DATA MINING

PERTEMUAN KE- 4

Proses Data Mining

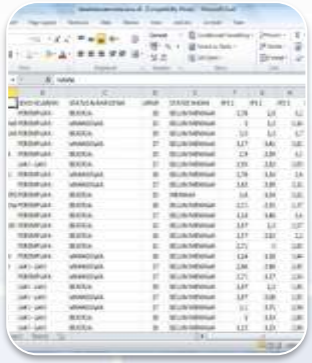
Proses Data Mining

1. Proses dan Tools Data Mining
2. Penerapan Proses Data Mining
3. **Evaluasi Model Data Mining**
4. Proses Data Mining berbasis CRISP-DM

3. EVALUASI MODEL DATA MINING

I.4

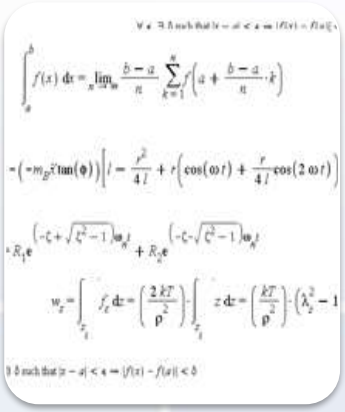
Proses Data Mining



1. Himpunan Data

(Pemahaman dan Pengolahan Data)

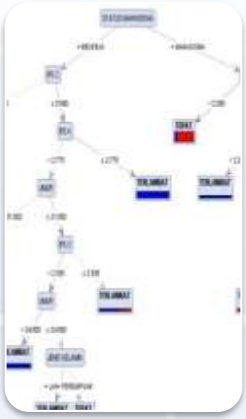
DATA PRE-PROCESSING
Data Cleaning
Data Integration
Data Reduction
Data Transformation



2. Metode Data Mining

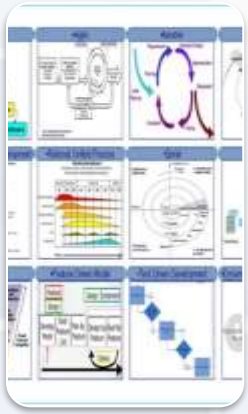
(Pilih Metode Sesuai Karakter Data)

Estimation
Prediction
Classification
Clustering
Association



3. Pengetahuan

(Pola/Model/Rumus/ Tree/Rule/Cluster)



4. Evaluation

(Akurasi, AUC, RMSE, Lift Ratio,…)

Versi : 01

Evaluasi Data Mining

1. Estimation:
   - Error: Root Mean Square Error (RMSE), MSE, MAPE, etc
2. Prediction/Forecasting (Prediksi/Peramalan):
   - Error: Root Mean Square Error (RMSE) , MSE, MAPE, etc
3. Classification:
   - Confusion Matrix: Accuracy
   - ROC Curve: Area Under Curve (AUC)
4. Clustering:
   - Internal Evaluation: Davies–Bouldin index, Dunn index,
   - External Evaluation: Rand measure, F-measure, Jaccard index, Fowlkes–Mallows index, Confusion matrix
5. Association:
   - Lift Charts: Lift Ratio
   - Precision and Recall (F-measure)

I.6 Kriteria Evaluasi dan Validasi Model

1. Akurasi
   - Ukuran dari seberapa baik model mengkorelasikan antara hasil dengan atribut dalam data yang telah disediakan
   - Terdapat berbagai model akurasi, tetapi semua model akurasi tergantung pada data yang digunakan
2. Kehandalan
   - Ukuran di mana model data mining diterapkan pada dataset yang berbeda
   - Model data mining dapat diandalkan jika menghasilkan pola umum yang sama terlepas dari data testing yang disediakan
3. Kegunaan
   - Mencakup berbagai metrik yang mengukur apakah model tersebut memberikan informasi yang berguna

Keseimbangan diantaranya ketiganya diperlukan karena belum tentu model yang akurat adalah handal, dan yang handal atau akurat belum tentu berguna

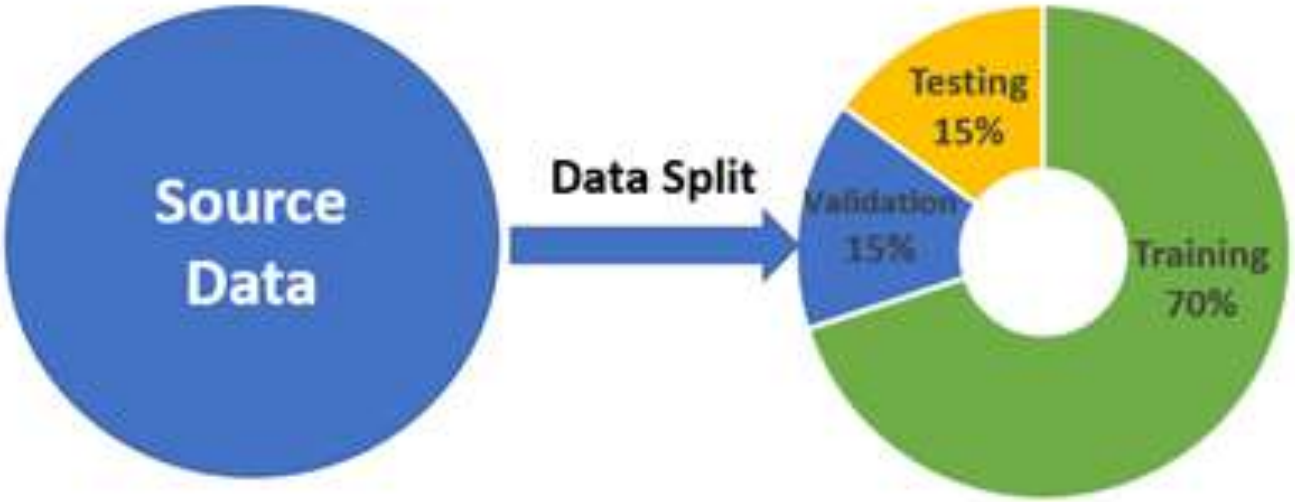
Evaluasi Model Data Mining

- Pembagian dataset:
  - Dua bagian: data training dan data testing
  - Tiga bagian: data training, data validation dan data testing

- Data training untuk pembentukan model, dan data testing digunakan untuk pengujian model

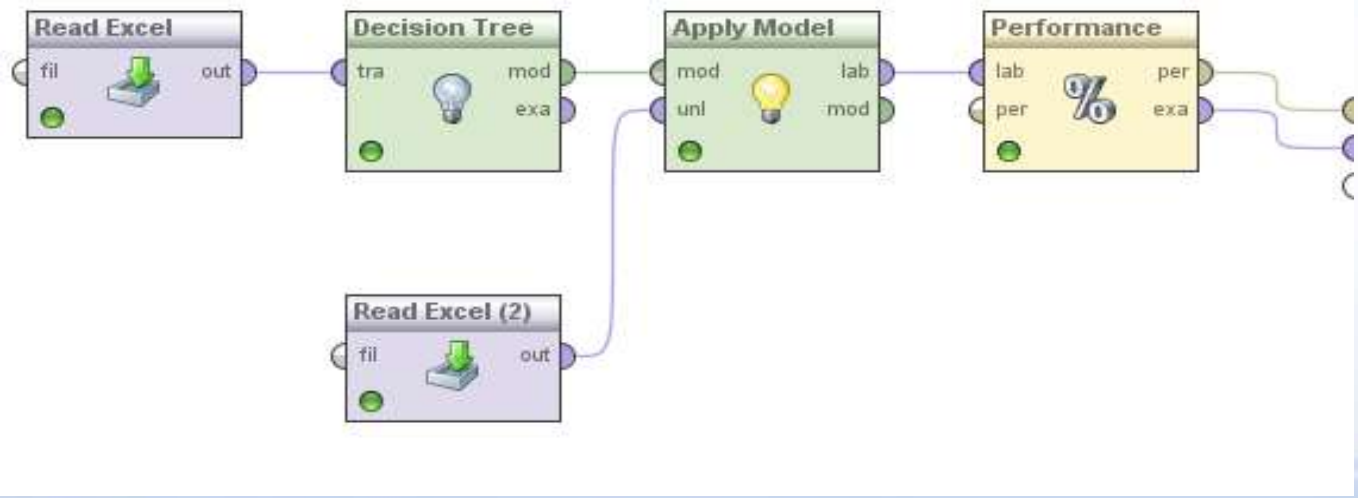
- Pemisahan data training dan testing
  1. Data dipisahkan secara manual
  2. Data dipisahkan otomatis dengan operator Split Data
  3. Data dipisahkan otomatis dengan X Validation

2 Pemisahan Data Manual



I.9 Latihan: Penentuan Kelayakan Kredit

- Gunakan dataset di bawah:
  - creditapproval-training.xls: untuk membuat model
  - creditapproval-testing.xls: untuk menguji model
- Data di atas terpisah dengan perbandingan: data testing (10%) dan data training (90%)
- Data training sebagai pembentuk model, dan data testing untuk pengujian model, ukur performancenya



Versi : 01

Confusion Matrix → Accuracy

accuracy: 90.00%

	true MACET	true LANCAR	class precision
pred. MACET	53	4	92.98%
pred. LANCAR	6	37	86.05%
class recall	89.83%	90.24%	

- pred MACET- true MACET: Jumlah data yang diprediksi macet dan kenyataannya macet (TP)
- pred LANCAR-true LANCAR: Jumlah data yang diprediksi lancar dan kenyataannya lancar (TN)
- pred MACET-true LANCAR: Jumlah data yang diprediksi macet tapi kenyataannya lancer (FP)
- pred LANCAR-true MACET: Jumlah data yang diprediksi lancar tapi kenyataanya macet (FN)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{53 + 37}{53 + 37 + 4 + 6} = \frac{90}{100} = 90\%$$

Versi : 01

I.11 Precision and Recall, and F-measures

- Precision: exactness – what % of tuples that the classifier labeled as positive are actually positive

$$precision = \frac{TP}{TP + FP}$$

- Recall: completeness – what % of positive tuples did the classifier label as positive?

$$recall = \frac{TP}{TP + FN}$$

- Perfect score is 1.0
- Inverse relationship between precision & recall
- F measure (F1 or F-score): harmonic mean of precision and recall,

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

- Fß: weighted measure of precision and recall
  - assigns ß times as much weight to recall as to precision

$$F_\beta = \frac{(1 + \beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$$

Versi : 01

Penjelasan

1. True Positives (TP): Jumlah contoh positif yang benar-benar diidentifikasi sebagai positif oleh model atau sistem.
2. True Negatives (TN): Jumlah contoh negatif yang benar-benar diidentifikasi sebagai negatif oleh model atau sistem.
3. False Positives (FP): Jumlah contoh negatif yang salah diidentifikasi sebagai positif oleh model atau sistem.
4. False Negatives (FN): Jumlah contoh positif yang salah diidentifikasi sebagai negatif oleh model atau sistem.

Note:

- Precision, ukuran untuk memahami sejauh mana model atau sistem yang dianalisis memberikan hasil positif yang benar (true positive) dibandingkan dengan hasil positif yang salah (false positive).
- Recall, mengukur sejauh mana model atau sistem yang dianalisis mampu mengidentifikasi semua contoh positif yang sebenarnya.
- Precision, mencari keseimbangan antara precision dan recall, terutama ketika ada trade-off antara kedua metrik
- Weighted measure (pengukuran terponderasi), konsep dalam analisis data, statistik, dan evaluasi kinerja yang mengizinkan kita untuk memberikan bobot atau prioritas yang berbeda pada berbagai kelompok atau elemen data yang sedang dievaluasi.
- Accuracy (akurasi), metrik evaluasi yang digunakan untuk mengukur sejauh mana model klasifikasi atau sistem klasifikasi dapat mengidentifikasi contoh dengan benar secara keseluruhan. Akurasi mengukur persentase prediksi yang benar dari keseluruhan data yang dievaluasi.

