Course Definition

- Kursus Merekonstruksi Data ini adalah bagian dari Data Preparation, dan merupakan lanjutan dari modul 7.
- Data Preparation yang dibahasa adalah transformasi data yaitu:*
  - Representasi Fitur, dan
  - Rekayasa Fitur
- Ada beberapa teknik transformasi data yang digunakan sesuai kebutuhan dan ketersediaan/jenis data baik numerik maupun kategorik

Learning Objective

Dalam kursus ini diharapkan:

A. Peserta mampu menganalisis teknik transformasi data

B. Peserta mampu melakukan transformasi data

C. Peserta mampu membuat dokumentasi konstruksi data

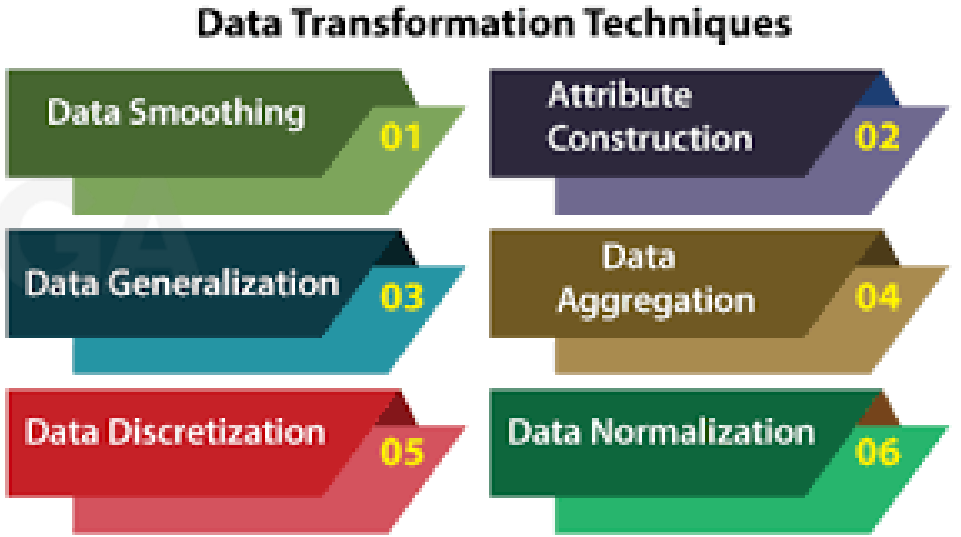
Data Transformation

- Representasi Fitur atau Pembelajaran Fitur:
  - Teknik-Teknik yang memungkinkan sistem bekerja otomatis menemukan representasi yang diperlukan (untuk deteksi fitur atau klasifikasi dari dataset),
  - menggantikan rekayasa fitur manual, dan
  - memungkinkan mesin mempelajari fitur dan menggunakannya untuk melakukan tugas tertentu.

- Rekayasa Fitur:
  - Proses mengubah data mentah menjadi fitur yang:
    - Mewakili masalah mendasar ke model prediktif,
    - menghasilkan akurasi model yang lebih baik pada data yang tidak terlihat.

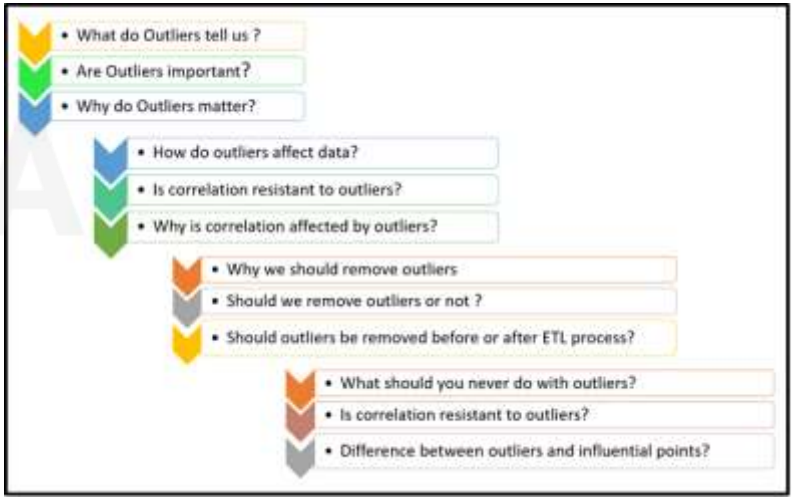
Pengertian Transformasi Data

- Transformasi Data adalah
  memformat, menskalakan
  mentah (dari format sum
  yang diperlukan, baik itu
  atau model pembelajaran



