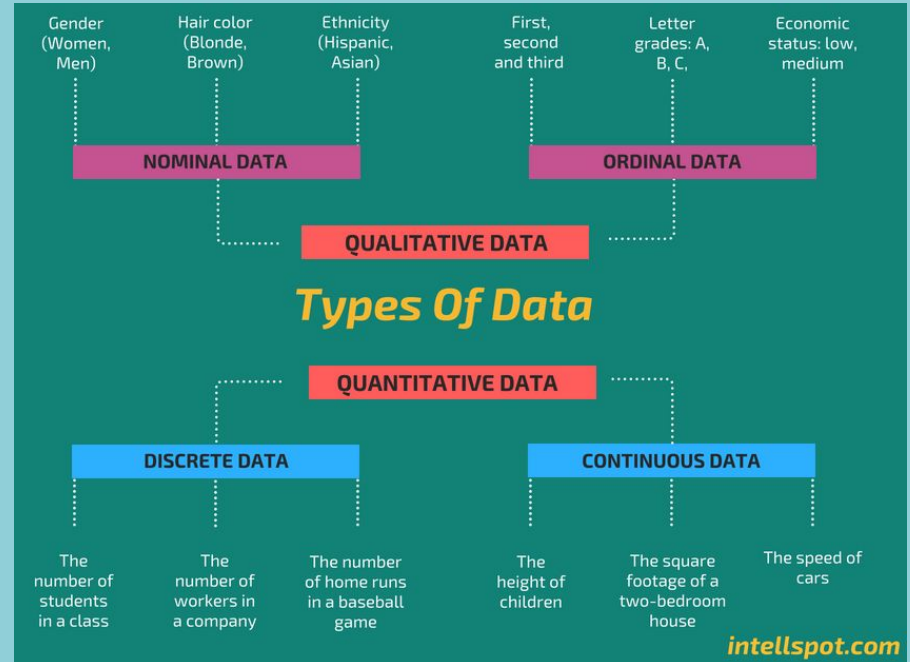
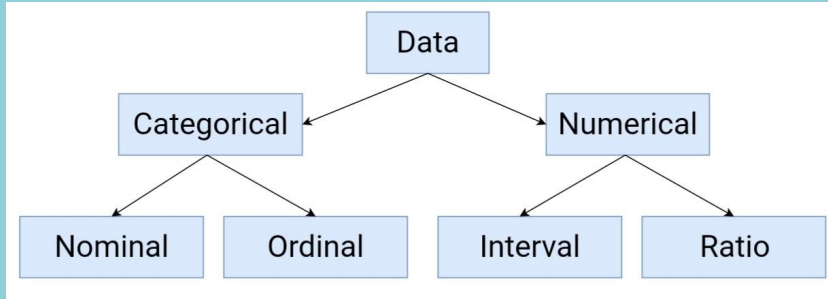
location	date_of_sale	property_size_sq_m	number of bedrooms	price	type
Clapham	12/4/1999	50	1	729000	apartment,1930s
Ashford	5/8/2017	119	3	699000	semi-detached,1970s
Stratford-on-Avon	29/3/2012	212	3	540000	detached,17th century
Canterbury	1/7/2009	95	2	529000	terraced,1960s
Camden	16/12/2001	54	1	616000	apartment,2000s
Rugby	1/3/2003	413	7	247000	detached, 19th century
Hampstead	5/3/2016	67	2	890000	terraced, 19th century

Data Type

Tipe data adalah konsep yang sangat penting dalam machine learning. Dengan mengetahui tipe data akan memudahkan dalam proses data preprocessing.


Kenapa tipe data penting:

- Untuk mengaplikasikan pengukuran statistic ke data dengan benar
- Menyimpulkan dengan benar asumsi tertentu dari dataset



Feature and Target

Seperti yang dijelaskan sebelumnya Feature adalah data dependen atau predictor yang digunakan untuk analysis. Lalu Target adalah dependen variable atau label, ini adalah data yang akan diprediksi.

