Telkom University

Fakultas Rekayasa Industri
School of Industrial Engineering
Telkom University

The Methods in Data Mining

Information Systems Undergraduate Program
School of Industrial Engineering

PLO & CLO to be achieved:
PLO01 – Able to analyze complex INFOKOM issues, defining and modeling requirements in the context of enterprises or society by applying knowledge in the fields of computing, information and communication technology, and other relevant disciplines

CLO07 – Students are able to apply fundamental statistical knowledge in the scope of information systems science .

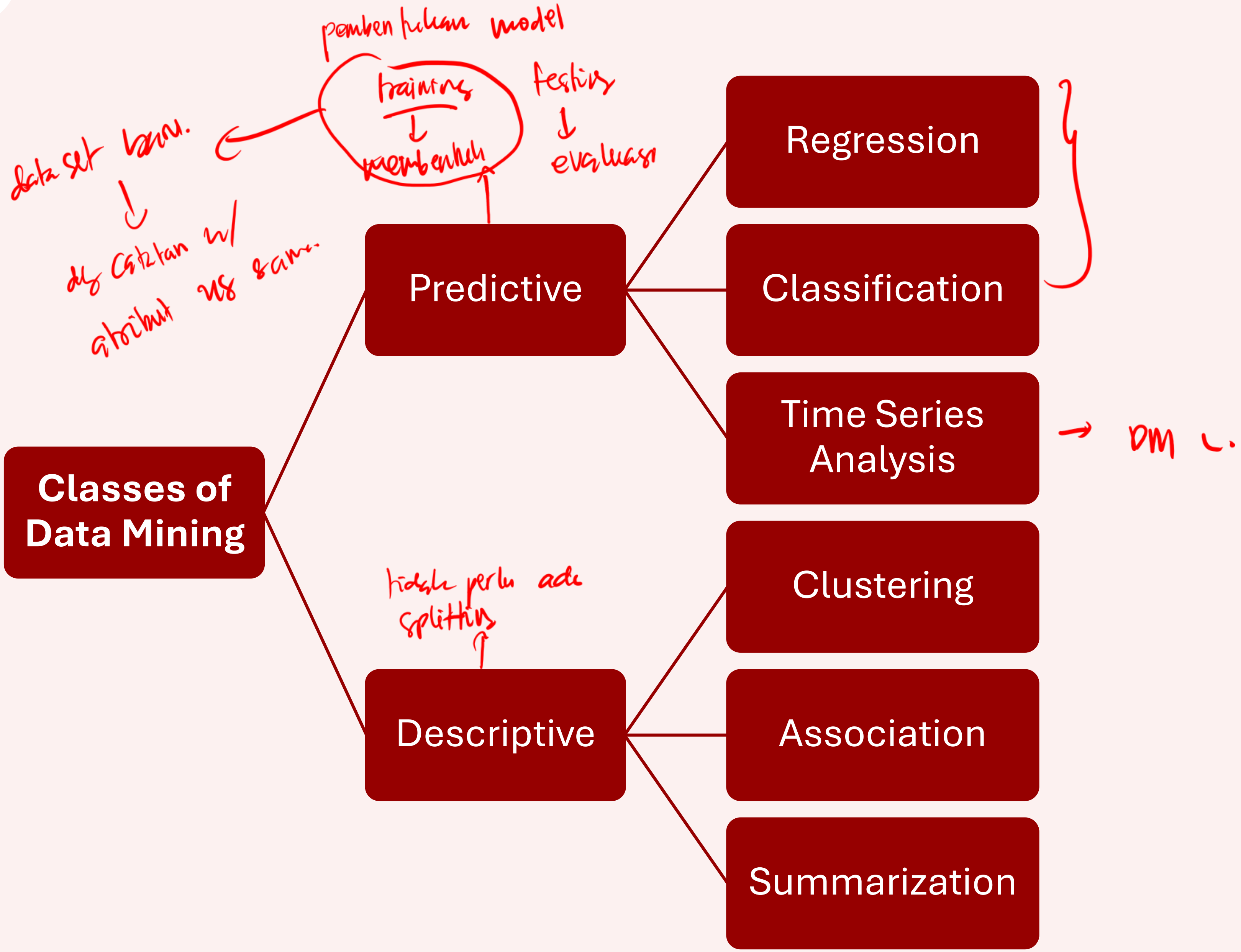
Outline:

1. **Clustering**
2. **Regression**
3. **Classification**

Reference:

1. James, Gareth et al. 2023. An Introduction Statistical Machine Learning With Applications in Python. Springer
2. Walpole, Ronald E., Myers, Raymond H., Myers, Sharon L. 2013. Probability & Statistics for Engineers & Scientists. 9th ed.. Pearson Education. United States of America
3. Han, Jiawei, Jian Pei, and Hanghang Tong. 2023. Data Mining Concepts and Techniques. 4th ed. ed. Beth LoGiudice. Cambridge: Katey Birtcher.
4. Moreira, J. M., Carvalho, A. C. P. L. F., & Horváth, T. (2018). A General Introduction to Data Analytics. In A General Introduction to Data Analytics. https://doi.org/10.1002/9781119296294