Mengatasi (*Handling*) *Outlier* (HO)

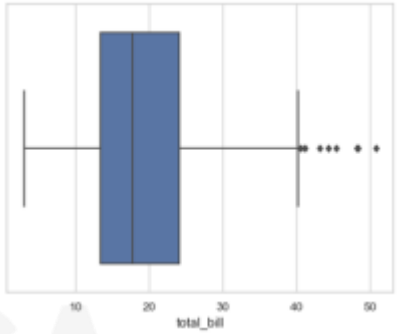
- Definisi *Outlier*:
  - Titik data yang sangat berbeda dari data lainnya.
  - Pengamatan yang menyimpang dari pola keseluruhan pada sampel.
- Penyebab:
  - Error percobaan, salah input, error instrumen, kesengajaan (untuk pengujian), error pemrosesan data, error sampling, kewajaran karena keanehan dalam data (bukan error).

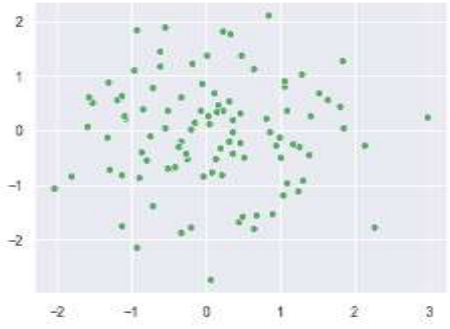


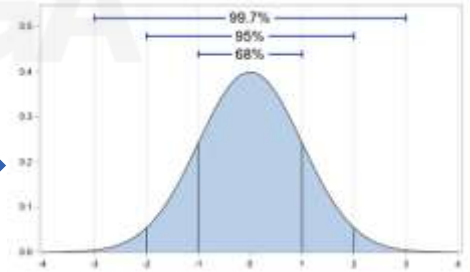
sumber gbr:
https://heartbeat.fritz.ai/hands-on-with-feature-engineering-techniques-dealing-with-outliers-fcc9f57cb63b
https://www.analyticsvidhya.com/blog/2021/03/zooming-out-a-look-at-outlier-and-how-to-deal-with-them-indata-science/

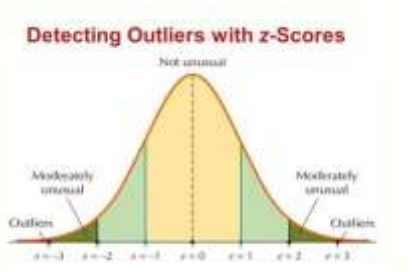
Deteksi Outlier

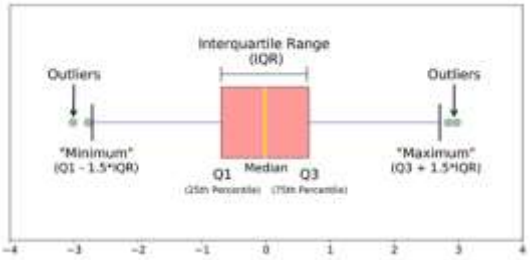
- Visualiasi dgn Boxplot dan Scatterplot
  - Sebagian besar titik data terletak di tengah, tetapi ada satu titik yang jauh dari pengamatan lainnya; ini bisa menjadi outlier.
- Distribusi Normal
  - Dalam distribusi normal, sekitar 99,7% data berada dalam tiga standar deviasi dari mean.
  - Jika ada pengamatan yang lebih dari tiga kali standar deviasi, kemungkinan itu adalah outlier.
- Z-score
- Inter Quantile Range (IQR)











sumber gbr:
https://heartbeat.fritz.ai/hands-on-with-feature-engineering-techniques-dealing-with-outliers-fcc9f57cb63b
https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/

