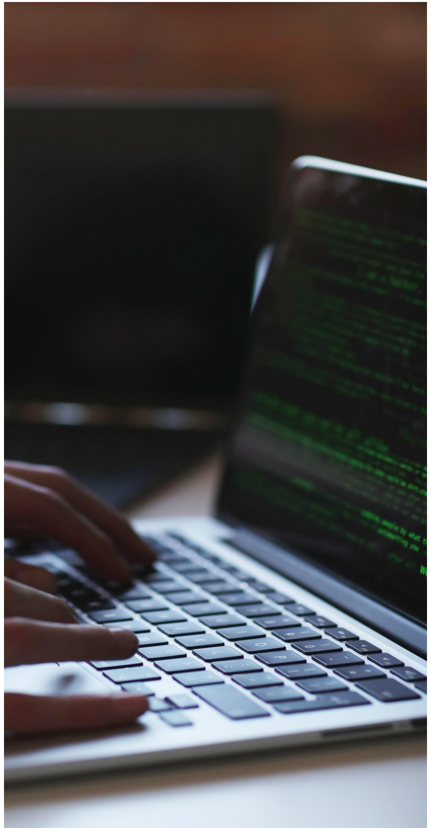
Course-Net
Empowering IT People

Introduction to Data Cleansing



APA YANG AKAN KITA PELAJARI

Data Preprocessing Method

- What is Data Preprocessing
- Data Preprocessing Intuition
- Data Preprocessing Methods
- Data Preprocessing Use Cases

Course-Net
Empowering IT People

EmpoweringITPeople

Apa itu Data Preprocessing Mengapa Kita Pelajari di Data Science ?