Sensitivity and Specificity

Binary classification should be both sensitive and specific as much as possible:

1. Sensitivity measures the proportion of true 'positives' that are correctly identified (True Positive Rate (TP Rate) or Recall)

$$Sensitivity = \frac{Number\ of\ 'True\ Positives'}{Number\ of\ 'True\ Positives' + Number\ of\ 'False\ Negatives''}$$

2. Specificity measures the proportion of true 'negatives' that are correctly identified (False Negative Rate (FN Rate or Precision)

$$Specificity = \frac{Number\ of\ 'True\ Negatives'}{Number\ of\ 'True\ Negatives' + Number\ of\ 'False\ Positives''}$$

Keterangan

- Sensitivity (sensitivitas), juga dikenal sebagai True Positive Rate (TPR) atau recall, adalah metrik evaluasi yang digunakan dalam konteks klasifikasi, khususnya dalam evaluasi kinerja model klasifikasi atau sistem deteksi. Sensitivity mengukur sejauh mana model mampu mengidentifikasi semua contoh positif yang sebenarnya.

- Specificity (spesifisitas), metrik evaluasi yang digunakan dalam konteks klasifikasi, terutama dalam evaluasi kinerja model klasifikasi atau sistem deteksi. Specificity mengukur sejauh mana model mampu mengidentifikasi semua contoh negatif yang sebenarnya.

PPV and NPV

We need to know the probability that the classifier will give the correct diagnosis, but the sensitivity and specificity do not give us this information

- Positive Predictive Value (PPV) is the proportion of cases with 'positive' test results that are correctly diagnosed

$$PPV = \frac{Number\ of\ 'True\ Positives'}{Number\ of\ 'True\ Positives' + Number\ of\ 'False\ Positives''}$$

- Negative Predictive Value (NPV) is the proportion of cases with 'negative' test results that are correctly diagnosed

$$NPV = \frac{Number\ of\ 'True\ Negatives'}{Number\ of\ 'True\ Negatives' + Number\ of\ 'False\ Negatives'}.$$

Explanation (1)

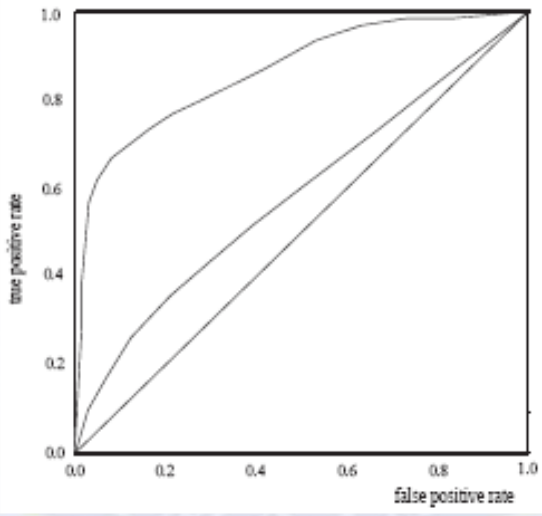
- Positive Predictive Value (PPV), metrik evaluasi yang mengukur sejauh mana hasil positif yang dihasilkan oleh suatu model klasifikasi atau sistem deteksi adalah benar positif.
- PPV mengukur persentase prediksi positif yang benar dari keseluruhan prediksi positif yang dilakukan oleh model.
- Positive Predictive Value menjadi penting dalam situasi di mana mengidentifikasi contoh positif dengan benar adalah prioritas tinggi dan kesalahan false positives (mengidentifikasi contoh negatif sebagai positif) memiliki konsekuensi yang serius atau berpotensi berbahaya. Ini terutama relevan dalam konteks di mana kesalahan positif palsu dapat memiliki dampak yang mahal atau berbahaya, seperti dalam pengujian medis atau dalam deteksi keamanan.

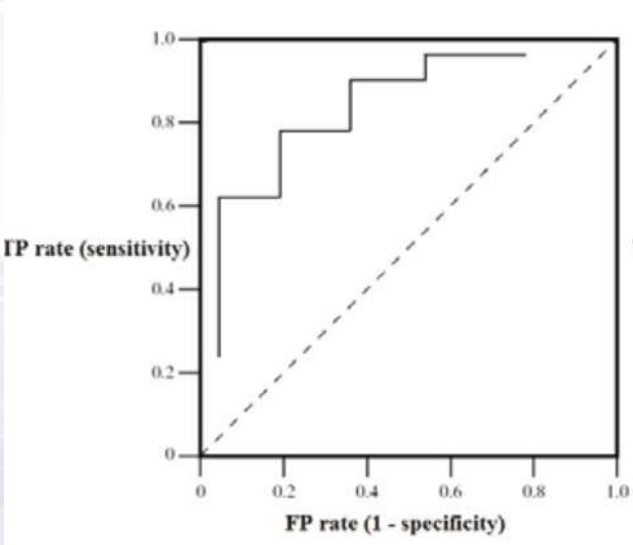
Explanation (2)

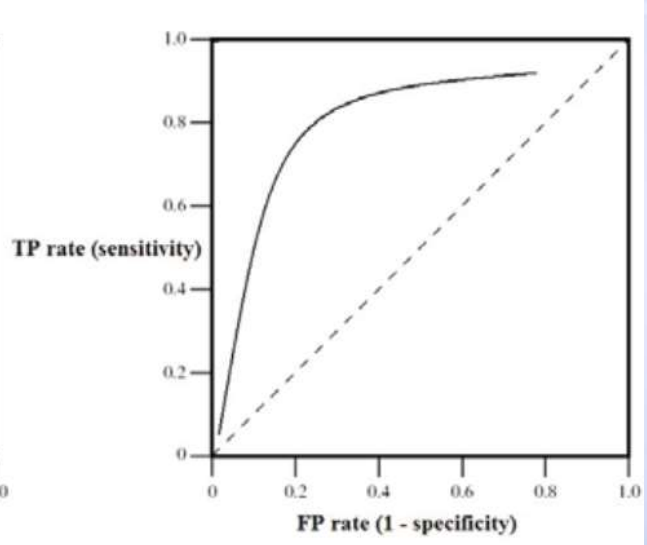
- Negative Predictive Value (NPV), metrik evaluasi yang digunakan dalam konteks klasifikasi, khususnya dalam evaluasi kinerja model klasifikasi atau sistem deteksi. NPV mengukur sejauh mana hasil negatif yang dihasilkan oleh suatu model klasifikasi adalah benar negatif. NPV mengukur persentase prediksi negatif yang benar dari keseluruhan prediksi negatif yang dilakukan oleh model.
- Negative Predictive Value menjadi penting dalam situasi di mana mengidentifikasi contoh negatif dengan benar adalah prioritas tinggi dan kesalahan false negatives (mengidentifikasi contoh positif sebagai negatif) memiliki konsekuensi yang serius atau berpotensi berbahaya.

I.19 Kurva ROC - AUC (Area Under Curve)

- ROC (Receiver Operating Characteristics) curves: for visual comparison of classification models
  - Originated from signal detection theory
- ROC curves are two-dimensional graphs in which the TP rate is plotted on the Y-axis and the FP rate is plotted on the X-axis
- ROC curve depicts relative trade-offs between benefits ('true positives') and costs ('false positives')
- Two types of ROC curves: discrete and continuous







Versi : 01

Penjelasan (1)

- **ROC (Receiver Operating Characteristic)** adalah sebuah grafik yang digunakan dalam analisis statistik, evaluasi kinerja model klasifikasi, dan sistem deteksi. Grafik ROC digunakan untuk mengilustrasikan dan mengevaluasi kemampuan model untuk membedakan antara dua kelas (biasanya positif dan negatif) pada berbagai titik ambang batas (threshold) yang berbeda.
- Grafik ROC umumnya memiliki dua sumbu:

1. **Sumbar X (Horizontal):** Tingkat False Positive Rate (FPR), yang mengukur sejauh mana model mengklasifikasikan contoh negatif sebagai positif. Ini dihitung sebagai False Positives dibagi oleh jumlah True Negatives ditambah False Positives.
2. **Sumbar Y (Vertikal):** Tingkat True Positive Rate (TPR), yang juga dikenal sebagai sensitivitas. Ini mengukur sejauh mana model mengklasifikasikan contoh positif sebagai positif. Ini dihitung sebagai True Positives dibagi oleh jumlah True Positives ditambah False Negatives.

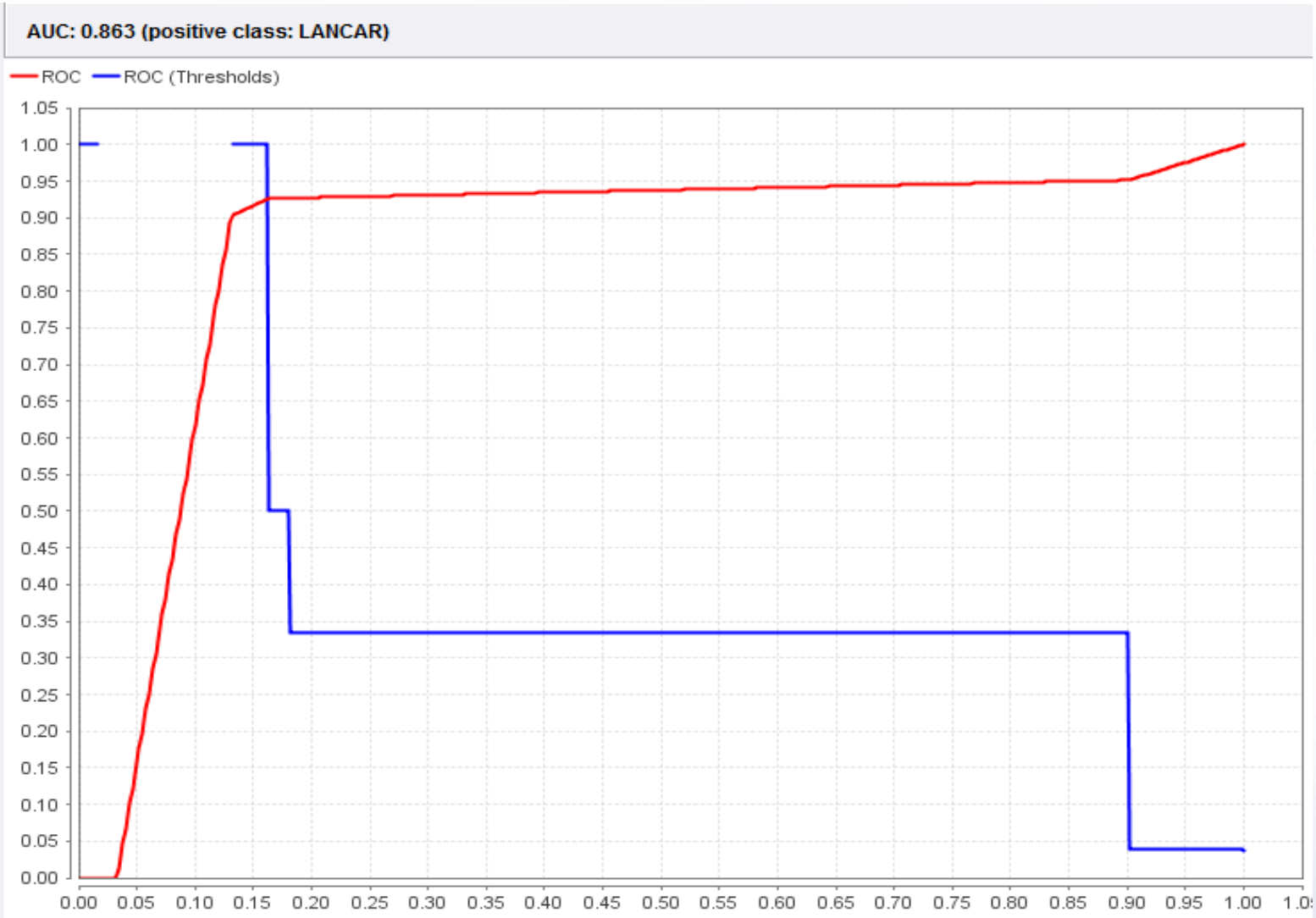
Penjelasan (2)