Data Transformation

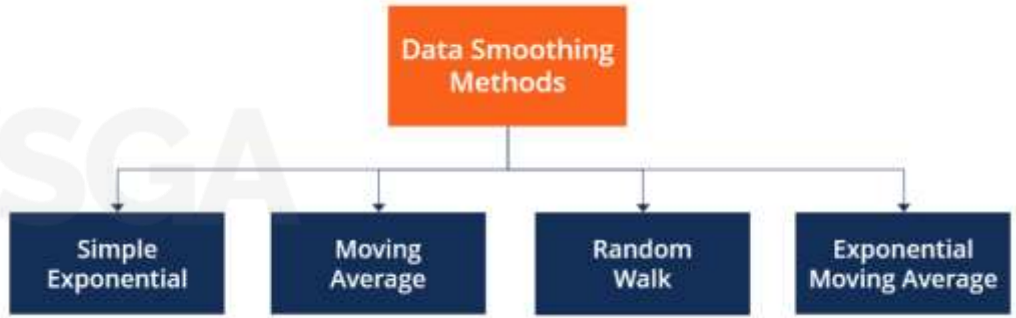
- Representasi Fitur atau Pembelajaran Fitur:
  - Teknik-Teknik yang memungkinkan sistem bekerja otomatis menemukan representasi yang diperlukan (untuk deteksi fitur atau klasifikasi dari dataset),
  - menggantikan rekayasa fitur manual, dan
  - memungkinkan mesin mempelajari fitur dan menggunakannya untuk melakukan tugas tertentu.

- Rekayasa Fitur:
  - Proses mengubah data mentah menjadi fitur yang:
    - Mewakili masalah mendasar ke model prediktif,
    - menghasilkan akurasi model yang lebih baik pada data yang tidak terlihat.

Smoothing

- Smoothing (penghalusan) data mengacu pada pendekatan statistik untuk menghilangkan outlier dari kumpı

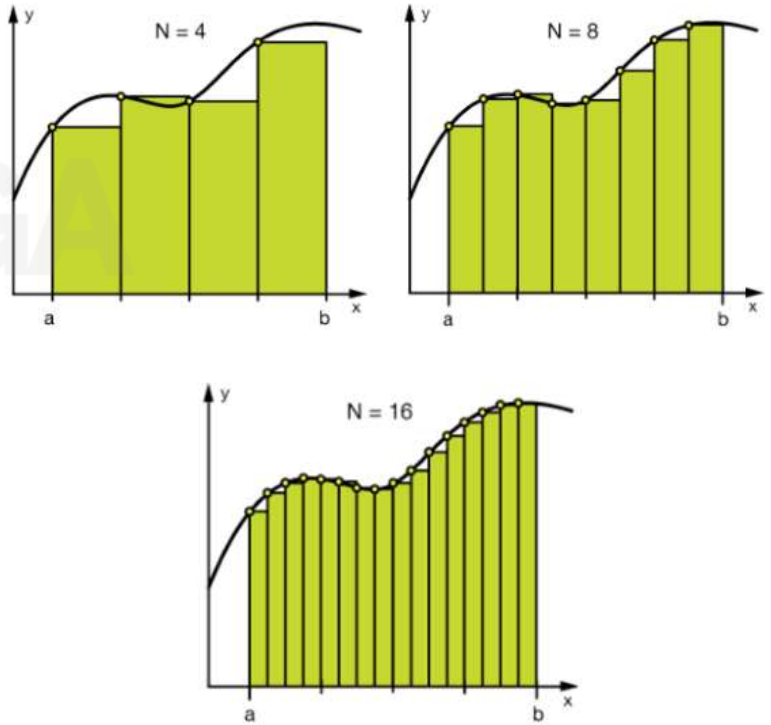
- Teknik Smoothing
  - Simple Exponential
  - Moving Average
  - Random Walk
  - Exponential Moving Average