Introduction to the Methods in Data Mining



Classes of Data Mining

- Predictive
  - Regression
  - Classification
  - Time Series Analysis
- Descriptive
  - Clustering
  - Association
  - Summarization

pembentukan model

training
↓
pembentukan

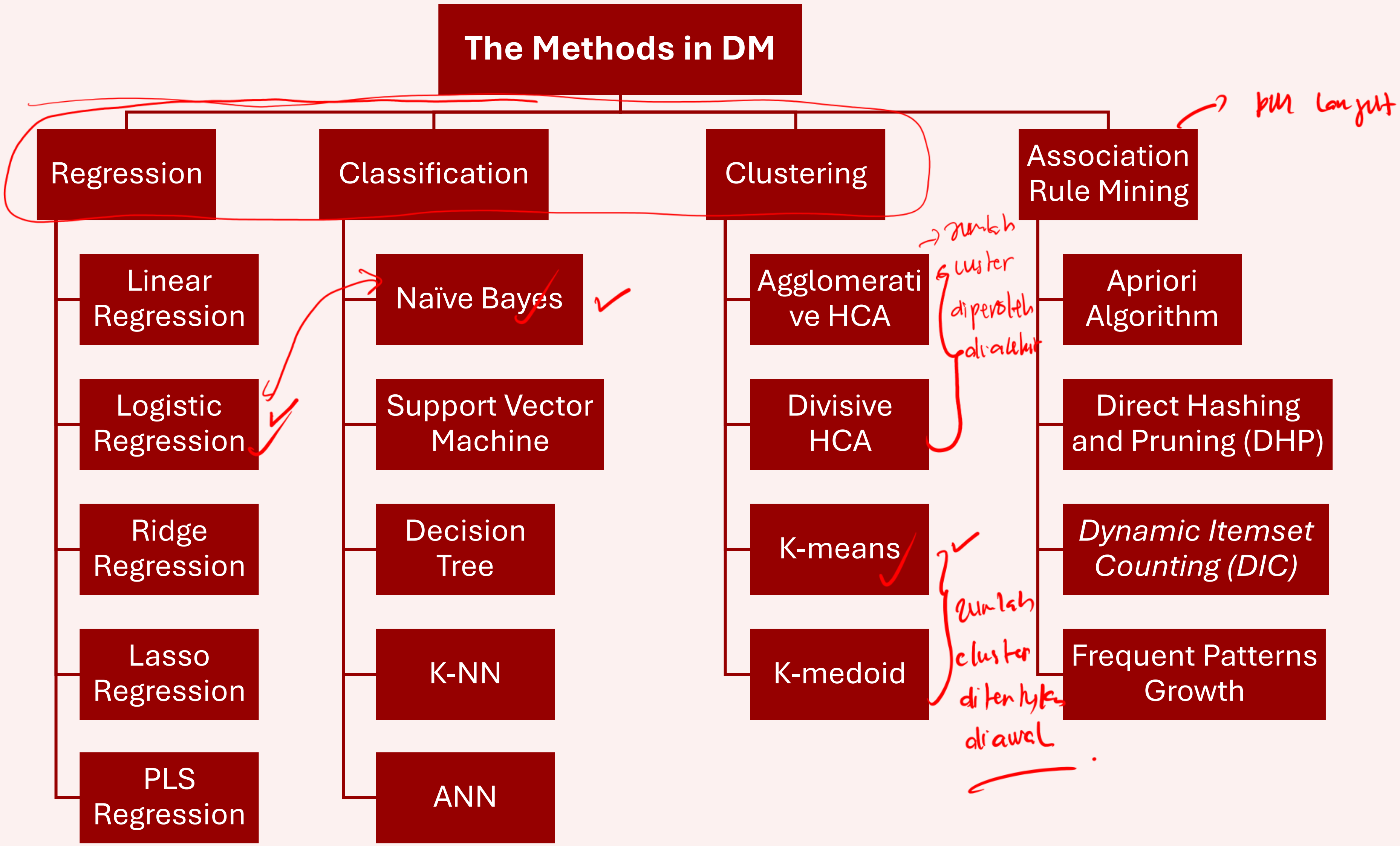
testing
↓
evaluasi

data set baru.
↓
dg catatan w/ atribut yg sama..

tidak perlu ada splitting

→ DM L.

The Methods in DM



- Regression
  - Linear Regression
  - Logistic Regression ✓
  - Ridge Regression
  - Lasso Regression
  - PLS Regression
- Classification
  - Naïve Bayes ✓
  - Support Vector Machine
  - Decision Tree
  - K-NN
  - ANN
- Clustering
  - Agglomerative HCA
  - Divisive HCA
  - K-means ✓
  - K-medoid
- Association Rule Mining
  - Apriori Algorithm
  - Direct Hashing and Pruning (DHP)
  - *Dynamic Itemset Counting (DIC)*
  - Frequent Patterns Growth

→ DM lanjut

→ jumlah cluster diperoleh diakhir

✓ jumlah cluster ditentukan diawal

Clustering

→ Clustering   Kelompok 1 ✓   Kelompok 2 ✓   Kelompok 3 ✓

- Unsupervised learning is a type of machine learning where the algorithm is given data without any explicit instructions or labeled examples.
- Its goal is to identify patterns, structures, or relationships in the data independently.
- Unsupervised learning deals with unlabeled data, meaning the algorithm must work without prior guidance.
- Clustering is one of algorithms/ methods in unsupervised learning
- ✓ Clustering is the process of dividing a set of data objects into subsets called clusters, where objects in the same cluster are similar and different from those in other clusters.
- The goal of clustering is to organize data into meaningful subgroups so that data points within the same cluster are more similar than those in other clusters.
- Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity.

Requirement in Clustering

- **Scalability**
  Many clustering algorithms excel with small datasets but struggle with large databases containing millions or billions of objects, as seen in web search. Sampling from these extensive datasets can lead to biased results, highlighting the need for scalable clustering algorithms.
- **Ability to deal with different types of attributes**
  Many algorithms are designed for clustering numeric data, but there is a growing need for techniques that can manage mixed data types, such as binary, nominal, ordinal, and numerical data, along with various object types like text, graphs, sequences, images, and videos.
- **Discovery of clusters with arbitrary shape**
  Many clustering algorithms use Euclidean or Manhattan distance measures, which typically identify spherical clusters of similar size and density. However, clusters can take on various shapes.