- AUC (Area Under the Curve) adalah metrik evaluasi yang digunakan untuk mengukur kinerja model klasifikasi dalam konteks kurva karakteristik operasi penerima (ROC), atau kadang-kadang disebut juga sebagai AUC-ROC.
- AUC-ROC mengukur sejauh mana model mampu membedakan antara kelas positif dan negatif. Nilai AUC-ROC berkisar antara 0 hingga 1, di mana:
- Jika AUC-ROC = 1, ini menunjukkan bahwa model memiliki kemampuan sempurna untuk membedakan antara kelas positif dan negatif.
- Jika AUC-ROC = 0,5, ini menunjukkan kinerja model yang sama dengan keberuntungan atau pengambilan keputusan acak (tidak ada kemampuan membedakan).
- Jika AUC-ROC < 0,5, ini menunjukkan bahwa model kurang baik daripada pengambilan keputusan acak (kebalikan dari prediksi positif dan negatif).
- **Sebagai aturan umum**, semakin besar nilai AUC-ROC, semakin baik kinerja model klasifikasi dalam membedakan kelas positif dan negatif.

Penjelasan (3)

- Selain grafik ROC itu sendiri, metrik penting lainnya yang berkaitan adalah area di bawah kurva ROC (ROC AUC atau AUC-ROC). Nilai AUC-ROC adalah area di bawah kurva ROC, dan semakin mendekati 1, semakin baik kinerja model klasifikasi. Nilai 0,5 menunjukkan kinerja acak, sedangkan nilai di bawah 0,5 menunjukkan kinerja yang buruk.
- Grafik ROC dan AUC-ROC adalah alat yang berguna dalam mengevaluasi dan membandingkan kinerja berbagai model klasifikasi, terutama ketika Anda ingin memahami sejauh mana model mampu membedakan antara kelas positif dan negatif pada berbagai tingkat ambang batas.

I.23 Kurva ROC - AUC (Area Under Curve)



Versi : 01

Guide for Classifying the AUC

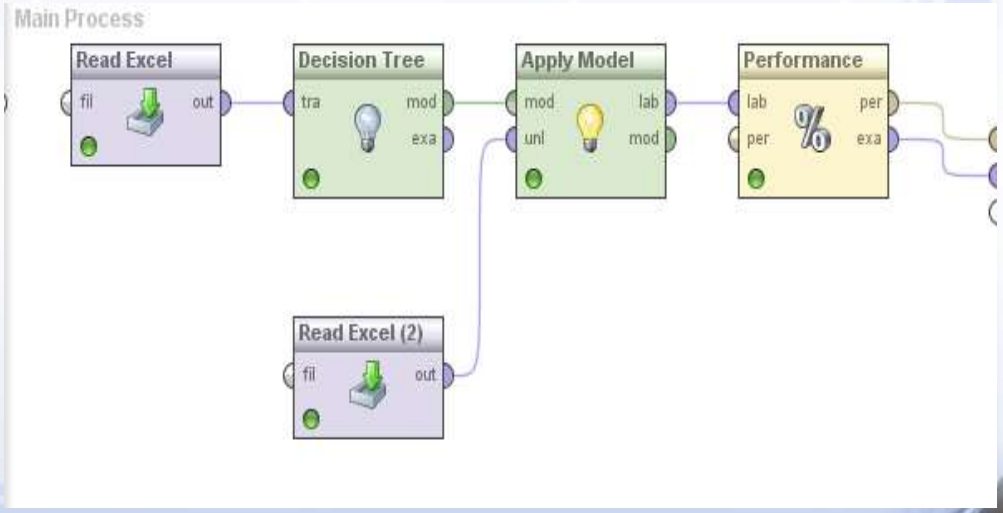
1. 0.90 - 1.00 = excellent classification
2. 0.80 - 0.90 = good classification
3. 0.70 - 0.80 = fair classification
4. 0.60 - 0.70 = poor classification
5. 0.50 - 0.60 = failure

(Gorunescu, 2011)

Latihan: Deteksi Serangan Jaringan

- Gunakan dataset di bawah:
  - intrusion-training.xls: untuk membuat model
  - intrusion-testing.xls: untuk menguji model
- Data di atas terpisah dengan perbandingan: data testing (10%) dan data training (90%)
- Jadikan data training sebagai pembentuk model/pola/knowledge, dan data testing untuk pengujian model
- Ukur performance (AUC dan Accuracy)

	C4.5
Accuracy	58%
AUC	0.86



Root Mean Square Error

- The square root of the mean/average of the square of all of the error

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- The use of RMSE is very common and it makes an excellent general purpose error metric for numerical predictions
- To construct the RMSE, we first need to determine the residuals
  - Residuals are the difference between the actual values and the predicted values
  - We denoted them by
  - where is the obser $\hat{y}_i - y_i$ e for the *i*th observation and
  - i y_i e predicted value
- The $\hat{y}_i$ an be positive or negative as the predicted value under or over estimates the actual value
- You then use the RMSE as a measure of the spread of the *y* values about the predicted *y* value

Root Mean Square Error (1)

- **Root Mean Square Error (RMSE)** adalah sebuah metrik yang digunakan untuk mengukur sejauh mana perbedaan antara nilai yang diamati (data aktual atau hasil prediksi) dengan nilai yang diprediksi oleh suatu model atau algoritma.
- RMSE umumnya digunakan dalam konteks evaluasi kinerja model statistik atau machine learning, terutama dalam masalah regresi.

Root Mean Square Error (2)

- **RMSE** mengukur seberapa baik model dapat memprediksi nilai target dengan memperhitungkan selisih antara nilai sebenarnya dan nilai yang diprediksi oleh model. Untuk menghitung RMSE, langkah-langkah berikut biasanya diikuti:

1. Ambil selisih antara setiap nilai sebenarnya (y) dan nilai yang diprediksi (ŷ) oleh model.
2. Kuadratkan setiap selisih.
3. Hitung rata-rata dari kuadrat selisih tersebut.
4. Akar kuadratkan hasil rata-rata tersebut untuk mendapatkan RMSE.

Root Mean Square Error (3)

- **Rumus RMSE:**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Di mana:

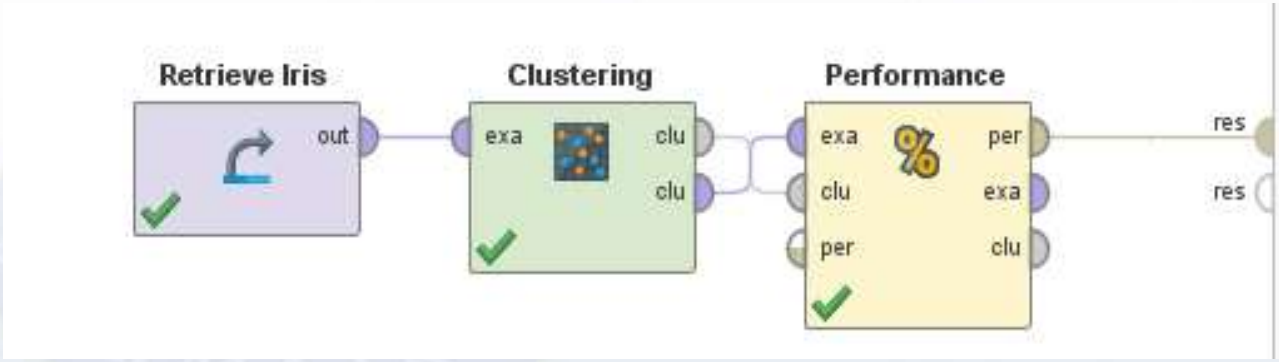
- n adalah jumlah data yang dievaluasi.
- y_i adalah nilai sebenarnya.
- $\hat{y}_i$ adalah nilai yang diprediksi oleh model.

Root Mean Square Error (4)

- RMSE menghasilkan skor non-negatif, dan semakin kecil nilai RMSE, semakin baik kinerja modelnya.
- RMSE memberikan informasi tentang seberapa akurat model dalam memprediksi data aktual.
- Model dengan RMSE yang lebih kecil biasanya lebih baik dalam melakukan prediksi.
- Namun, penting juga untuk mempertimbangkan konteks masalah dan tujuan akhir saat mengevaluasi kinerja model, karena dalam beberapa kasus, nilai RMSE yang kecil mungkin tidak mencerminkan kinerja yang baik secara keseluruhan.

Latihan: Klastering Jenis Bunga Iris

1. Lakukan training pada data iris (ambil dari repositories rapidminer) dengan menggunakan algoritma clustering k-means
2. Ukur performance-nya dengan *Cluster Distance Performance*, cek dan analisis nilai yang keluar Davies Bouldin Indeks (DBI)
3. Lakukan pengubahan pada nilai k pada parameter k-means dengan memasukkan berbagai nilai: 3, 4, 5, 6, 7



k	DBI
2	0.405
3	0.666
4	0.764
5	0.806
6	0.910
7	0.999

Davies–Bouldin index (DBI)

- The Davies–Bouldin index (DBI) (introduced by David L. Davies and Donald W. Bouldin in 1979) is a metric for evaluating clustering algorithms
- This is an internal evaluation scheme, where the validation of how well the clustering has been done is made using quantities and features inherent to the dataset
- As a function of the ratio of the within cluster scatter, to the between cluster separation, a lower value will mean that the clustering is better
- This affirms the idea that no cluster has to be similar to another, and hence the best clustering scheme essentially minimizes the Davies–Bouldin index
- This index thus defined is an average over all the *i* clusters, and hence a good measure of deciding how many clusters actually exists in the data is to plot it against the number of clusters it is calculated over
- The number *i* for which this value is the lowest is a good measure of the number of clusters the data could be ideally classified into

Davies–Bouldin index (DBI)

- **Davies-Bouldin Index (DBI)** adalah sebuah metrik yang digunakan untuk mengukur kualitas partisi dalam data clustering.
- **Metrik ini bertujuan** untuk mengevaluasi sejauh mana partisi atau pengelompokan data (clustering) yang dihasilkan oleh algoritma clustering seperti K-Means atau Hierarchical Clustering.
- **Tujuannya** adalah untuk mengukur seberapa baik data telah dikelompokkan berdasarkan kedekatan antara cluster dan seberapa jauh antara cluster-cluster tersebut.
- **DBI mengukur** sejauh mana cluster yang satu dari yang lain.
- **Semakin rendah nilai DBI**, semakin baik partisi atau pengelompokan data, karena berarti cluster-cluster tersebut memiliki kedekatan yang tinggi satu sama lain dan jarak antara cluster-cluster tersebut relatif besar.