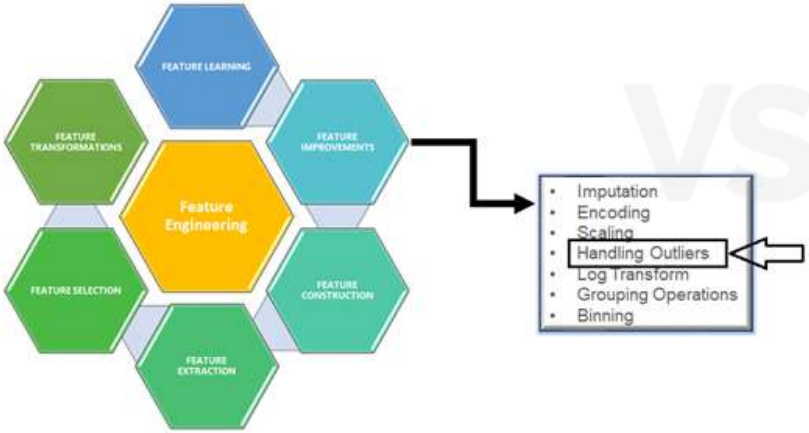
Binning

- Pro:
  - Dapat diterapkan pada data kategorik dan numerik.
  - Model lebih robust dan mencegah overfitting.
- Kontra:
  - Meningkatnya biaya kinerja perhitungan.
  - Mengorbankan informasi.
  - Untuk kolom data numerik, dapat menyebabkan redudansi untuk beberapa algoritma.
  - Untuk kolom data kategorik, label dengan frekuensi rendah berdampak negatif pada robustness model statistik.
  - Untuk ukuran data dengan 100 ribu baris, disarankan menggabungkan label/kolom dengan record yang < 100 menjadi kategori baru, misal “Lain-lain”.

Ilustrasi binning untuk data numerik



Rekayasa Fitur



Outlier

Mengatasi (*Handling*) *Outlier* (HO)

- Jenis/kategori:
  - Univariate vs multivariate
  - Parametrik vs non-parametrik
- Deteksi dan Cari Outlier dengan:
  - Visualisasi
  - Distribusi normal

Teknik Mengatasi Outlier:

- Trimming
- Winsorizing
- Imputing
- Discretization
- Censoring
- Z-score
- Linear Regression Model

Teknik HO: Trimming (Pangkas) vs Winsorizing

- Nama lain: Truncation (Potong)
- Definisi: Menghapus outlier dari dataset
- Perlu memutuskan metrik untuk menentukan outlier.

Definisi:
Mengganti outlier dari dataset dengan nilai persentil setiap ujung/batas atas dan bawah.

Trimming vs Winsorizing

- Contoh kasus: laporan jumlah pasien yang ditangani tiap dokter/bulan (di bawah 100 pasien), dengan 4% dilaporkan lebih dari 100 pasien.

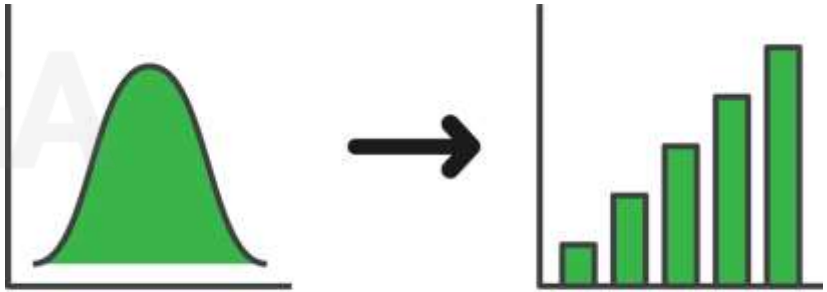
range [5,100]. Outlier tidak dibuang, namun dimasukan ke dalam range terdekat. Sehingga nilai mean mengecil. Median dan N tidak berubah.