Requirement in Clustering

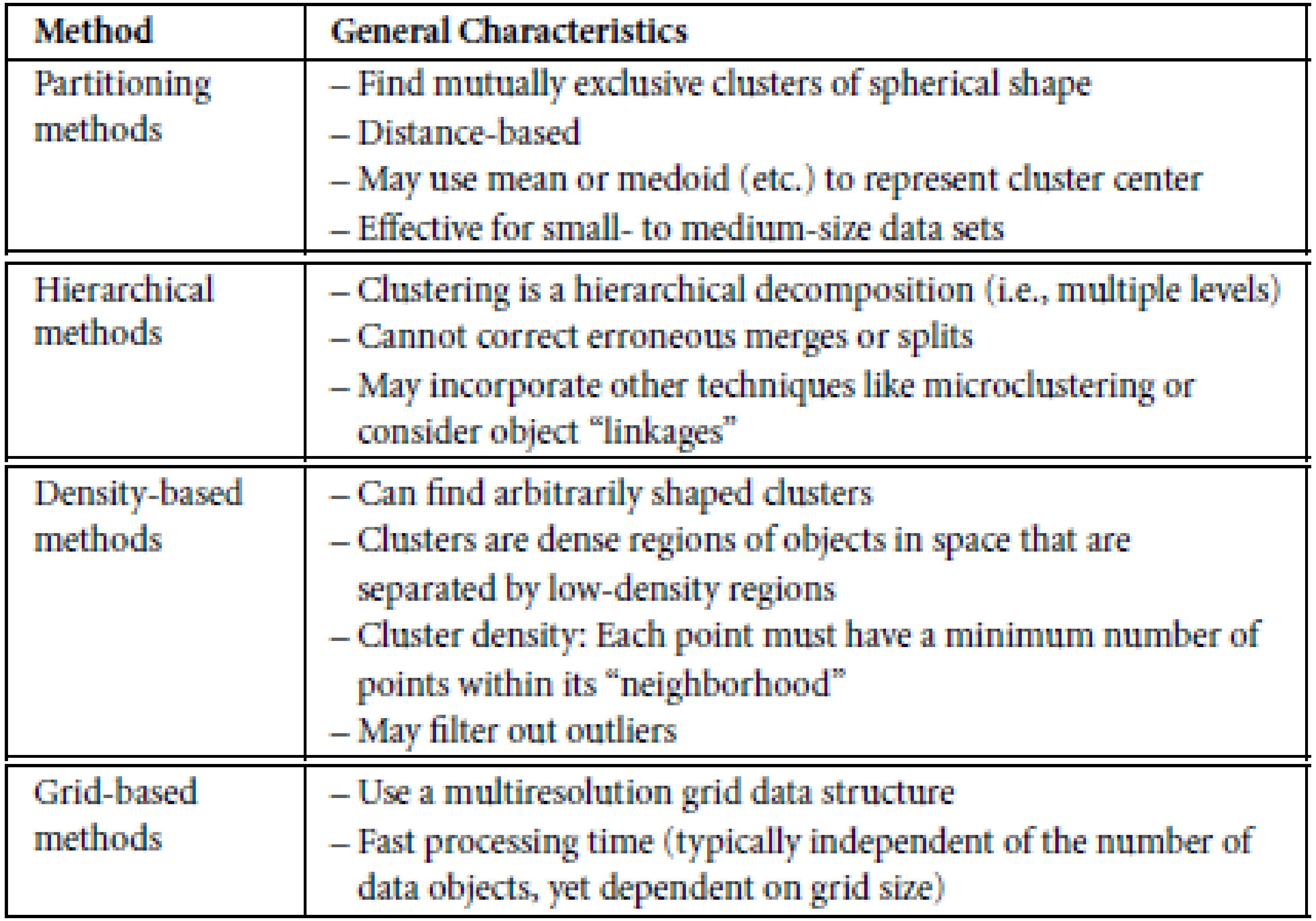
- **Requirements for domain knowledge to determine input parameters**
  Many clustering algorithms need users to specify parameters, such as the number of clusters, based on domain knowledge. This makes results sensitive to these parameters, which are difficult to define, especially in high-dimensional datasets. This requirement increases user burden and makes it harder to ensure clustering quality.
- **Ability to deal with noisy data**
  Most real-world datasets have outliers and may contain missing, unknown, or erroneous data. Clustering algorithms often struggle with this noise, resulting in poor-quality clusters, highlighting the need for robust methods that can effectively handle noise.
- **Incremental clustering and insensitivity to input order**
  In many applications, incremental updates can occur at any time, but some clustering algorithms cannot incorporate these updates and must recompute clusters from scratch. Moreover, the order of input data can significantly impact results, resulting in different clustering based on the presentation order.

Requirement in Clustering

- **Capability of clustering high-dimensionality data**
  Datasets can have numerous dimensions or attributes. In document clustering, each keyword is a dimension, often resulting in thousands of dimensions. Most clustering algorithms handle low-dimensional data well, but finding clusters in high-dimensional spaces is challenging due to data sparsity and skewness.
- **Constraint-based clustering**
  Real-world clustering often involves constraints. The challenge is to find data groups that exhibit good clustering behavior while meeting these constraints.
- **Interpretability and usability**
  Clustering results should be interpretable, comprehensible, and usable, often requiring alignment with specific semantic interpretations and applications.