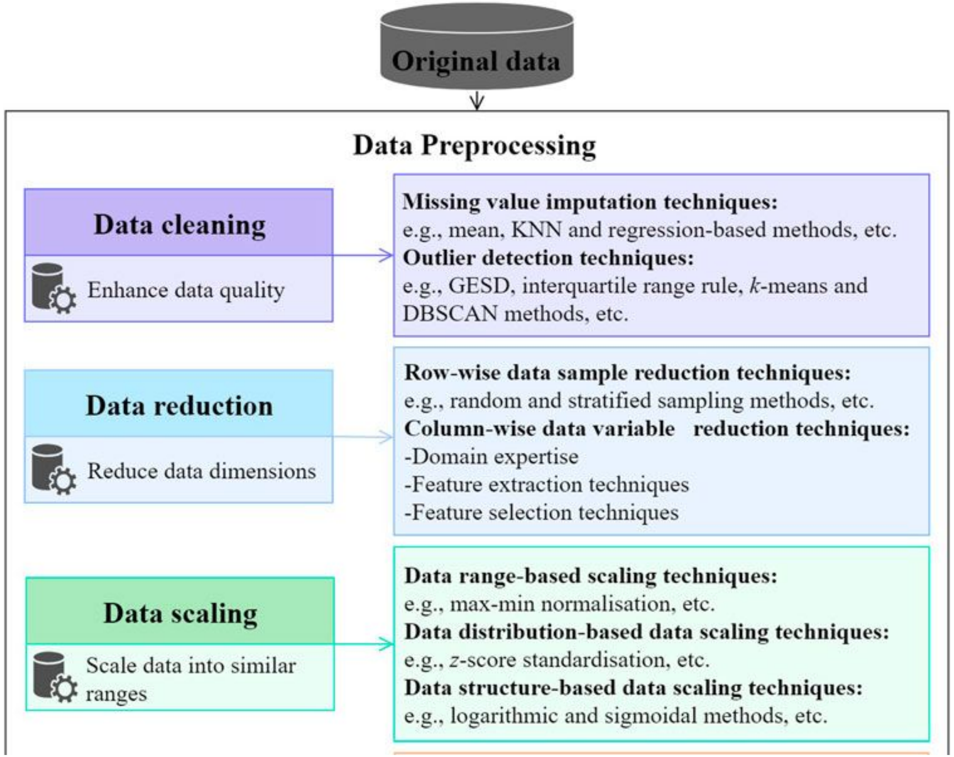
Course-Net Empowering IT People

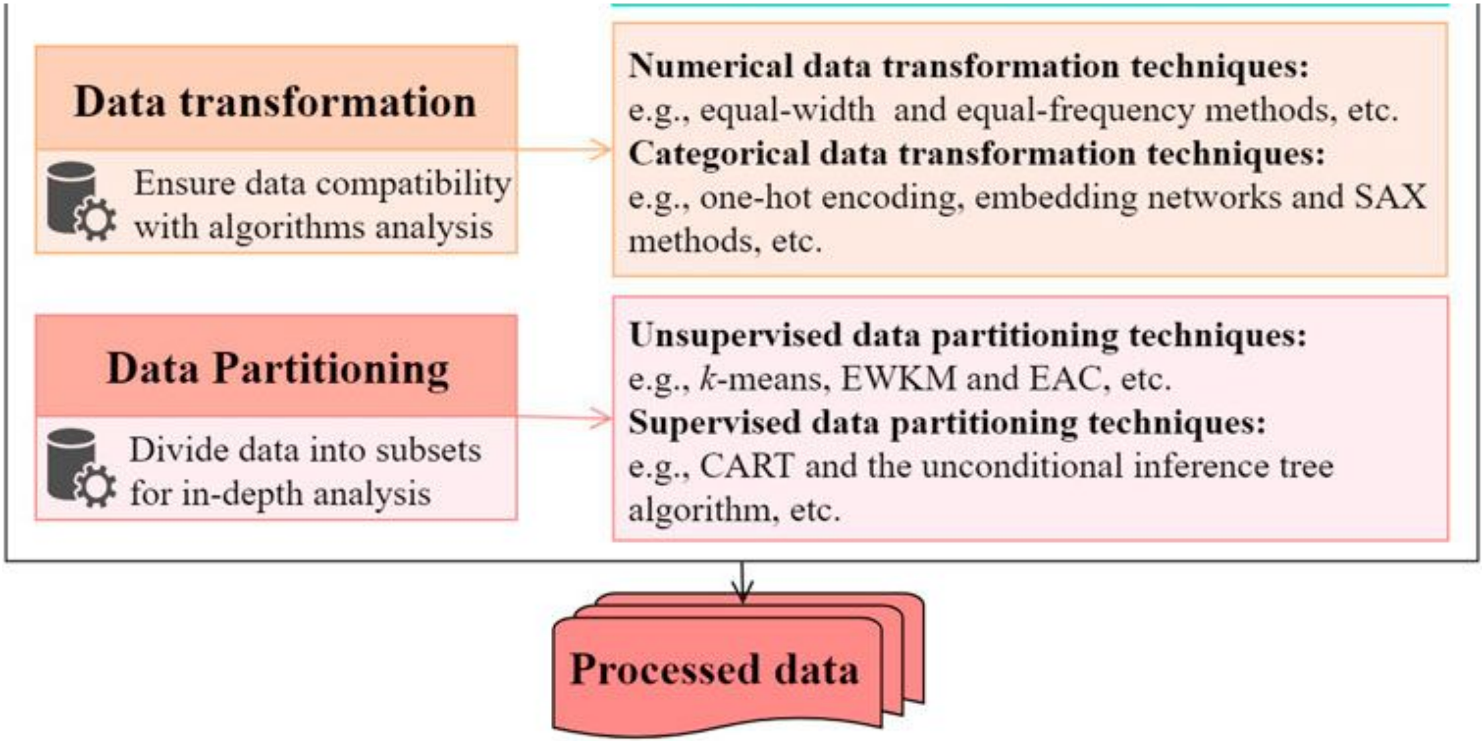
EmpoweringITPeople

Data Preprocessing: Bring data up to a level of quality such that it can reliably be used for the production of statistical models or statements

Course-Net
Empowering IT People

EmpoweringITPeople




Data transformation
Ensure data compatibility with algorithms analysis
Numerical data transformation techniques:
e.g., equal-width and equal-frequency methods, etc.
Categorical data transformation techniques:
e.g., one-hot encoding, embedding networks and SAX methods, etc.
Data Partitioning
Divide data into subsets for in-depth analysis
Unsupervised data partitioning techniques:
e.g., k-means, EWKM and EAC, etc.
Supervised data partitioning techniques:
e.g., CART and the unconditional inference tree algorithm, etc.
Processed data