Rumus DBI

Rumus Davies-Bouldin Index (DBI) untuk data yang telah dikelompokkan menjadi k cluster adalah:

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

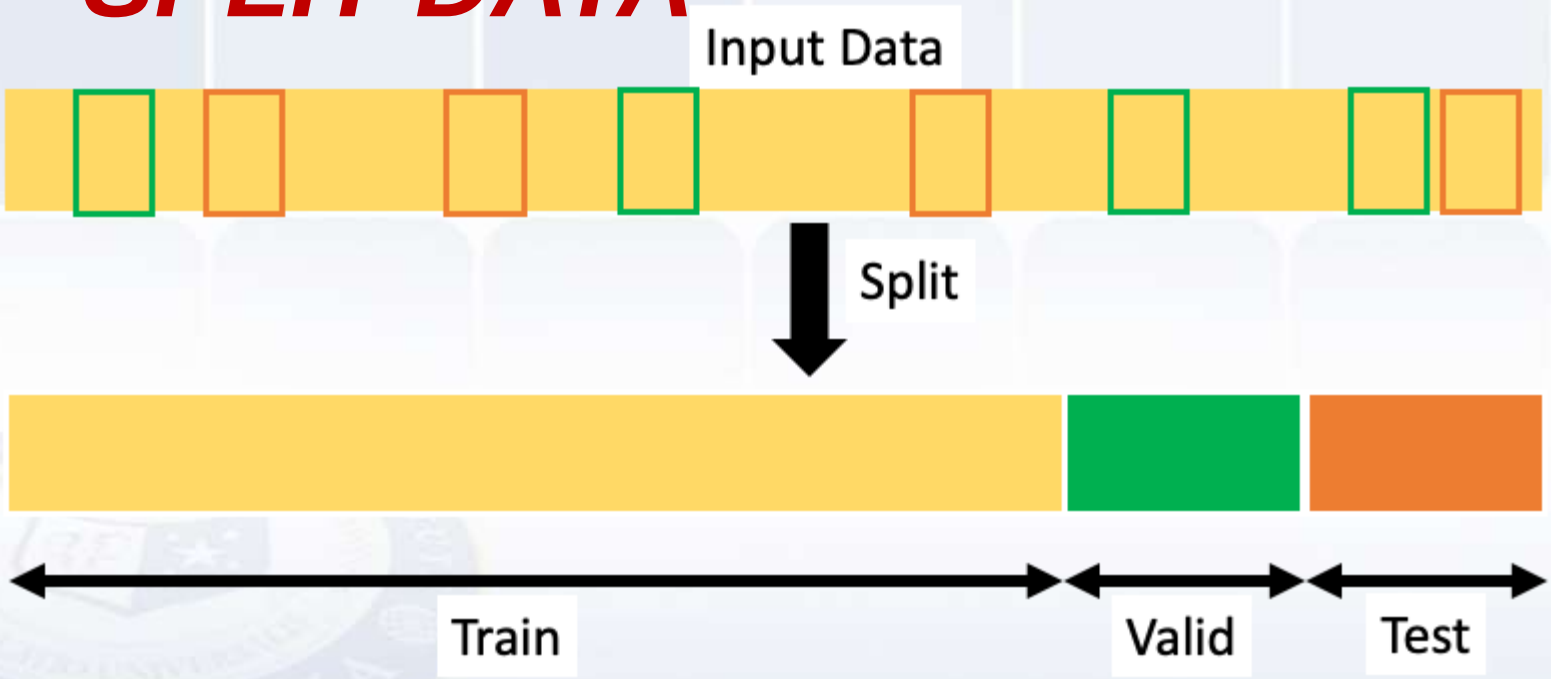
Di mana:

- k adalah jumlah cluster.
- σ_i adalah ukuran dispersi (variance) dari cluster ke-i.
- $d(c_i, c_j)$ adalah jarak antara pusat cluster c_i dan c_j.

Goals DBI

- **Tujuan utama** dari DBI adalah untuk mencari kombinasi cluster yang memberikan nilai DBI terendah.
- **Dalam konteks DBI**, semakin kecil nilai DBI, semakin baik kualitas pengelompokan data.
- **Sebagai hasilnya**, DBI membantu dalam pemilihan jumlah cluster yang optimal dan evaluasi kualitas pengelompokan yang dihasilkan oleh algoritma clustering.

3. PEMISAHAN DATA OTOMATIS DENGAN OPERATOR *SPLIT DATA*

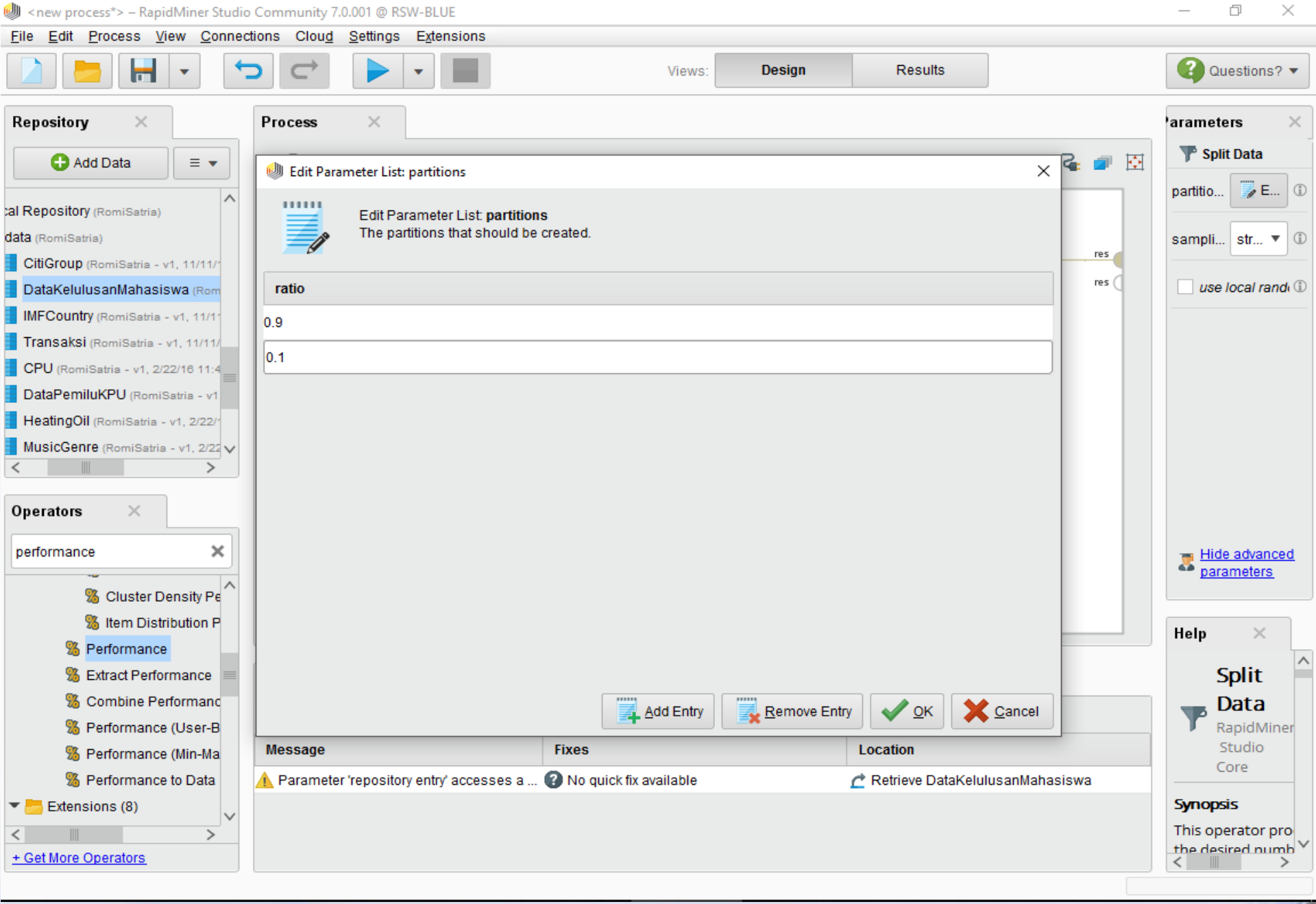


Split Data Otomatis

- The Split Data operator takes a dataset as its input and delivers the subsets of that dataset through its output ports
- The sampling type parameter decides how the examples should be shuffled in the resultant partitions:
  1. Linear sampling: Divides the dataset into partitions without changing the order of the examples
  2. Shuffled sampling: Builds random subsets of the dataset
  3. Stratified sampling: Builds random subsets and ensures that the class distribution in the subsets is the same as in the whole dataset

Split Data Otomatis

- **Operator Split Data** mengambil kumpulan data sebagai masukannya dan mengirimkan subkumpulan kumpulan data tersebut melalui port keluarannya
- **Parameter tipe pengambilan sampel** memutuskan bagaimana contoh harus diacak di partisi yang dihasilkan:
  - **Pengambilan sampel linier:** Membagi kumpulan data menjadi beberapa partisi tanpa mengubah urutan contoh Pengambilan sampel acak: Membuat subkumpulan acak dari kumpulan data
  - **Pengambilan sampel bertingkat:** Membangun subset acak dan memastikan bahwa distribusi kelas di subset sama dengan di seluruh dataset
  - **Pengambilan sampel bertingkat:** Membangun subset acak dan memastikan bahwa distribusi kelas di subset sama dengan di seluruh dataset


<new process*> – RapidMiner Studio Community 7.0.001 @ RSW-BLUE
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Questions?
Repository
Add Data
Process
Parameters
Split Data
Edit Parameter List: partitions
Edit Parameter List: partitions
The partitions that should be created.
ratio
0.9
0.1
Add Entry
Remove Entry
OK
Cancel
Operators
performance
Performance
Extract Performance
Performance to Data
Extensions (8)
+ Get More Operators
Message
Fixes
Location
Parameter 'repository entry' accesses a ...
No quick fix available
Retrieve DataKelulusanMahasiswa
Help
Split Data
RapidMiner Studio Core
Synopsis
Hide advanced parameters