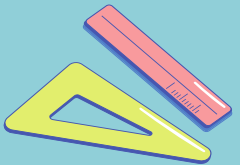
BRAND	TYPE	CYLINDER	ENG-SIZE	STROKE		PRICE	RISK
Brand-A	sedan	four	109	3.4	 $f(x)$	13950	POS
Brand-A	sedan	five	136	3.4		17450	POS
Brand-B	sedan	four	108	2.8		16430	POS
Brand-B	sedan	four	108	2.8		16925	POS
Brand-C	hatchback	three	61	3.03		5151	NEG
Brand-C	hatchback	four	90	3.11		6295	NEG
Brand-D	hatchback	four	90	3.23		5572	NEG
Brand-D	hatchback	four	90	3.23		6377	NEG

Input

Label

How Important is Data Preprocessing

Data dunia nyata cenderung incomplete, noisy, dan inconsistent. Hal ini dapat menyebabkan rendahnya kualitas data yang dikumpulkan dan selanjutnya rendahnya kualitas model yang dibangun di atas data. Untuk mengatasi masalah ini, Data Preprocessing menyediakan operasi yang dapat mengatur data ke dalam bentuk yang tepat untuk pemahaman yang lebih baik. Kita tidak dapat memahami perilaku atau tren data. Oleh karena itu, kita perlu mengubah atau mengaturnya agar menjadi format yang tepat dengan menggunakan Data Preprocessing. Dengan dilakukannya data preprocessing maka dapat meningkatkan efisiensi model dan meningkatkan performa model.





DATA CLEANSING

Missing Value, adalah data yang tidak lengkap diperlihatkan dengan Na di dataframe.

Untuk mengetahui apakah ada missing value dalam dataset dapat menggunakan :

`Dataframe.isnull().sum()`

```
# Checking missing value for each feature
print('Checking missingg value for each feature:')
print(dataset.isnull().any(),'\n')
print(dataset.isnull().sum(),'\n')

# Counting total missing value
print('\nCounting total missing value:')
print(dataset.isnull().sum().sum())
```

executed in 29ms, finished 11:15:27 2021-10-30

Checking missingg value for each feature:

Administrative	True
Administrative_Duration	True
Informational	True
Informational_Duration	True
ProductRelated	True
ProductRelated_Duration	True
BounceRates	True
ExitRates	True
PageValues	False
SpecialDay	False
Month	False
OperatingSystems	False
Browser	False
Region	False
TrafficType	False
VisitorType	False
Weekend	False
Revenue	False
dtype:	bool

Administrative	14
Administrative_Duration	14
Informational	14
Informational_Duration	14
ProductRelated	14
ProductRelated_Duration	14
BounceRates	14
ExitRates	14
PageValues	0
SpecialDay	0
Month	0
OperatingSystems	0
Browser	0
Region	0
TrafficType	0
VisitorType	0
Weekend	0
Revenue	0
dtype:	int64

Counting total missing value:
112

Lalu untuk handling menggunakan Imputation

```
print('\nAfter Imputation:')  
# Fill missing value with mean of feature value  
dataset.fillna(dataset.mean(), inplace=True)  
# Checking missing value for each feature  
print(dataset.isnull().sum())  
#print(dataset.isna().sum())  
# Counting total missing value  
print(dataset.isna().sum().sum())
```


Hasilnya:

```
After Imputation:  
Administrative      0  
Administrative_Duration  0  
Informational      0  
Informational_Duration  0  
ProductRelated     0  
ProductRelated_Duration  0  
BounceRates        0  
ExitRates          0  
PageValues         0  
SpecialDay         0  
Month              0  
OperatingSystems   0  
Browser            0  
Region             0  
TrafficType        0  
VisitorType        0  
Weekend            0  
Revenue            0  
dtype: int64  
0
```