Course-Net Empowering IT People

EmpoweringITPeople

Missing Data

User	Device	OS	Transactions
A	Mobile	Android	5
B	Mobile	Android	3
C	NA	iOS	2
D	Tablet	Android	1
E	Mobile	iOS	4
F	NA	Android	2
G	Tablet	Android	4

User	Device	OS	Transactions
A	Mobile	Android	5
B	Mobile	Android	3
C	Tablet	iOS	1
D	Tablet	Android	1
E	Mobile	iOS	4
F	Mobile	Android	2
G	Tablet	Android	4

Device	OS		Avg. Transactions
	#Android	#iOS	
Mobile	2	1	4
Tablet	2	0	2.5
Missing	2		

Device	OS		Avg. Transactions
	#Android	#iOS	
Mobile	3	1	3.5
Tablet	2	1	2

Methods :

1. Discard the record
2. Value Imputations

Imputations

Methods :

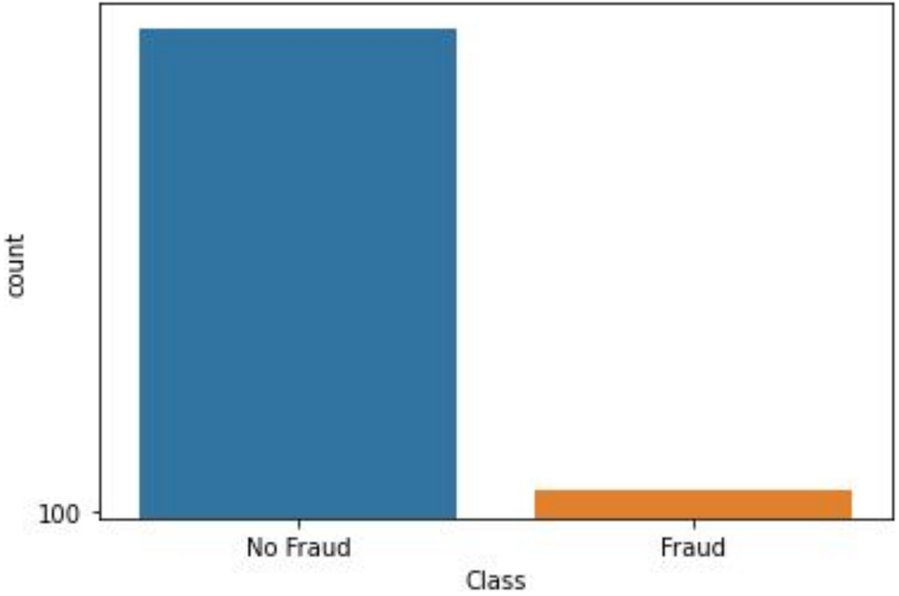
Numerical Data

1. Mean (Sensitive to outliers)
2. Median (Unsensitive to outliers)

Discrete Data

1. Mode

Class Imbalance

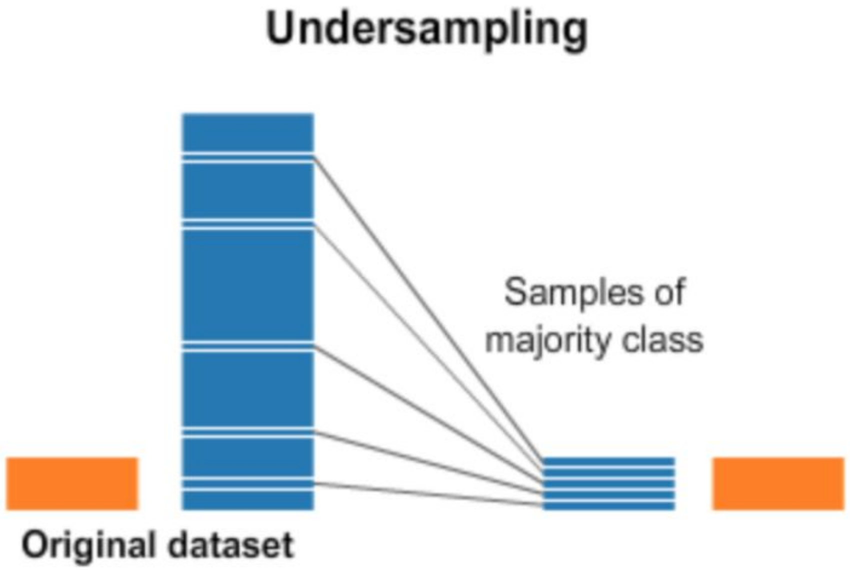


Example: Credit Card Fraud Datasets

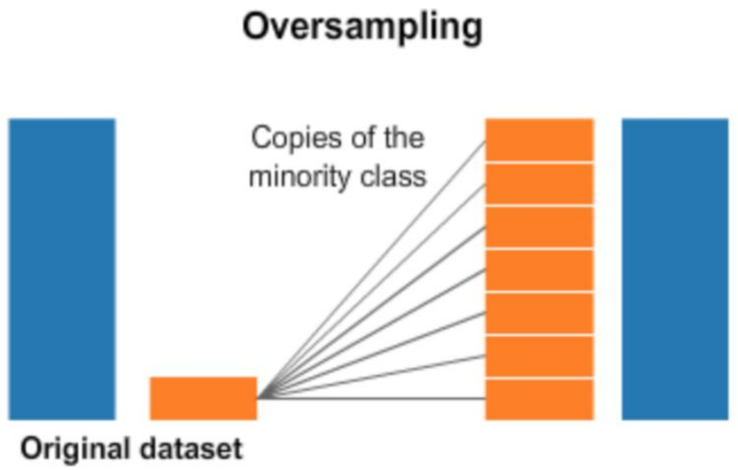
Methods :

1. Undersampling
2. Oversampling

Undersampling



Oversampling



Methods :

1. **SMOTE**
(Synthetic Minority Over-sampling TEchnique)

1. **ADASYN**
(Adaptive Synthetic Sampling Method for Imbalanced Data)

Normalization and Scaling

Normalization : Adjust the values of numeric data to a common scale without changing the range where as scaling shrinks or stretches the data to fit within a specific range.

Scaling : Useful to compare two different variables on equal grounds especially with variables which use distance measures

Normalization

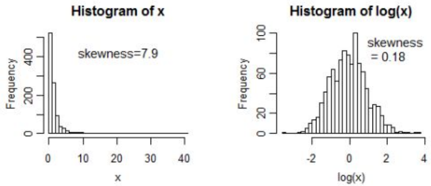
Normalization is good to use when you know that the distribution of your data does not follow a **Gaussian distribution**. This can be **useful in algorithms that do not assume any distribution of the data** like K-Nearest Neighbors and Neural Networks.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Scaling