Membuang data yang berada di luar range. Nilai N berubah, mean mengecil, median masih dinilai yg sama.

sumber gbr: https://protobi.com/post/extreme-values-winsorize-trim-or-retain

Discretization

- Definisi:
  - Proses mengubah fungsi, model dan variabel kontinu menjadi diskret (data kontinu di *Ukur* (measured) vs data kontinu di *Hitung* (counted)).
- Nama lain: Binning.
- Dasar Pertimbangan:
  - Data kontinu memiliki derajat kebebasan (DoF) yang tak hingga.
  - Data kontinu lebih mudah dipahami dan disimpan dalam bentuk kategori/grup
    - misal berat badan < 65 kg (ringan); 65 – 80 kg (mid); > 80 kg (berat).

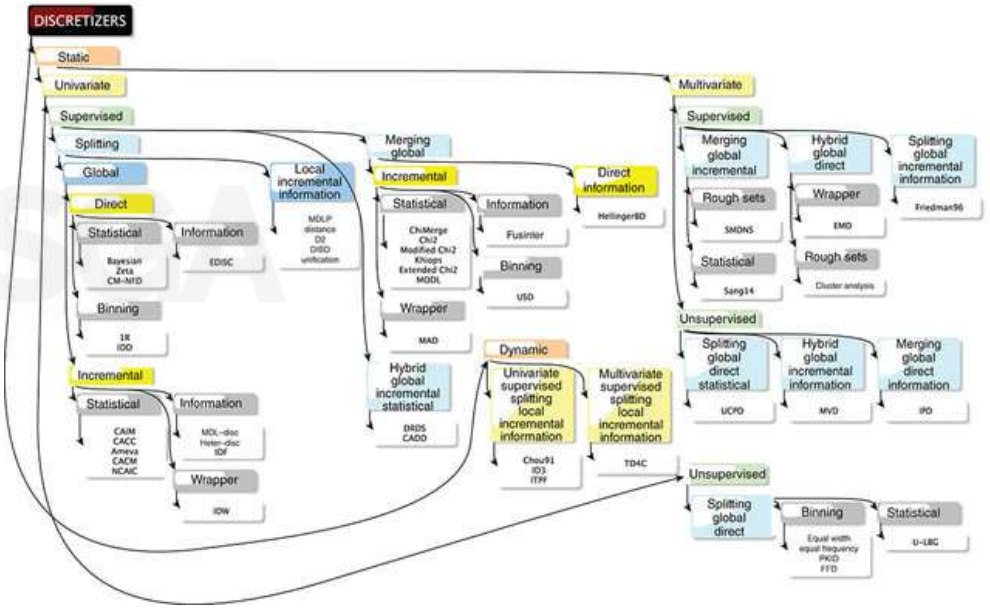


Discretization Process

Discretization

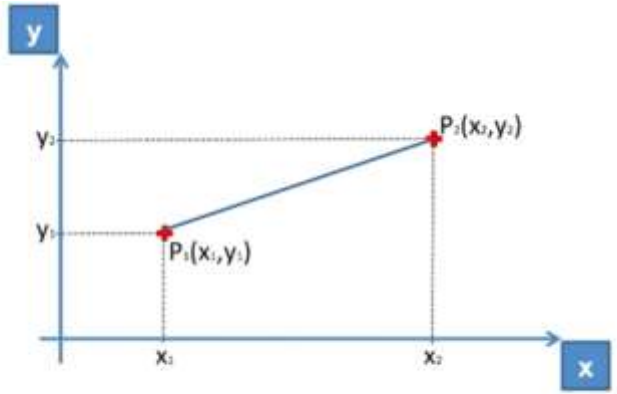
- Jenis:
  - Supervised
    - Decision Tree.
  - Unsupervised
    - Equal-width discretization.
    - Equal-frequency discretization.
    - K-means discretization.
  - Lainnya
    - Custom discretization.