Clustering Methods

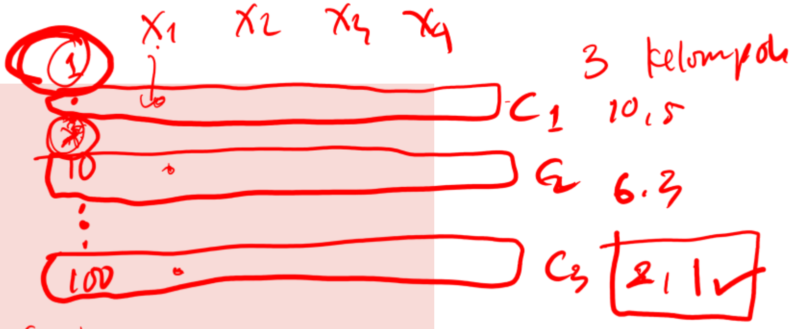
This section just described the partitioning method, especially k-means clustering



Method	General Characteristics
Partitioning methods	– Find mutually exclusive clusters of spherical shape – Distance-based – May use mean or medoid (etc.) to represent cluster center – Effective for small- to medium-size data sets
Hierarchical methods	– Clustering is a hierarchical decomposition (i.e., multiple levels) – Cannot correct erroneous merges or splits – May incorporate other techniques like microclustering or consider object "linkages"
Density-based methods	– Can find arbitrarily shaped clusters – Clusters are dense regions of objects in space that are separated by low-density regions – Cluster density: Each point must have a minimum number of points within its "neighborhood" – May filter out outliers
Grid-based methods	– Use a multiresolution grid data structure – Fast processing time (typically independent of the number of data objects, yet dependent on grid size)

The Distance in Clustering

3 kelompok / 3 cluster.



① n = 2 → data ke-2

dr setiap x → 4

$x_{1.2}$; $x_{2.2}$ ...

$x_{4.1}$

② → n = 10 →

1. **Euclidean Distance** ✓

   Formula:

   pusat / centroid

$$d(p,q) = \sum_{i=1}^{n} (p_i - q_i)^2$$ ✓ → jarak (+)

1000

   Use Cases: Suitable for continuous numerical data and is the default distance metric for K-means.

2. **Manhattan Distance** ✓

   Formula:

$$d(p,q) = \sum_{i=1}^{n} |p_i - q_i|$$ → (+)

   Use Cases: Useful for high-dimensional spaces and when you want to measure distance in a grid-like path.

Where:

p_i is point one of the observation

q_i is point two of the observation

n is the number of the observation

C1 →

Application of Clustering

- **Market Segmentation**
  Businesses use clustering to segment customers based on purchasing behavior, demographics, or preferences.
- **Image Segmentation**
  Clustering techniques are applied in image processing to segment images into distinct regions based on color or texture, aiding in object recognition and classification.
- **Anomaly Detection**
  Clustering helps identify outliers in data, which is useful in fraud detection, network security, and monitoring for unusual patterns in various fields.
- **Social Network Analysis**
  Clustering identifies communities or groups within social networks based on user interactions, providing insights into social dynamics and user behavior.
- **Telecommunications**
  Clustering analyzes call data records to identify patterns of usage or detect fraudulent activities.

K-Means Clustering

$\boxed{100} \rightarrow$ k yg bisa terbentuk
≤ 100

- K-means clustering is a simple and elegant approach for partitioning n observation of a dataset into k distinct, non-overlapping clusters, where $k \leq n$.
- K-means clustering is an effective and widely used algorithm for grouping similar data points into clusters.
- The algorithm groups data points based on their similarity, aiming to minimize the variance within each cluster while maximizing the variance between clusters.

$k_1 \rightarrow k_2$
$\vdots$

The Characteristics

- **Centroid-Based Clustering**
  Each cluster is represented by a centroid, which is the mean of all data points assigned to that cluster.
- **Sensitivity to Initialization**
  Results can vary based on the initial placement of centroids, potentially leading to different clustering outcomes.
- **Sensitivity to Outliers**
  K-means is sensitive to outliers, which can skew the position of centroids and affect the quality of the clusters.
- **Iterative Process**
  K-means follows an iterative process consisting of two main steps: assignment and update.

$S \rightarrow -1$ sd 1

← jumlah c tidak cocok → jumlah c cocok.

How K-Means Works

$k=3$

$k=2$

$C_1 \rightarrow n=2$ $C_2 = n_4$

1. **Initialization**
   - Select the number of clusters, k, which is predetermined by the user.
   - Randomly initialize k centroids from the dataset. These centroids will act as the starting points for each cluster.

1, 2, ..., 505

		x_1	x_2	x_3								
C_2 ←	1	4	5	7	3	3	2	1	5	3	2	2
C_1	2	6	8	5	4	5	6	7	6	4	2	7

2. **Assignment Step**
   - For each data point, calculate its distance to each centroid (typically using Euclidean distance).
   - Assign each data point to the cluster of the nearest centroid. This initial grouping creates preliminary clusters.

n_1 thd C_2

$$d_2 = \sqrt{1+1+0} = \boxed{\sqrt{2}}$$

n_1 thd C_1

$$d_1 = \sqrt{(4-5)^2 + (5-8)^2 + (7-5)^2} = \sqrt{4+9+4} = \boxed{\sqrt{17}}$$

3. **Update Step**
   - For each cluster, update the centroid by calculating the mean (average) position of all the data points within that cluster.
   - This updated centroid represents the new "center" of the cluster.
4. **Iterative Process**
   - Repeat assignment and update step until convergence. The convergence criteria is centroids no longer move significantly between iterations, meaning that data points stay in the same clusters.