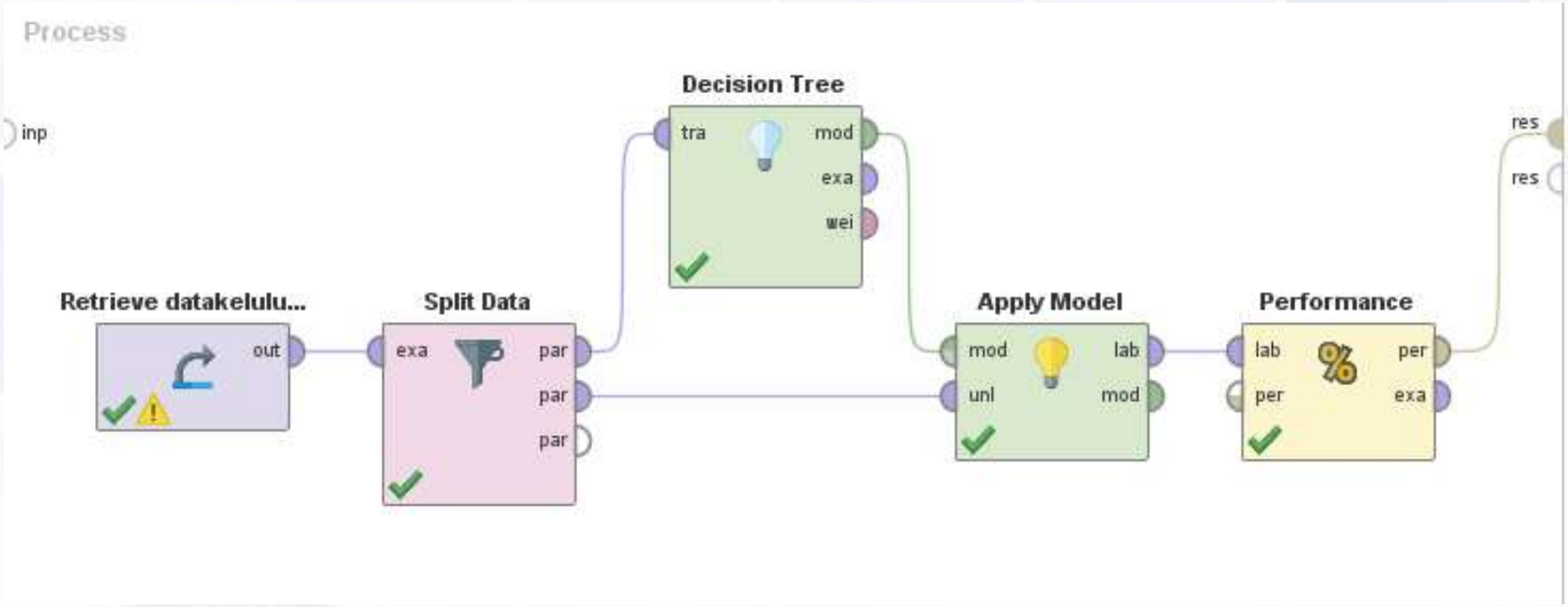
Versi : 01

Latihan: Prediksi Kelulusan Mahasiswa

1. Dataset: datakelulusanmahasiswa.xls
2. Pisahkan data menjadi dua secara otomatis (Split Data): data testing (10%) dan data training (90%)
3. Ujicoba parameter pemisahan data baik menggunakan Linear Sampling, Shuffled Sampling dan Stratified Sampling
4. Jadikan data training sebagai pembentuk model/pola/knowledge, dan data testing untuk pengujian model
5. Terapkan algoritma yang sesuai dan ukur performance dari model yang dibentuk

I.41

Proses Prediksi Kelulusan Mahasiswa

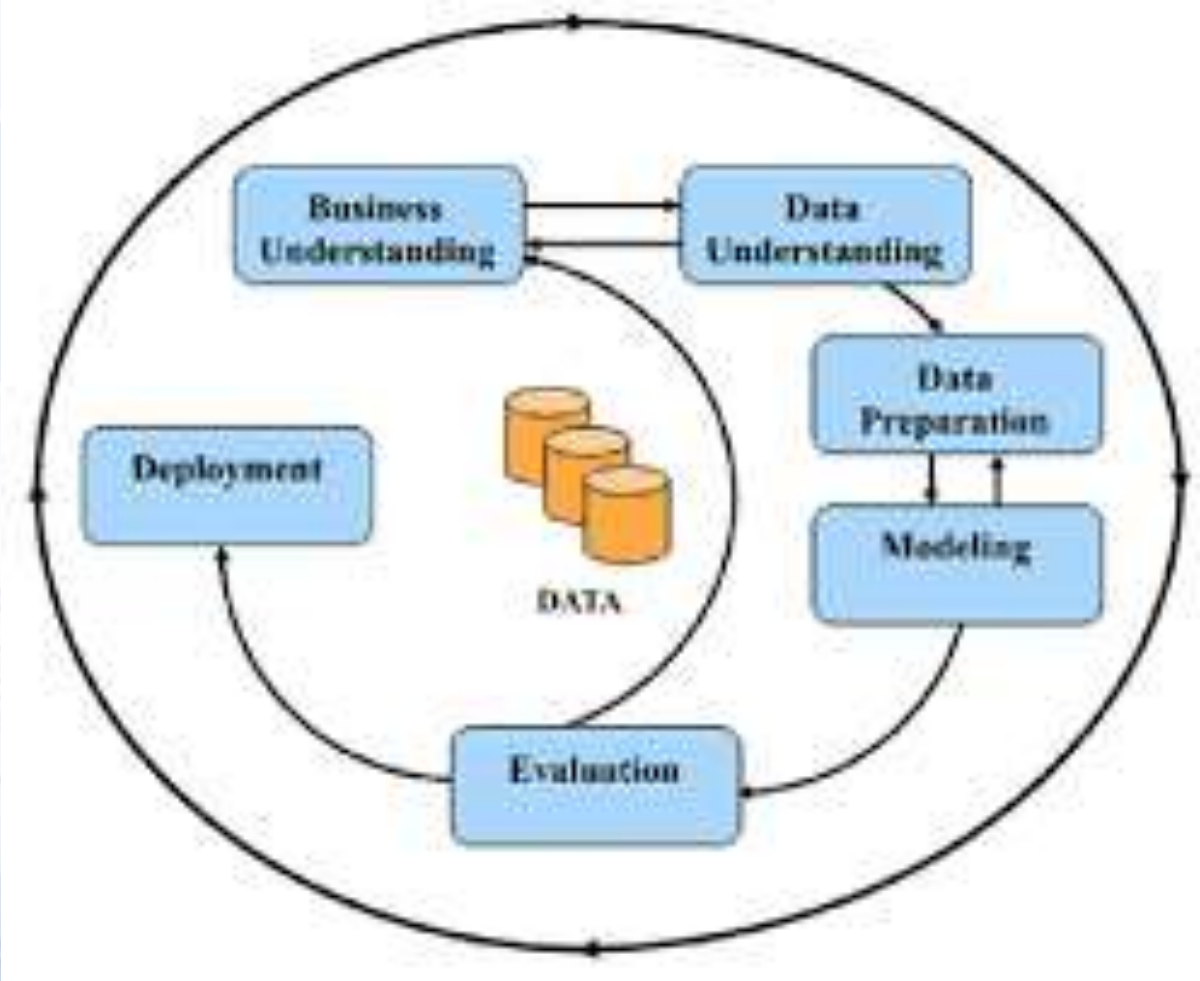


Versi : 01

Data Mining Standard Process

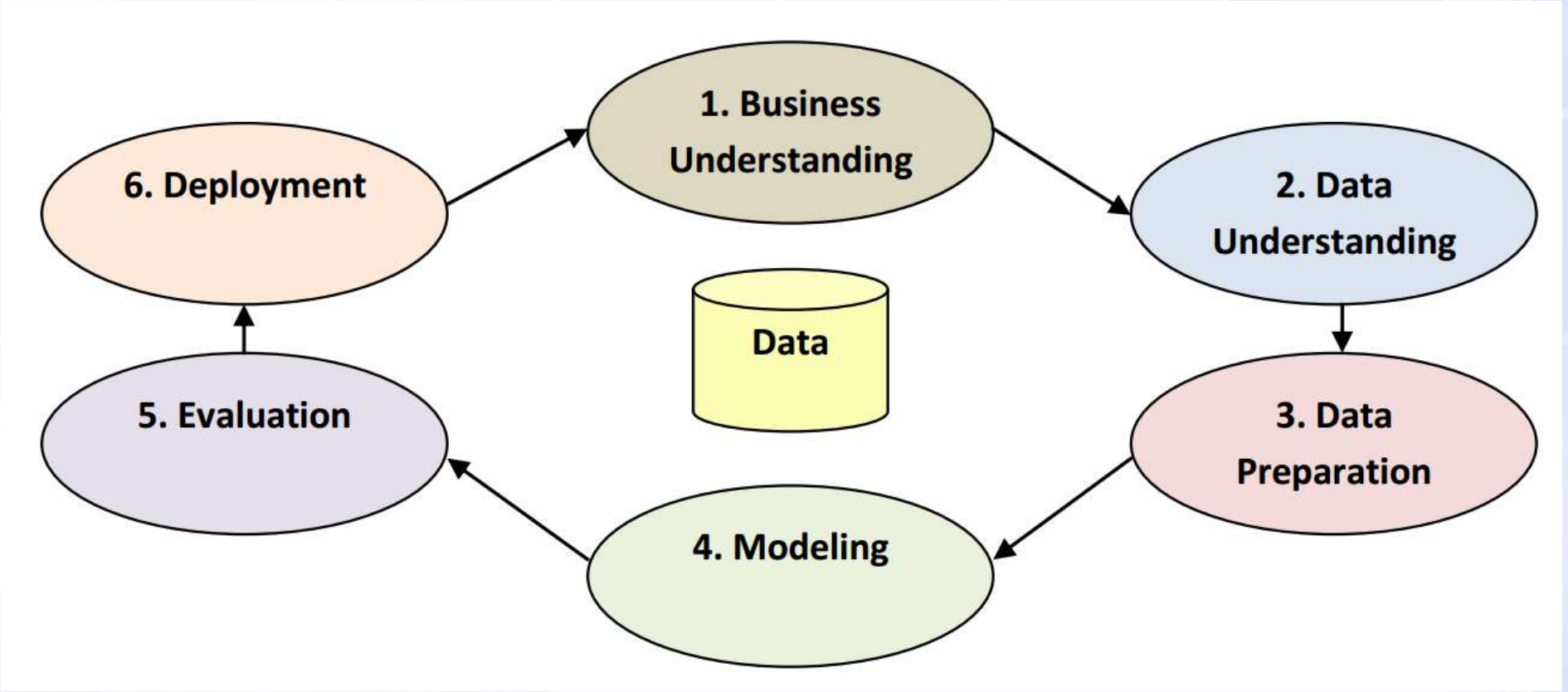
- Dunia industri yang beragam bidangnya memerlukan proses yang standard yang mampu mendukung penggunaan data mining untuk menyelesaikan masalah bisnis
- Proses tersebut harus dapat digunakan di lintas industry (cross-industry) dan netral secara bisnis, tool dan aplikasi yang digunakan, serta mampu menangani strategi pemecahan masalah bisnis dengan menggunakan data mining
- Pada tahun 1996, lahirlah salah satu standard proses di dunia data mining yang kemudian disebut dengan: the Cross-Industry Standard Process for Data Mining (CRISP–DM) *(Chapman, 2000)*

4. PROSES DATA MINING BERBASIS METODOLOGI CRISP-DM



Versi : 01

CRISP-DM


1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment
Data

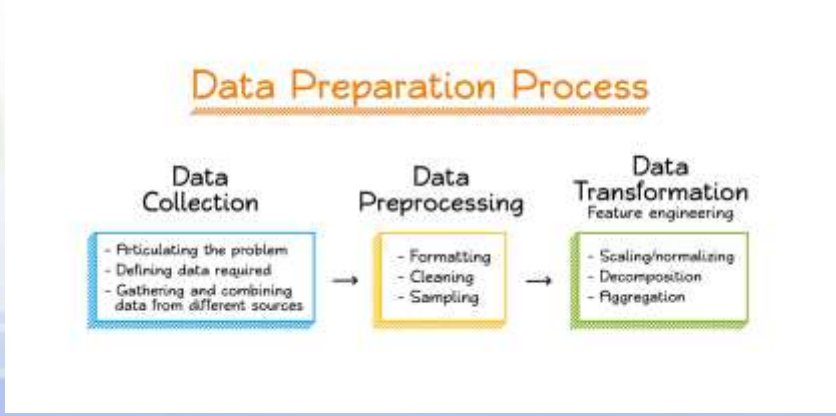
Definisi CRISP-DM

- **CRISP-DM adalah** singkatan dari Cross-Industry Standard Process for Data Mining, yang merupakan sebuah metodologi atau kerangka kerja yang digunakan dalam proses data mining dan analisis data.
- **Metodologi CRISP-DM terdiri dari** serangkaian tahap yang terstruktur untuk memandu para profesional dalam merancang, mengembangkan, dan mengevaluasi proyek-proyek data mining.
- **Kerangka kerja ini bersifat siklus,** yang berarti prosesnya dapat berulang dan beradaptasi dengan perubahan.