Taksonomi Discretization



Scaling (Penskalaan)

- Dasar:
  - Sering diabaikan oleh pemula di Data Science.
  - Data numerik (biasanya) tidak memiliki range. range "Usia" vs range "Gaji" tidak sama (karakteristik berbeda). *Usia* memiliki rentang dari 1 sampai 150 (dalam tahun), sedangkan *Gaji* memiliki rentang dari 10 ribu sampai 100 ribu (dalam dolar). Untuk itu membandingkan perlu *scaling*.
  - Beberapa algoritma Machine Learning (regresi linear dan logistik dan Neural Network; SVM, KNN, K-means; LDA; PCA) yang menggunakan teknik optimasi Euclidian Distance 2 poin (titik).



Euclidean Distance between P_1 and $P_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

- Dengan menggunakan rumus Euclidean Distance diatas, maka jelas bahwa hasil perhitungan pada kolom *Usia* dan *Gaji* akan memiliki jarak (distance) yang sangat jauh. Disinilah proses Feature Scaling dibutuhkan.
- Feature Scaling adalah suatu cara untuk membuat numerical data pada dataset memiliki rentang nilai (scale) yang sama. Tidak ada lagi satu variabel data yang mendominasi variabel data lainnya.

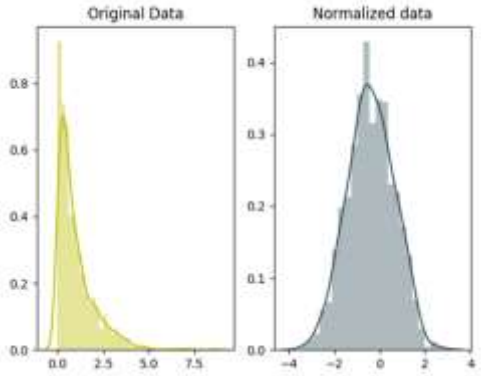
sumber gbr: https://medium.com/machine-learning-id/melakukan-feature-scaling-pada-dataset-229531bb08de

Scaling: Jenis

Standardisation	Normalisation
$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$	$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$

Scaling: Normalisasi

- Nama Lain: Min-Max Scaling.
- Definisi: Teknik penskalaan di mana nilai-nilai digeser dan diubah skalanya sehingga nilainya berkisar antara 0 dan 1 (rentang [0,1]).



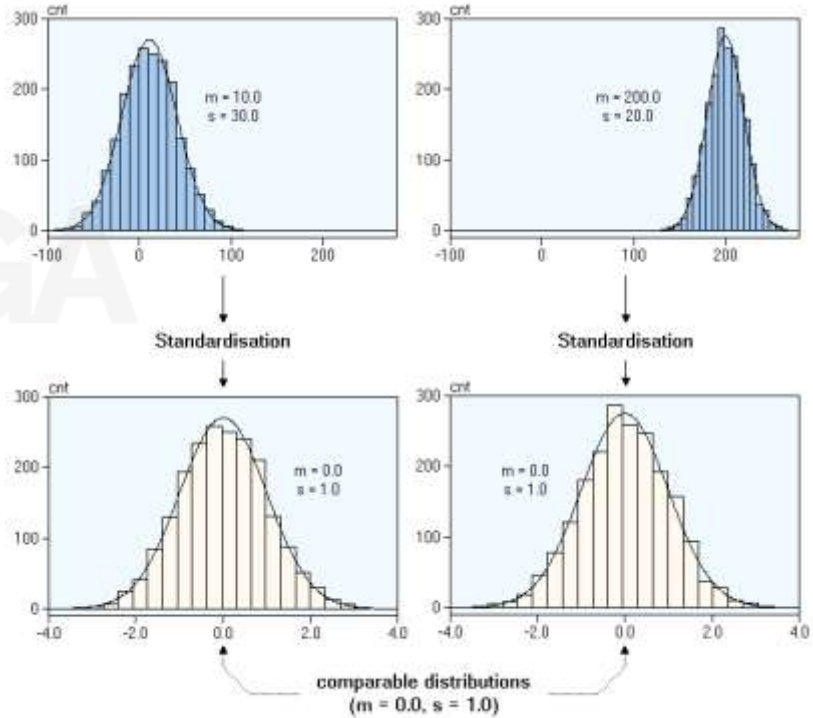
$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Di sini, Xmax dan Xmin masing-masing adalah nilai maksimum dan minimum dari fitur.