The Advantages and Disadvantages

Advantages:

1. **Efficiency and Speed:** K-means is computationally efficient and has a linear time complexity, making it suitable for large datasets.
2. **Simplicity:** The algorithm is straightforward and easy to implement, widely understood, and accessible.
3. **Scalability:** K-means performs well with high-dimensional data and can easily scale to large datasets.

Disadvantages:

1. **Sensitivity to Initialization:** Poor initialization of centroids can lead to suboptimal clustering results.
2. **Sensitivity to Outliers:** Outliers and noise can distort the cluster centroids, affecting accuracy.
3. **Limited to Numeric Data**: It's best suited for numerical data and requires adaptations for categorical data.

Regression

- General Linear Model (GLM) is a statistical model that used to describe the relationship between a dependent variable and one or more independent variables.
- The basic form of the GLM is:

$$\enclose{circle}{Y} = \underline{X\beta + \overset{\checkmark}{\varepsilon}}$$

- Like the GLM, regression describes the relationship between a response/ dependent and predictor/ independent variables
- So, we can say that regression is a part of the GLM.

Logistic Regression

A / B

yes / No

multinomial → 3
tentu bisa.

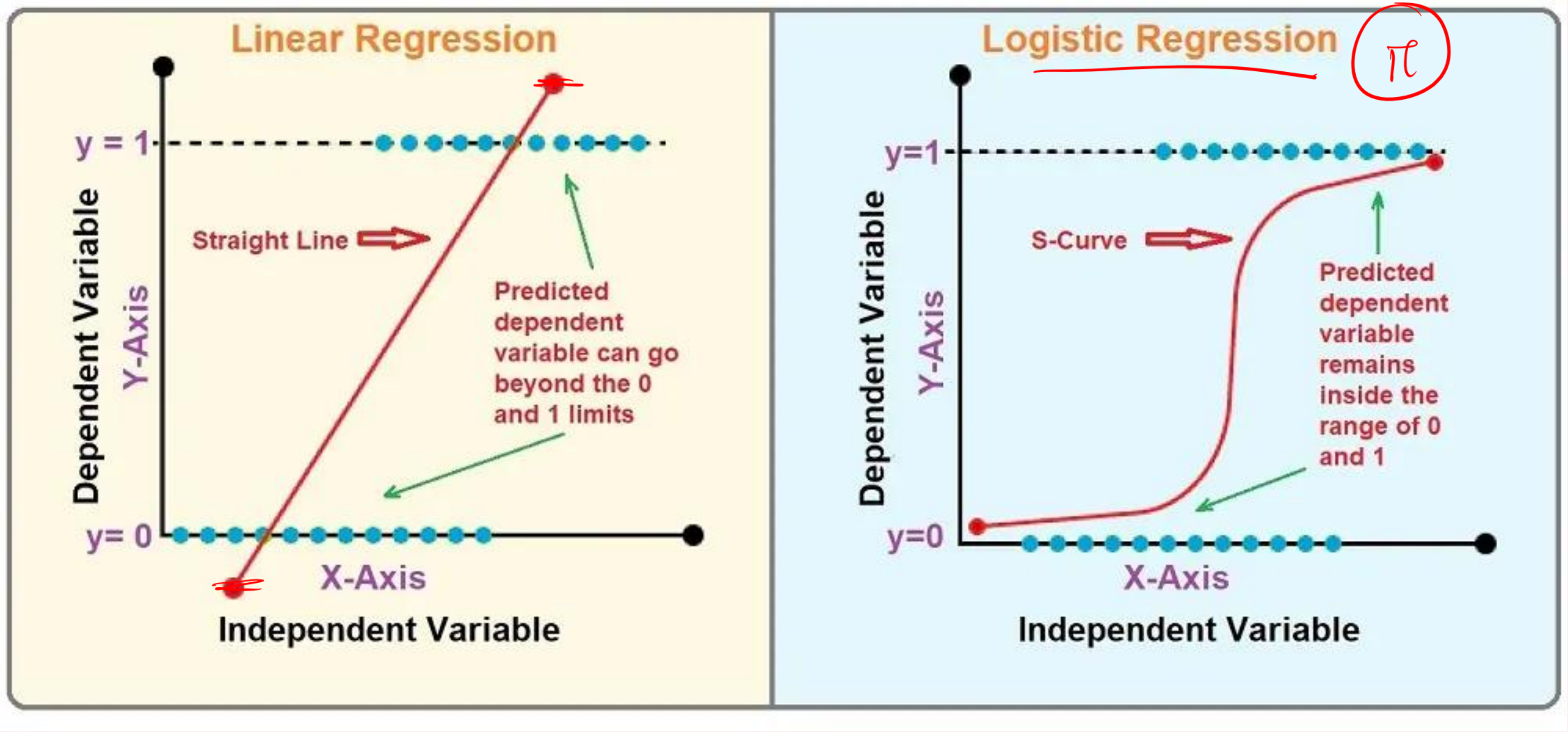
- Logistic regression is a statistical method used for binary classification problems, where the outcome or dependent variable is categorical and typically represents two classes.

→ 0 – 1

- It predicts the probability that a given input point belongs to a certain class, using a logistic function (also called the sigmoid function) to map any real-valued number into the interval between 0 and 1.
- Probabilistic Output: It outputs the probability of the input belonging to a class, which can be thresholded to make a decision (if the probability ≥ 0.5, then classify as 1).