Scaling is good to use when you know that the distribution of your data does follow a **Gaussian distribution.** This can be **useful in algorithms that do assume any distribution of the data** like Linear Regression, Logistic Regression, Linear Discriminant Analysis

Methods to Transforms Data:



Log Transformation

$$z = \frac{x - min(x)}{[\max(x) - min(x)]}$$

Min-Max Scaling

$$x' = \frac{x - \bar{x}}{\sigma}$$

Z-Score Normalization

$$x' = \frac{x - \text{mean}(x)}{\max(x) - \min(x)}$$

Mean Normalization

$$||\vec{v}|| = \sqrt{v_x * v_x + v_y * v_y}$$

Unit Vector Transformation

$$(4.3) \quad y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & if\ \lambda \neq 0, \\ \ln(y_i) & if\ \lambda = 0, \end{cases}$$

Box Cox Transformation

Course-Net Empowering IT People #EmpoweringITPeople

Categorical Encoding

Typically, any structured dataset includes multiple columns – a combination of numerical as well as categorical variables. A machine can only understand the numbers. It cannot understand the text. That's essentially the case with Machine Learning algorithms too.

Categorical encoding is a process of converting categories to numbers.

Label Encoding

In this technique, each label is assigned a unique integer based on alphabetical ordering.

Country	Age	Salary
India	44	72000
US	34	65000
Japan	46	98000
US	35	45000
Japan	23	34000

Before

Country	Age	Salary
0	44	72000
2	34	65000
1	46	98000
2	35	45000
1	23	34000

After

One Hot Encoding

In this technique, It simply creates additional features based on the number of unique values in the categorical feature. Every unique value in the category will be added as a feature.