Data Encoding

Dalam data kategorikal yang bertipe object tidak dapat diproses ke dalam model. Maka perlu dilakukan Encoding untuk merubah nilainya menjadi numeric.

id	color			
1	red			
2	blue			
3	green			
4	blue			



id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Contoh:

```
from sklearn.preprocessing import LabelEncoder
import numpy as np

# Convert feature/column 'Month'
LE = LabelEncoder()
dataset['Month'] = LE.fit_transform(dataset['Month'])
print(LE.classes_)
print(np.sort(dataset['Month'].unique()))
print('')

# Convert feature/column 'VisitorType'
dataset['VisitorType'] = LE.fit_transform(dataset['VisitorType'])
print(LE.classes_)
print(np.sort(dataset['VisitorType'].unique()))
```

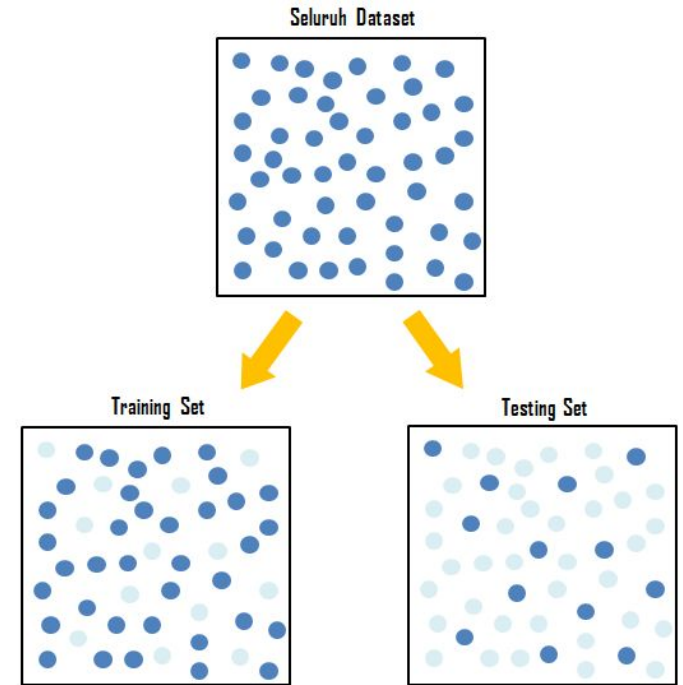
executed in 29ms, finished 11:15:32 2021-10-30

```
['Aug' 'Dec' 'Feb' 'Jul' 'June' 'Mar' 'May' 'Nov' 'Oct' 'Sep']
[0 1 2 3 4 5 6 7 8 9]

['New_Visitor' 'Other' 'Returning_Visitor']
[0 1 2]
```

Train Test Split

Train/test split adalah salah satu metode yang dapat digunakan untuk mengevaluasi performa model machine learning. Metode evaluasi model ini membagi dataset menjadi dua bagian yakni bagian yang digunakan untuk training data dan untuk testing data dengan proporsi tertentu.



Contoh:

```
from sklearn.model_selection import train_test_split
# splitting the X, and y
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
# checking the shapes
print("Shape of X_train :", X_train.shape)
print("Shape of y_train :", y_train.shape)
print("Shape of X_test :", X_test.shape)
print("Shape of y_test :", y_test.shape)
```

executed in 73ms, finished 11:15:32 2021-10-30

```
Shape of X_train : (9864, 17)
Shape of y_train : (9864,)
Shape of X_test : (2466, 17)
Shape of y_test : (2466,)
```

Feature Scaling

Kenapa harus melakukan feature scaling:

- Data dengan skala yang sama akan menjamin algoritma pembelajaran memperlakukan semua feature dengan adil
- Data dengan skala yang sama dan centered akan mempercepat algoritma pembelajaran
- Data dengan skala yang sama akan mempermudah interpretasi beberapa model ML

Kapan menggunakan feature scaling:

- Gunakan feature scaling jika model ML yang digunakan terpengaruhi oleh skala data (KNN, Logistic Regression, SVM)
- Gunakan Standardization bila tahu bahwa data memiliki sebaran normal/Gaussian
- Gunakan Standardization bila model yang kita pakai punya asumsi tentang normalitas (e.g. regresi linear)
- Gunakan normalization apabila tidak memenuhi 2 kriteria di atas.