- Ketika nilai X adalah nilai minimum dalam kolom, pembilangnya adalah 0, dan karenanya X' adalah 0.
- Sebaliknya, ketika nilai X adalah nilai maksimum dalam kolom, pembilangnya sama dengan penyebutnya sehingga nilai X' adalah 1.
- Jika nilai X berada di antara nilai minimum dan maksimum, maka nilai X' berada di antara 0 dan 1.

Scaling: Standardisasi

- Tujuan: Berfokus pada mengubah data mentah menjadi informasi yang dapat digunakan sebelum dianalisis.
- Definisi: Teknik yang menskalakan data shingga memiliki mean = 0 dan standar deviasi =1
- Kontra:
  - Menambah langkah dalam data preparation
  - Waktu bertambah



sumber:
https://www.cnblogs.com/quinn-yann/p/9808247.html

Contoh Kasus *Scaling*

	Country	Age	Salary	Purchased
1	France	44	72000	No
2	Spain	27	48000	Yes
3	Germany	30	54000	No
4	Spain	38	61000	No
5	Germany	40		Yes
6	France	35	58000	Yes
7	Spain		52000	No
8	France	48	79000	Yes
9	Germany	50	83000	No
10	France	37	67000	Yes

The range of Age: 27 - 50

The range of Salary:48,000 - 83,000

```
dataset['Age'].min()
```

```
27.0
```

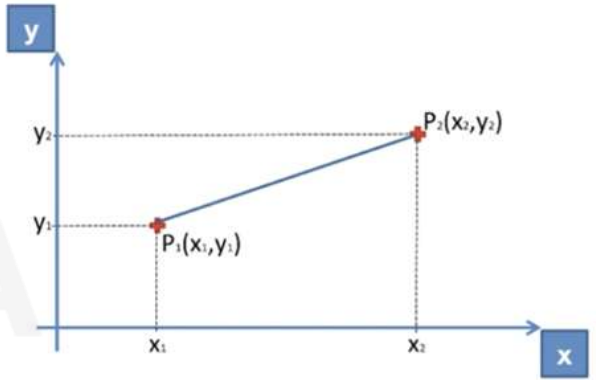
```
dataset['Salary'].min()
```

```
48000.0
```

```
dataset['Age'].max()
```

```
50.0
```

```
dataset['Salary'].max()
```



Euclidean Distance between P_1 and $P_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

Let x be the no. of Salary and y be the no. of Age

Example: x1&y1 are in row 2, x2&y2 are in row 9

(x2-x1)^2= (83000-48000)^2

=1225000000

(y2-y1)^2=(50-27)^2

=529

sumber gbr:
https://www.kdnuggets.com/2020/04/data-transformation-standardization-normalization.html

Contoh Kasus *Scaling*

- Ketika kita menghitung persamaan jarak (distance) Euclidean, jumlah $(x2-x1)^2$ jauh lebih besar daripada jumlah $(y2-y1)^2$ yang berarti jarak Euclidean akan didominasi oleh *Gaji* jika kita tidak menerapkan penskalaan. Perbedaan *Usia* berkontribusi lebih sedikit terhadap perbedaan keseluruhan.
- Oleh karena itu, kita harus menggunakan penskalaan untuk membawa semua nilai ke besaran yang sama dan dengan demikian, menyelesaikan masalah ini.