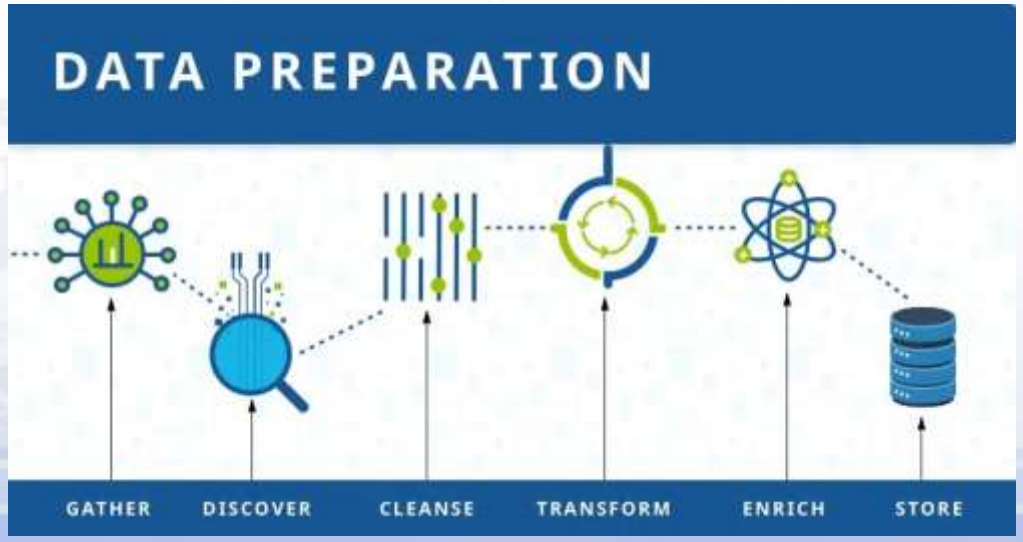
Metodologi CRISP-DM terdiri dari enam tahap utama:

1. **Business Understanding (Pemahaman Bisnis)**: Tahap ini fokus pada pemahaman terhadap tujuan bisnis dan masalah yang ingin dipecahkan melalui analisis data. Langkah-langkah awal melibatkan interaksi dengan pemangku kepentingan untuk memahami kebutuhan mereka.
2. **Data Understanding (Pemahaman Data)**: Pada tahap ini, data yang relevan dikumpulkan, dieksplorasi, dan dipahami. Ini termasuk memahami struktur data, keberadaan anomali, dan menentukan data apa yang diperlukan untuk mencapai tujuan proyek.

3. **Data Preparation (Persiapan Data)**: Data yang telah dipahami kemudian dipersiapkan untuk analisis. Ini mencakup pembersihan data, penggabungan data dari berbagai sumber, transformasi data, dan pemilihan atribut yang relevan.
4. **Modeling (Modeling)**: Tahap ini melibatkan pemilihan teknik atau algoritma data mining yang sesuai dan pembangunan model. Model ini digunakan untuk menganalisis data dan mencapai tujuan proyek.



5. **Evaluation (Evaluasi)**: Model yang telah dibangun dievaluasi untuk memastikan bahwa itu mencapai tujuan proyek dan kualitasnya cukup baik. Evaluasi melibatkan pengujian model dan pengukuran kinerja, serta penentuan apakah hasilnya sesuai dengan tujuan awal.
6. **Deployment (Penggunaan)**: Jika model dinyatakan berhasil pada tahap evaluasi, maka model tersebut diterapkan dalam lingkungan bisnis atau organisasi. Ini dapat berarti penggunaan model untuk pengambilan keputusan, integrasi model dalam sistem, atau tindakan lanjutan berdasarkan temuan.



1. Business Understanding

- Enunciate the project objectives and requirements clearly in terms of the business or research unit as a whole
- Translate these goals and restrictions into the formulation of a data mining problem definition
- Prepare a preliminary strategy for achieving these objectives
- Designing what you are going to build

2. Data Understanding

- Collect the data
- Use exploratory data analysis to familiarize yourself with the data and discover initial insights
- Evaluate the quality of the data
- If desired, select interesting subsets that may contain actionable patterns

3. Data Preparation

- Prepare from the initial raw data the final data set that is to be used for all subsequent phases
- Select the cases and variables you want to analyze and that are appropriate for your analysis
- Perform data cleaning, integration, reduction and transformation, so it is ready for the modeling tools

4. Modeling

- Select and apply appropriate modeling techniques
- Calibrate model settings to optimize results
- Remember that often, several different techniques may be used for the same data mining problem
- If necessary, loop back to the data preparation phase to bring the form of the data into line with the specific requirements of a particular data mining technique

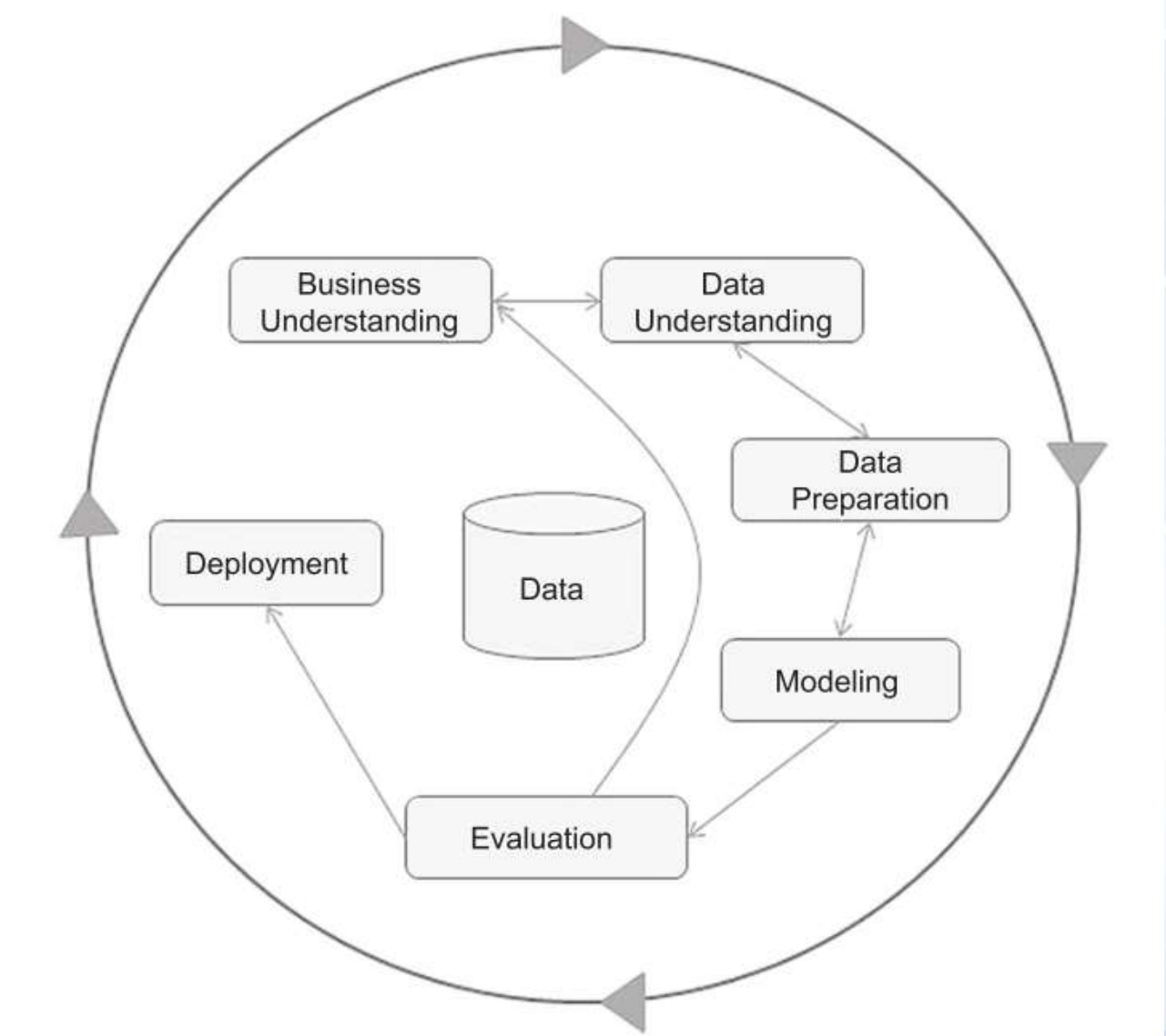
5. Evaluation

- Evaluate the one or more models delivered in the modeling phase for quality and effectiveness before deploying them for use in the field
- Determine whether the model in fact achieves the objectives set for it in the first phase
- Establish whether some important facet of the business or research problem has not been accounted for sufficiently
- Come to a decision regarding use of the data mining results

6. Deployment

- Make use of the models created:
  - model creation does not signify the completion of a project
- Example of a simple deployment:
  - Generate a report
- Example of a more complex deployment:
  - Implement a parallel data mining process in another department
- For businesses, the customer often carries out the deployment based on your model

CRISP-DM: Detail Flow

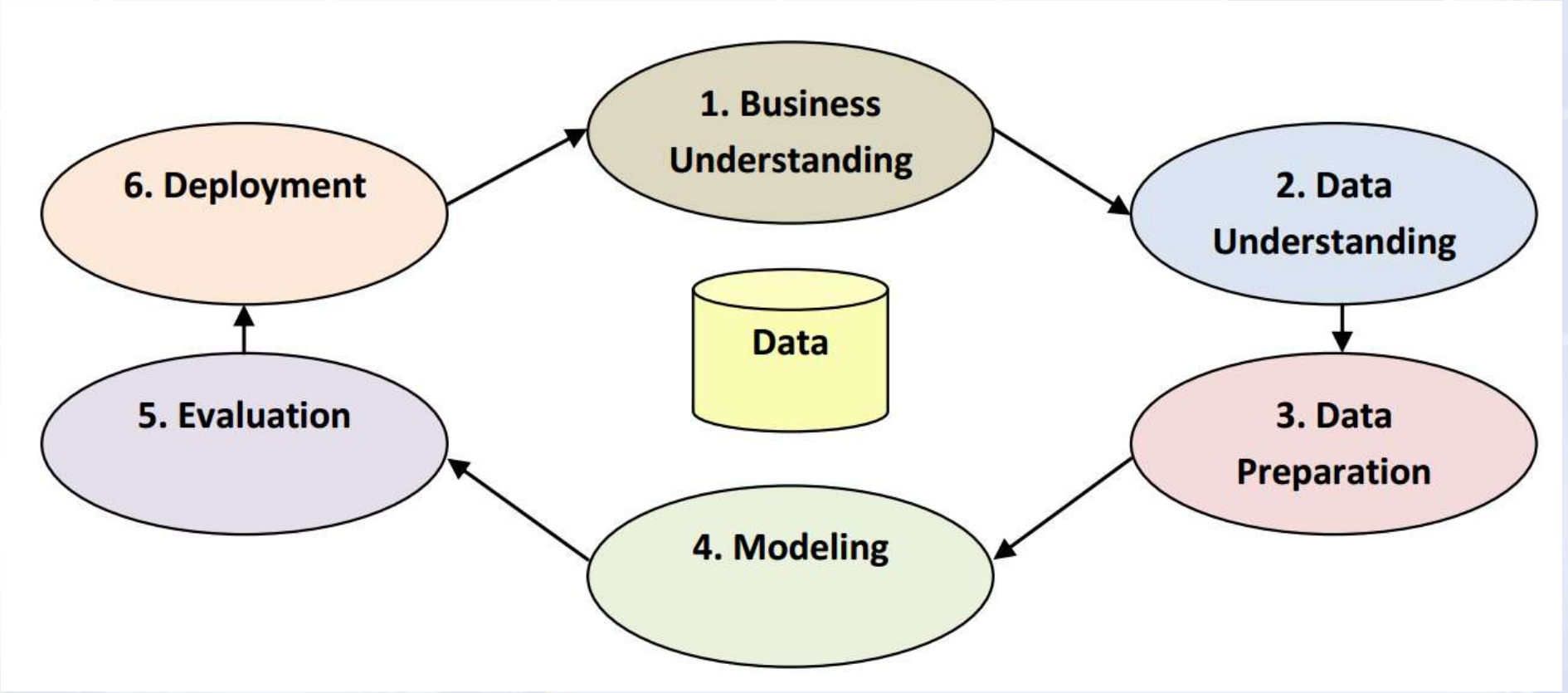


STUDI KASUS CRISP-DM

Kelulusan Mahasiswa di Universitas Suka Belajar

Dataset: datakelulusanmahasiswa.xls

CRISP-DM


1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment
Data

1. Business Understanding

- Problems:
  - Budi adalah Rektor di Universitas Suka Belajar
  - Universitas Suka Belajar memiliki masalah besar karena rasio kelulusan mahasiswa tiap angkatan sangat rendah
  - Budi ingin memahami dan membuat pola dari profile mahasiswa yang bisa lulus tepat waktu dan yang tidak lulus tepat waktu
  - Dengan pola tersebut, Budi bisa melakukan konseling, terapi, dan memberi peringatan dini kepada mahasiswa kemungkinan tidak lulus tepat waktu untuk memperbaiki diri, sehingga akhirnya bisa lulus tepat waktu
- Objective:
  - Menemukan pola dari mahasiswa yang lulus tepat waktu dan tidak