Country	Age	Salary
India	44	72000
US	34	65000
Japan	46	98000
US	35	45000
Japan	23	34000

0	1	2	Age	Salary
1	0	0	44	72000
0	0	1	34	65000
0	1	0	46	98000
0	0	1	35	45000
0	1	0	23	34000

Before

After

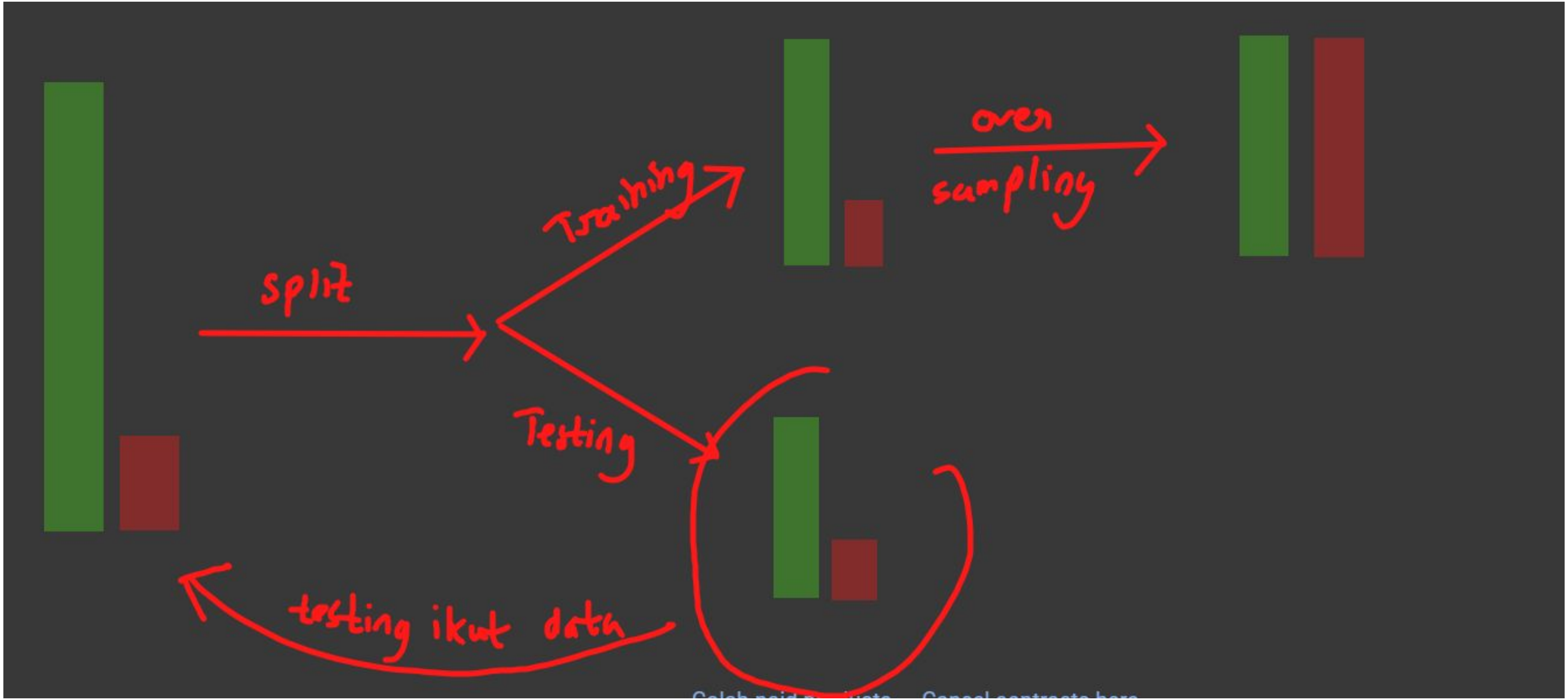
Label Encoding vs One Hot Encoding

Use Label Encoding when:

1. The categorical feature is ordinal
2. The number of categories is quite large as one hot encoding can lead to high memory consumption

Use One Hot Encoding when:

1. the categorical feature is not ordinal
2. The number of categories can be applied effectively


split
Training
Testing
over
sampling
testing ikut data

Course-Net Empowering IT People

EmpoweringITPeople

References