Yes
A

< 0.5 → 0
No


Linear Regression
y = 1
y= 0
Straight Line
Predicted dependent variable can go beyond the 0 and 1 limits
Dependent Variable
Y-Axis
X-Axis
Independent Variable
Logistic Regression
π
y=1
y=0
S-Curve
Predicted dependent variable remains inside the range of 0 and 1
Dependent Variable
Y-Axis
X-Axis
Independent Variable

Form of logistic regression (using sigmoid model)

- Simple logistic regression (single predictor)

? X hanya 1

$$p(X) = P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X)}} = \boxed{\frac{e^{(\beta_0+\beta_1 X)}}{1 + e^{(\beta_0+\beta_1 X)}}}$$

- Multiple logistic regression (two or more predictors)

$$p(X) = P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\cdots++\beta_p X_p)}} = \frac{e^{(\beta_0+\beta_1 X_1+\cdots++\beta_p X_p)}}{1 + e^{(\beta_0+\beta_1 X_1+\cdots++\beta_p X_p)}}$$

$$\downarrow$$

$$\log (P(Y=1|X)) = \log \left(\frac{e^{\beta x}}{1+e^{\beta x}} \right) \rightarrow \text{linear} \rightarrow$$

y → yg diprediksi

↳ kategori
- 0 — 180
- 1 — 570

The Important Things in ~~Linear~~ logistic. Regression

1. Model Fit (Goodness of fit model) ✓
   - Hosmer-Lemeshow Test -> A p-value greater than 0.05 suggests that there is no significant difference between observed and predicted values, indicating a good fit.
   - Pseudo R-Squared ✓
     - McFadden's R^2: This is the most common pseudo R-squared for logistic regression. It ranges between 0 and 1, with higher values indicating a better fit, although values between 0.2 and 0.4 are considered reasonable.
     - Cox & Snell R^2: It is another pseudo R-squared that compares the likelihood of the model to the null model. It can be difficult to interpret directly as it does not reach a maximum value of 1.
     - ✓ Nagelkerke R^2: This is an adjusted version of Cox & Snell R^2, which adjusts the scale to allow it to reach a maximum value of 1.
- Confusion matrix (accuracy, precision, recall, f1-score)

2. The relationship between the response and predictors -> using hypothesis testing

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

(there is no relationship between the response and predictors)

$$H_1 : at\ least\ one\ \beta_j\ is\ non-zero$$

(At least one predictor that relates to the response)

Statistics test: $likelihood\ ratio\ test\ (G\ test)$

$G\ test = -2 \times (loglikelihood\ of\ restricted\ model - loglikelihood\ of\ the\ full\ model)$

Rejected H_0 if $G\ test > \chi^2_{table}$

→ p-value < 0,05 → signifikan

3. Deciding on Important Variables (partial testing) -> backward, forward method

$$H_0: \beta_j = 0$$

(There is no relationship between X and Y)

$$H_1: \beta_j \ is \ non - zero$$

(There is a relationship between X and Y)

Statistics test: $Wald\ test\ (W)$

$$W = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \ or \ W^2 = \left(\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}\right)^2$$

Rejected H_0 if $W > Z_{table}$ or $W^2 > \chi^2_{table}$

p-value < 0,05 → signifikan

Assumption in Logistic Regression:

1. Independence: Observations must be independent of each other.
2. Linearity of the logit: The relationship between independent variables and the log odds of the dependent variable should be linear.
3. No Multicollinearity: Independent variables should not be highly correlated. Multicollinearity can be checked using Variance Inflation Factor (VIF) or correlation matrices

training : 80%   testing : 20%

How to modelling:

1. Split the dataset into training set and testing set
2. Model Training (Fitting) -> using training set
3. Model evaluation -> check the goodness of fit model using Pseudo R-Squared, Hosmer-Lemeshow test, or confusion matrix
4. Model interpretation -> check the significant variables using $G\ test$ and $Wald\ test$
5. Model testing -> using testing set -> apply the model to the testing dataset (or unseen data) to evaluate its performance and generalizability; use the same evaluation metrics to determine how well the model performs on the test data compared to the training data.

Classification

Classification:

- Predict the class of item
- Creating a model based on training dataset

Application:

- Credit approval
- Diagnosis of disease
- Fraud detection
- Churn detection

- Proses untuk menyatakan suatu objek ke salah satu kategori yg sudah didefinisikan sebelumnya.
- Proses pembelajaran fungsi target (model klasifikasi) yg memetakan setiap sekumpulan atribut x (input) ke salah satu class/label y yang didefinisikan sebelumnya.
  - Input : sekumpulan record (training set)
  - Setiap record terdiri atas sekumpulan atribut, salah satu atribut adalah class.
  - Mencari model utk atribut class sebagai fungsi dari nilai2 utk atribut yg lain.
- Tujuannya: record yang belum diketahui kelasnya akan diberi label seakurat mungkin.

Classification algorithms principles

- Distance-based algorithms : K-Nearest ✓ Neighbor
- Probability-based algorithms : Naïve Bayes, Logistic Regression ✓
- Search-based algorithms : Decision Tree → If condition
- Optimization-based algorithms : Artificial Neural Network, Support Vector Machines

↳ classification. 2 kategori

Bayesian Classification

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- P(H | X) Kemungkinan H benar jika X. X adalah kumpulah **atribut.**
- P(H) Kemungkinan H di data, independen terhadap X
- P (“Single” | $\overset{x_1}{\underline{\text{“muka sayu”}}}$, $\overset{x_2}{\underline{\text{“baju berantakan”}}}$, x_3 ← $\underline{\text{“jalan sendiri”)}}$ ✸nilainya besar
- P (“Non Single” | “muka ceria”, “baju rapi”, “jalan selalu berdua”) ✸nilainya besar
- P (“Single”) = jumlah single / jumlah mahasiwa

Bayesian Classification

$P(B|A)\,P(A)$

- P(H | X) ✹posterior
- P(H) ✹a priori
- P (X | H) probabilitas X, jika kita ketahui bahwa H benar ✹data training
- Kegiatan klasifikasi: kegiatan mencari P (H | X) yang paling maksimal
- Teorema Bayes:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})}$$

Klasifikasi

x_1 x_3 x_2

X = (“muka cerah”, “jalan sendiri”, “baju rapi”)

Kelasnya Single atau Non Single?