2. Data Understanding

I.59

- Untuk menyelesaikan masalah, Budi mengambil data dari sistem informasi akademik di universitasnya
- Data-data dikumpulkan dari data profil mahasiswa dan indeks prestasi semester mahasiswa, dengan atribut seperti di bawah
  1. NAMA
  2. JENIS KELAMIN: Laki-Laki atau Perempuan
  3. STATUS MAHASISWA: Mahasiswa atau Bekerja
  4. UMUR:
  5. STATUS NIKAH: Menikah atau Belum Menikah
  6. IPS 1: Indeks Prestasi Semester 1
  7. IPS 2: Indeks Prestasi Semester 1
  8. IPS 3: Indeks Prestasi Semester 1
  9. IPS 4: Indeks Prestasi Semester 1
  10. IPS 5: Indeks Prestasi Semester 1
  11. IPS 6: Indeks Prestasi Semester 1
  12. IPS 7: Indeks Prestasi Semester 1
  13. IPS 8: Indeks Prestasi Semester 1
  14. IPK: Indeks Prestasi Kumulatif
  15. STATUS KELULUSAN: Terlambat atau Tepat

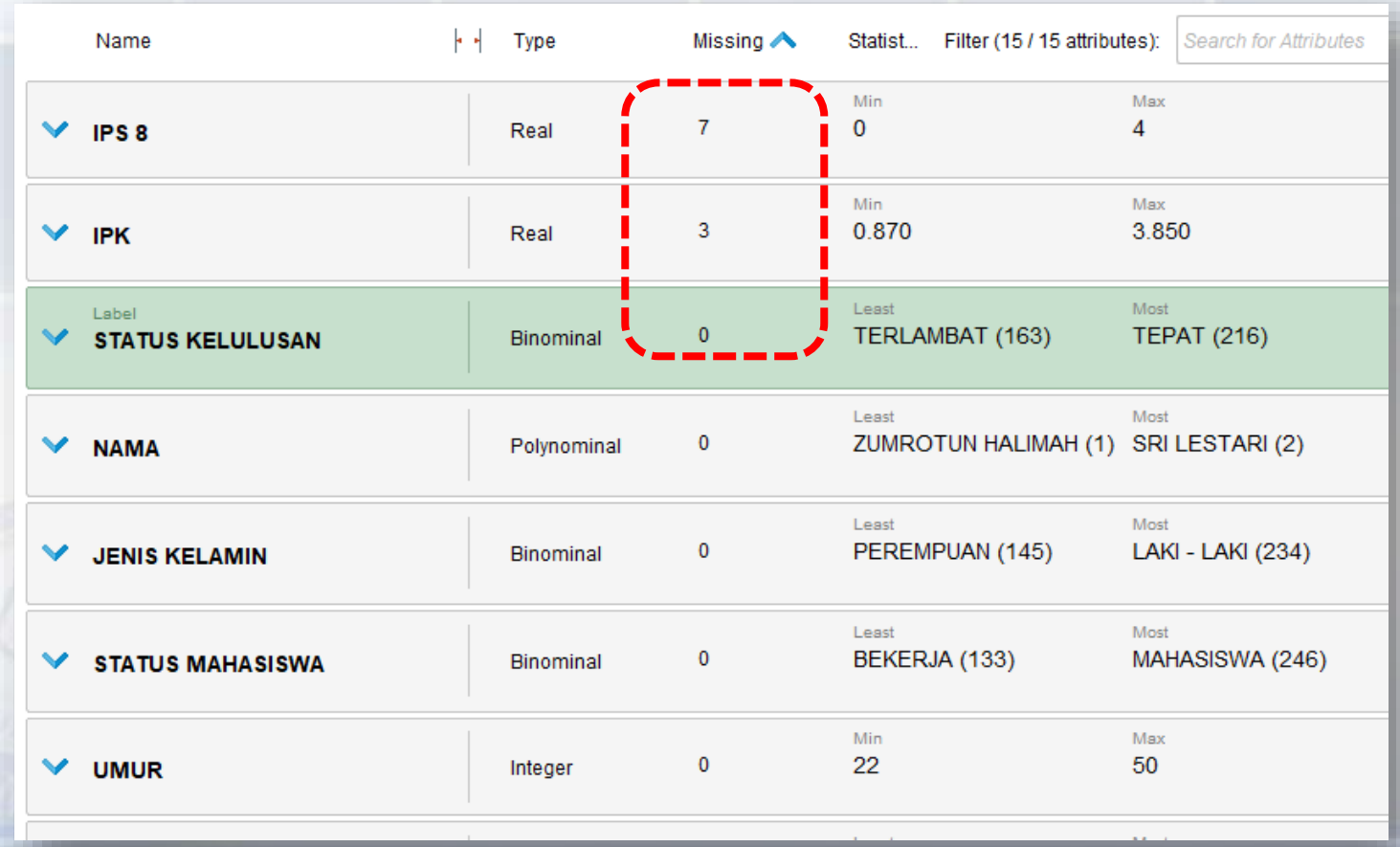
3. Data Preparation

Data set: datakelulusanmahasiswa.xls

Row No.	STATUS KEL...	NAMA	JENIS KELA...	STATUS MA...	UMUR	STATUS NIK...	IPS 1	IPS 2
1	TERLAMBAT	ANIK WIDAYA...	PEREMPUAN	BEKERJA	28	BELUM MENI...	2.760	2.800
2	TERLAMBAT	DWI HESTYN...	PEREMPUAN	MAHASISWA	32	BELUM MENI...	3	3.300
3	TERLAMBAT	MURYA ARIE...	PEREMPUAN	BEKERJA	29	BELUM MENI...	3.500	3.300
4	TERLAMBAT	NANIK SUSA...	PEREMPUAN	MAHASISWA	27	BELUM MENI...	3.170	3.410
5	TERLAMBAT	RIFKA ISTIQF...	PEREMPUAN	BEKERJA	29	BELUM MENI...	2.900	2.890
6	TERLAMBAT	SUHARYONO	LAKI - LAKI	BEKERJA	27	BELUM MENI...	2.950	2.820
7	TEPAT	FARIKHATUN...	PEREMPUAN	MAHASISWA	26	BELUM MENI...	2.760	3.140
8	TEPAT	FIFI SUNALISA	PEREMPUAN	MAHASISWA	27	BELUM MENI...	2.620	2.890
9	TERLAMBAT	HENDRIK M...	PEREMPUAN	BEKERJA	25	MENIKAH	3.600	3.540
10	TERLAMBAT	IMAM AGUNG...	PEREMPUAN	BEKERJA	28	BELUM MENI...	2.710	2.550
11	TERLAMBAT	IMAM SANTO...	PEREMPUAN	BEKERJA	27	BELUM MENI...	3.140	3.460
12	TERLAMBAT	IRFAN EKO ...	PEREMPUAN	BEKERJA	32	BELUM MENI...	2.670	2.300
13	TERLAMBAT	IWAN HAMBALI	PEREMPUAN	BEKERJA	26	BELUM MENI...	2.570	2.820
14	TERLAMBAT	M SYAIFULLAH	PEREMPUAN	BEKERJA	31	BELUM MENI...	2.710	3

3. Data Preparation

- Terdapat 379 data mahasiswa dengan 15 atribut
- Missing Value sebayak 10 data, dan tidak terdapat data noise



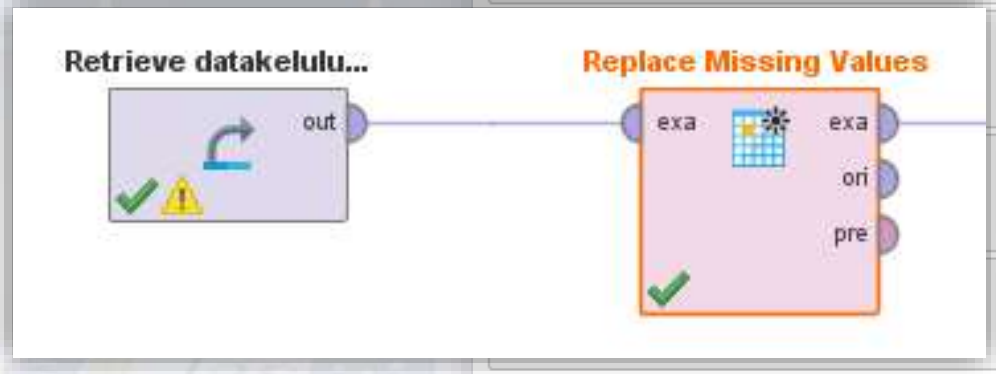
Name	Type	Missing	Statist...	Filter (15 / 15 attributes): Search for Attributes
IPS 8	Real	7	Min 0	Max 4
IPK	Real	3	Min 0.870	Max 3.850
Label STATUS KELULUSAN	Binominal	0	Least TERLAMBAT (163)	Most TEPAT (216)
NAMA	Polynominal	0	Least ZUMROTUN HALIMAH (1)	Most SRI LESTARI (2)
JENIS KELAMIN	Binominal	0	Least PEREMPUAN (145)	Most LAKI - LAKI (234)
STATUS MAHASISWA	Binominal	0	Least BEKERJA (133)	Most MAHASISWA (246)
UMUR	Integer	0	Min 22	Max 50

Versi : 01

3. Data Preparation

I.62

- Missing Value dipecahkan dengan menambahkan data dengan nilai rata-rata
- Hasilnya adalah data bersih tanpa missing value