Contoh Feature Scaling:

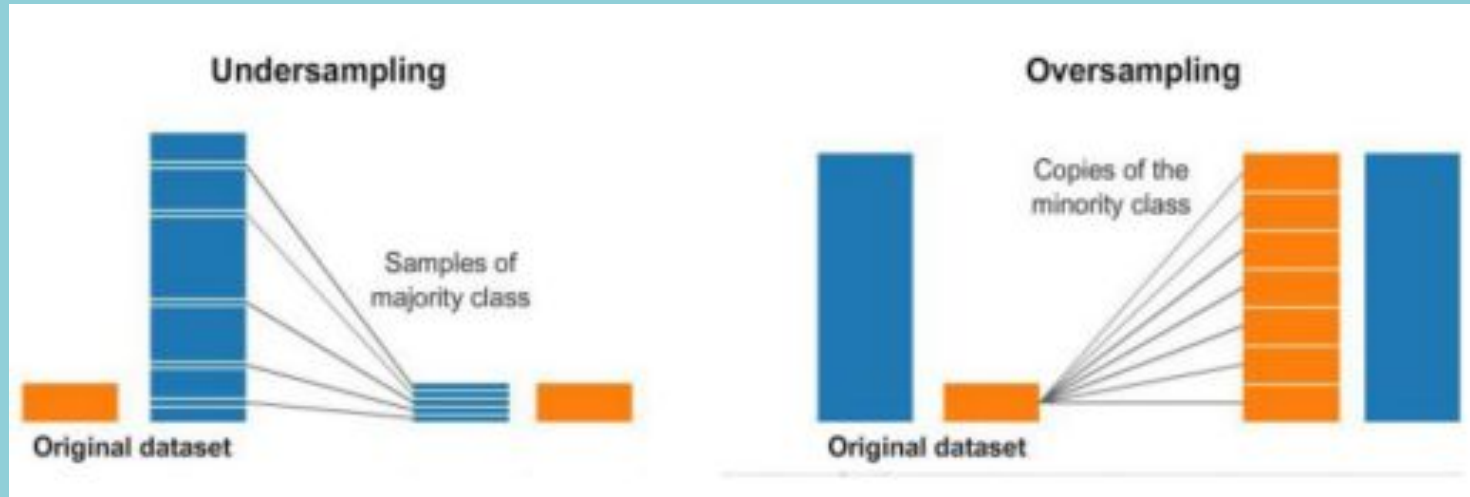
```
# Data Rescaling
from sklearn import preprocessing
data_scaler = preprocessing.MinMaxScaler(feature_range=(0,1))
housing[['RM', 'LSTAT', 'PTRATIO', 'MEDV']] = data_scaler.fit_transform(housing[['RM', 'LSTAT', 'PTRATIO', 'MEDV']])
```

executed in 28ms, finished 11:15:34 2021-10-30



Imbalanced

Data yang memiliki rasio yang tidak berimbang antara data satu dengan data lainnya dapat dikatakan sebagai imbalanced. Dengan begitu dataset harus dibuat balance dengan hanya menggunakan **Training Dataset**.





Melakukan Resampling

1. Over-Sampling

Melakukan generate pada rare class sehingga jumlah dari rare class sama dengan jumlah abundant class.

2. Under-Sampling

Melakukan seleksi pada abundant class secara acak/ random sehingga abundant class nilainya berkurang sampai dengan jumlahnya sama dengan rare class.

Thank you!

Our Team

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