Cari P(H|X) yang paling besar:

(“Single” | “muka cerah”, “jalan sendiri”, “baju rapi”)

Atau

(“**Non** Single” | “muka cerah”, “jalan sendiri”, “baju rapi”)

Harus memaksimalkan (C_i: kelas ke i)

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

Karena P(X) konstan untuk setiap C*i* maka bisa ditulis, pencarian max untuk:

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

Naïve Bayes Classifier

- Penyederhanaan masalah: Tidak ada kaitan antar atribut “jalan sendiri” tidak terkait dengan “muka sayu”

$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|\overset{\checkmark}{C_i}) = P(x_1|\overset{\checkmark}{C_i}) \times P(x_2|C_i) \times \overset{\checkmark}{\ldots} \times P(x_n|C_i)$$

X_1: atribut ke-1 (“jalan sendiri”)

X_n: atribut ke-n

Naïve Bayes Classifier

- Jika bentuknya kategori ,
  $P(x_k|C_i)$ = jumlah kelas C_i yang memiliki x_k dibagi $| C_i |$ (jumlah anggota kelas C_i di data contoh)
- Jika bentuknya continous dapat menggunakan distribusi gaussian

Contoh Naïve Bayes

x_1 x_2 x_3 x_4

buy-c → output

age	income	student	redit_rating	com
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30 ①	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30 ②	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

target output

$$\frac{3/\cancel{14}}{5/\cancel{14}}$$

$n : \frac{5}{14}$

$y : \frac{9}{14}$

$$\frac{2/\cancel{14}}{9/\cancel{14}}$$

Contoh Naïve Bayes

P(Ci):

P(buys_computer = “yes”) = 9/14 = 0.643

P(buys_computer = “no”) = 5/14= 0.357

yes beli komputer.

$\frac{2}{14}$ $\frac{9}{14}$

Training: Hitung P(X|Ci) untuk setiap kelas

- P(age = “<=30” | buys_computer = “yes”) = 2/9 = 0.222
- P(age = “<= 30” | buys_computer = “no”) = 3/5 = 0.6 ✓
- P(income = “medium” | buys_computer = “yes”) = 4/9 = 0.444
- P(income = “medium” | buys_computer = “no”) = 2/5 = 0.4 ✓
- P(student = “yes” | buys_computer = “yes) = 6/9 = 0.667
- P(student = “yes” | buys_computer = “no”) = 1/5 = 0.2 ✓
- P(credit_rating = “fair” | buys_computer = “yes”) = 6/9 = 0.667
- P(credit_rating = “fair” | buys_computer = “no”) = 2/5 = 0.4 ✓

$P(X|C_i) x P(C_i$

$P(C_i|X) = \frac{P(C_i \cap X)}{P(X)}$

Yes → 0,55

no → 0,22

Klasifikasi: X = (age <= 30 , income = medium, student = yes, credit_rating = fair) ✓

P(X|Ci) : P(X|buys_computer = “yes”) = 0.222 x 0.444 x 0.667 x 0.667 = $\boxed{0.044}$

P(X|buys_computer = “no”) = 0.6 x 0.4 x 0.2 x 0.4 = 0.019 ✓

P(X|Ci)*P(Ci) :

P(X|buys_computer = “yes”) * P(buys_computer = “yes”) = 0.028 ← ✓

P(X|buys_computer = “no”) * P(buys_computer = “no”) = 0.007 ✓

Pro, Cons Naïve Bayes

- Keuntungan
  - Mudah untuk dibuat
  - Hasil bagus
- Kerugian
  - Asumsi independence antar atribut membuat akurasi berkurang (karena biasanya ada keterkaitan)

		AC	
		1	0
predict	1	0	1
	0	2	1

Metriks Evaluasi

Y act.	Predict
1	0
0	0
0	1
1	0

Akurasi : $\frac{1}{4}$ = 25% = 0,25

presisi (+) : $\frac{0}{1}$ = 0 ✓

Recall (+) : $\frac{0}{2}$ = 0 ✓

		Actual Values	
		1 (Postive)	0 (Negative)
Predicted Values	1 (Postive)	**TP** (True Positive)	**FP** (False Positive) Type I Error
	0 (Negative)	**FN** (False Negative) Type II Error	**TN** (True Negative)