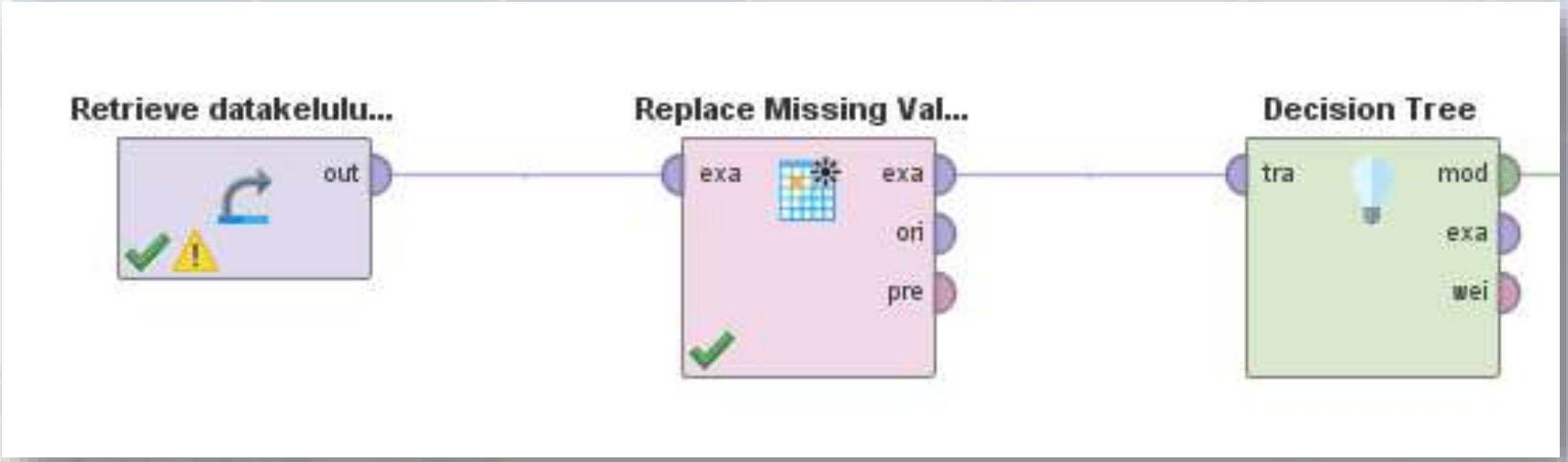


Name	Type	Missing	Statist...	Filter (15 / 15 attributes):
Label STATUS KELULUSAN	Binominal	0	Least TERLAMBAT (163)	Most TEPA
NAMA	Polynominal	0	Least ZUMROTUN HALIMAH (1)	Most SRI LE
	Binominal	0	Least PEREMPUAN (145)	Most LAKI -
	Binominal	0	Least BEKERJA (133)	Most MAHA
	Integer	0	Min 22	Max 50
STATUS NIKAH	Binominal	0	Least MENIKAH (8)	Most BELUM
IPS 1	Real	0	Min 0.330	Max 3.790

Versi : 01

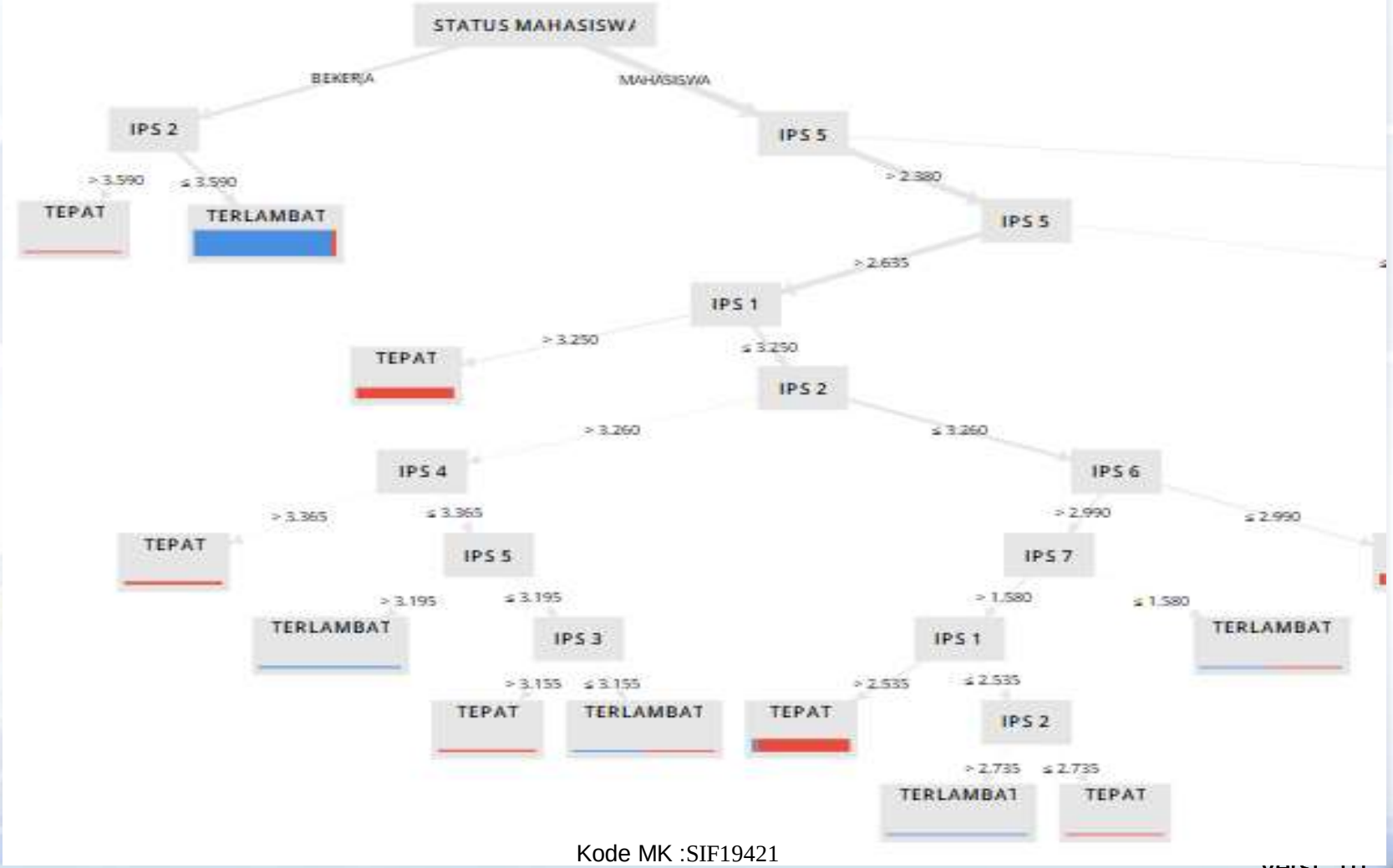
4. Modeling

- Modelkan dataset dengan Decision Tree
- Pola yang dihasilkan bisa berbentuk tree atau if-then



4. Modeling

Hasil pola dari data berupa berupa decision tree (pohon keputusan)



5. Evaluation

Hasil pola dari data berupa berupa peraturan if-then

```
STATUS MAHASISWA = BEKERJA
|   IPS 2 > 3.590: TEPAT {TERLAMBAT=0, TEPAT=2}
|   IPS 2 ≤ 3.590: TERLAMBAT {TERLAMBAT=127, TEPAT=4}
STATUS MAHASISWA = MAHASISWA
|   IPS 5 > 2.380
|   |   IPS 5 > 2.635
|   |   |   IPS 1 > 3.250: TEPAT {TERLAMBAT=0, TEPAT=50}
|   |   |   IPS 1 ≤ 3.250
|   |   |   |   IPS 2 > 3.260
|   |   |   |   |   IPS 4 > 3.365: TEPAT {TERLAMBAT=0, TEPAT=10}
|   |   |   |   |   IPS 4 ≤ 3.365
|   |   |   |   |   |   IPS 5 > 3.195: TERLAMBAT {TERLAMBAT=4, TEPAT=0}
|   |   |   |   |   |   IPS 5 ≤ 3.195
|   |   |   |   |   |   |   IPS 3 > 3.155: TEPAT {TERLAMBAT=0, TEPAT=5}
|   |   |   |   |   |   |   IPS 3 ≤ 3.155: TERLAMBAT {TERLAMBAT=1, TEPAT=1}
|   |   |   |   IPS 2 ≤ 3.260
|   |   |   |   |   IPS 6 > 2.990
|   |   |   |   |   |   IPS 7 > 1.580
|   |   |   |   |   |   |   IPS 1 > 2.535: TEPAT {TERLAMBAT=3, TEPAT=58}
|   |   |   |   |   |   |   IPS 1 ≤ 2.535
|   |   |   |   |   |   |   |   IPS 2 > 2.735: TERLAMBAT {TERLAMBAT=2, TEPAT=0}
|   |   |   |   |   |   |   |   IPS 2 ≤ 2.735: TEPAT {TERLAMBAT=0, TEPAT=2}
|   |   |   |   |   |   IPS 7 ≤ 1.580: TERLAMBAT {TERLAMBAT=1, TEPAT=1}
|   |   |   |   |   IPS 6 ≤ 2.990: TEPAT {TERLAMBAT=0, TEPAT=51}
|   |   IPS 5 ≤ 2.635
|   |   |   IPS 3 > 2.480
|   |   |   |   IPS 1 > 2.920: TEPAT {TERLAMBAT=0, TEPAT=5}
|   |   |   |   IPS 1 ≤ 2.920
|   |   |   |   |   IPS 3 > 3.075: TEPAT {TERLAMBAT=0, TEPAT=2}
|   |   |   |   |   IPS 3 ≤ 3.075: TERLAMBAT {TERLAMBAT=6, TEPAT=0}
|   |   |   IPS 3 ≤ 2.480: TEPAT {TERLAMBAT=0, TEPAT=11}
```

5. Evaluation

- Atribut atau faktor yang paling berpengaruh adalah Status Mahasiswa, IPS2, IPS5, IPS1

- Atribut atau faktor yang tidak berpengaruh adalah Nama, Jenis Kelamin, Umur, IPS6. IPS7, IPS8

6. Deployment

- Budi membuat program peningkatan disiplin dan pendampingan ke mahasiswa di semester awal (1-2) dan semester 5, karena faktor yang paling menentukan kelulusan mahasiswa ada di dua semester itu
- Budi membuat peraturan melarang mahasiswa bekerja paruh waktu di semester awal perkuliahan, karena beresiko tinggi di kelulusan tepat waktu
- Budi membuat program kerja paruh waktu di dalam kampus, sehingga banyak pekerjaan kampus yang bisa intens ditangani, sambil mendidik mahasiswa supaya memiliki pengalaman kerja. Dan yang paling penting mahasiswa tidak meninggalkan kuliah karena pekerjaan
- Budi memasukkan pola dan model yang terbentuk ke dalam sistem informasi kademik, dimana sistem dibuat cerdas, sehingga bisa mengirimkan email analisis pola secara otomatis ke mahasiswa sesuai profilnya

Latihan

- Pahami dan lakukan eksperimen berdasarkan seluruh studi kasus yang ada di buku Data Mining for the Masses (*Matthew North*)

- Pahami bahwa metode CRISP-DM membantu kita memahami penggunaan metode data mining yang lebih sesuai dengan kebutuhan organisasi

Tugas Menyelesaikan Masalah Organisasi

- Analisis masalah dan kebutuhan yang ada di organisasi lingkungan sekitar anda
- Kumpulkan dan review dataset yang tersedia, dan hubungkan masalah dan kebutuhan tadi dengan data yang tersedia (analisis dari 5 peran data mining)
  - Bila memungkinkan pilih beberapa peran sekaligus untuk mengolah data tersebut, misalnya: lakukan association (analisis faktor), sekaligus estimation atau clustering
- Lakukan proses CRISP-DM untuk menyelesaikan masalah yang ada di organisasi sesuai dengan data yang didapatkan
  - Pada proses data preparation, lakukan data cleaning (replace missing value, replace, filter attribute) sehingga data siap dimodelkan
  - Lakukan juga komparasi algoritma untuk memilih algoritma terbaik

Review dan Latihan

☺ END ☺