Presisi (-) : $\frac{1}{3}$ = 0,33 ✓

Recall (-) : $\frac{1}{2}$ = 0,5 = 80% ✓

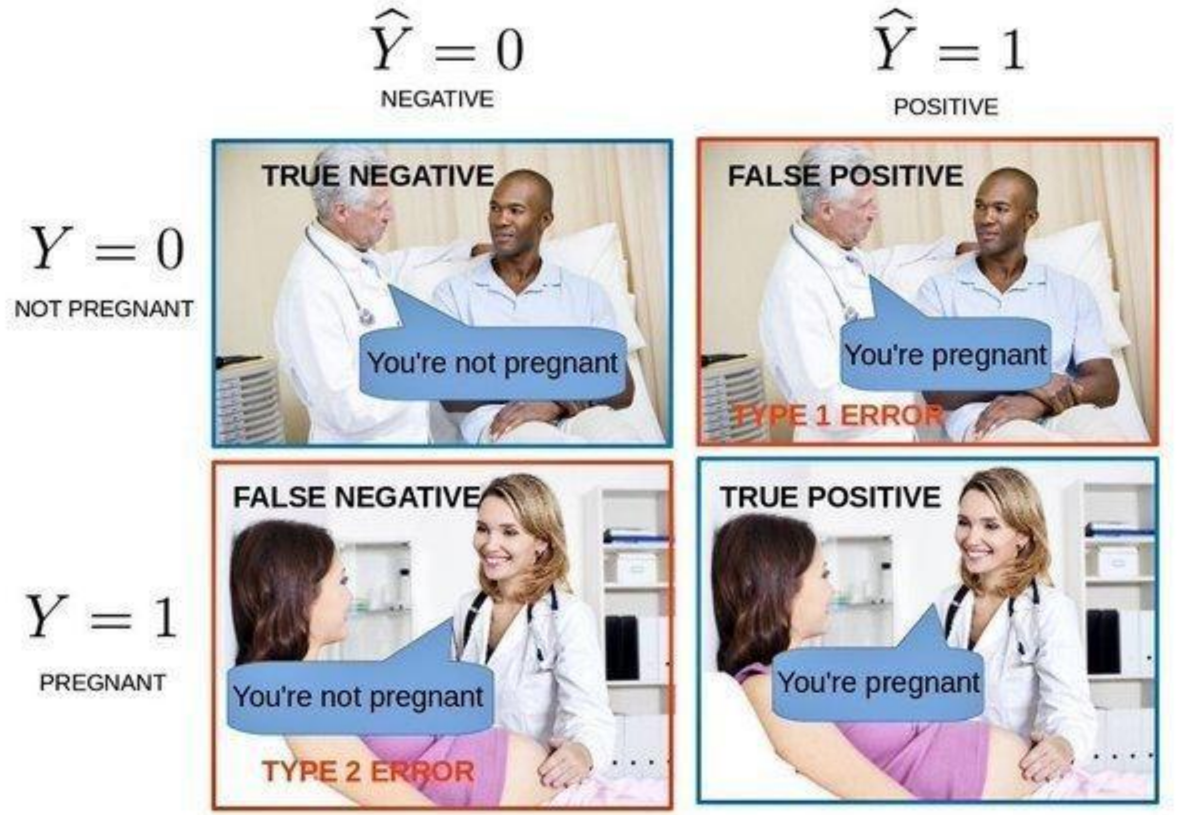
Accuracy = TP+TN/(TP+FP+FN+TN)
Precision = TP/(TP+FP)
Recall = TP/TP+FN
F1 Score/F-measure = 2*(Recall * Precision) / (Recall + Precision)
ROC - AUC

- ***True Positive*** **(TP)**
  Merupakan data positif yang diprediksi benar. Contohnya, pasien menderita kanker (*class* 1) dan dari model yang dibuat memprediksi pasien tersebut menderita kanker (*class* 1).
- ***True Negative*** **(TN)**
  Merupakan data negatif yang diprediksi benar. Contohnya, pasien tidak menderita kanker (*class* 2) dan dari model yang dibuat memprediksi pasien tersebut tidak menderita kanker (*class* 2).
- ***False Postive*** **(FP) — Type I Error**
  Merupakan data negatif namun diprediksi sebagai data positif. Contohnya, pasien tidak menderita kanker (*class* 2) tetapi dari model yang telah memprediksi pasien tersebut menderita kanker (*class* 1).
- ***False Negative*** **(FN) — Type II Error**
  Merupakan data positif namun diprediksi sebagai data negatif. Contohnya, pasien menderita kanker (*class* 1) tetapi dari model yang dibuat memprediksi pasien tersebut tidak menderita kanker (*class* 2).

F1 score (+) = $\frac{2(\text{Recall} * \text{Presisi})}{(\text{Recall} + \text{Presisi})}$ = 0

= F1 score (-) = $\frac{2(0,5 \times 0,33)}{0,83}$

Confusion Matrix



Receiver Operating Characteristic Curves (ROC curve)

Receiver operating characteristic curves are a useful visual tool for comparing two classification models.

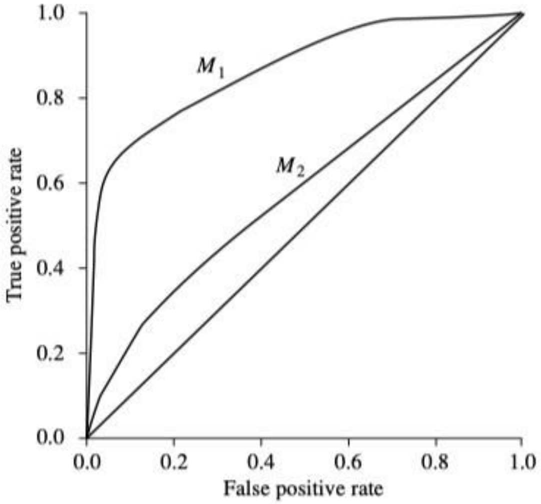


Figure 8.20 ROC curves of two classification models, M_1 and M_2. The diagonal shows where, for every true positive, we are equally likely to encounter a false positive. The closer an ROC curve is to the diagonal line, the less accurate the model is. Thus, $M1$ is more accurate here.

The Methods in Data Mining (Regression)

Thankyou