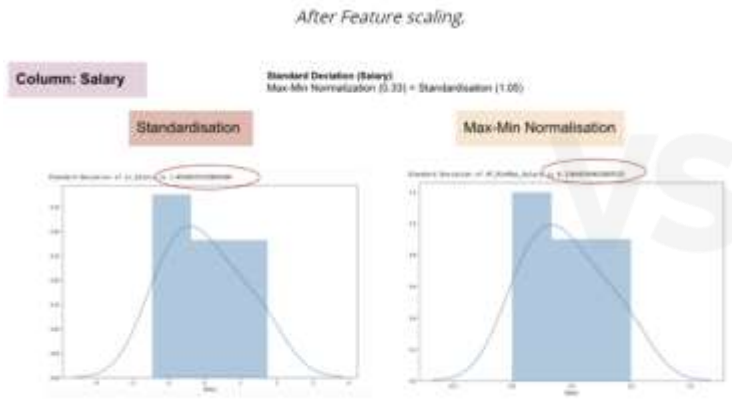
Standardisation

	Age	Salary
0	0.758874	7.494733e-01
1	-1.711504	-1.438178e+00
2	-1.275555	-8.912655e-01
3	-0.113024	-2.532004e-01
4	0.177609	6.632192e-16
5	-0.548973	-5.266569e-01
6	0.000000	-1.073570e+00
7	1.340140	1.387538e+00
8	1.630773	1.752147e+00
9	-0.258340	2.937125e-01

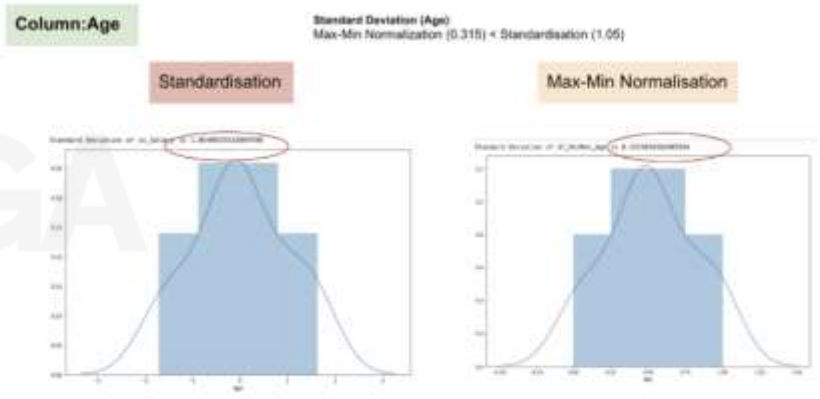
Max-Min Normalization

	Age	Salary
0	0.739130	0.685714
1	0.000000	0.000000
2	0.130435	0.171429
3	0.478261	0.371429
4	0.565217	0.450794
5	0.347826	0.285714
6	0.512077	0.114286
7	0.913043	0.885714
8	1.000000	1.000000
9	0.434783	0.542857

Contoh Kasus *Scaling*



Normal distribution and Standard Deviation of Salary.



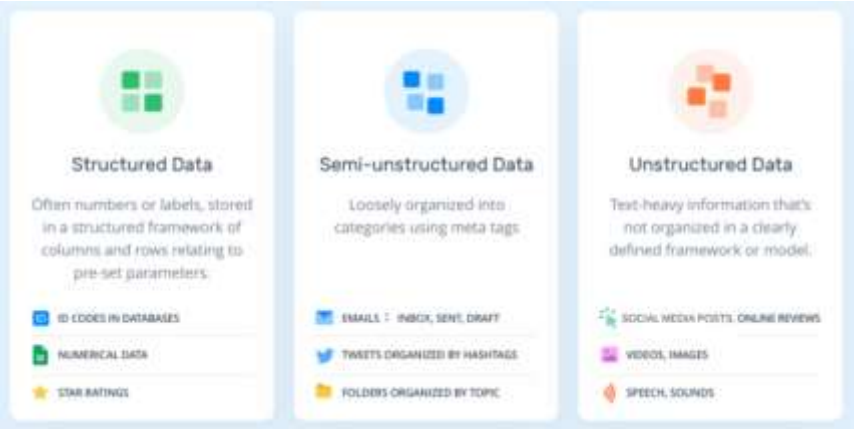
Normal distribution and Standard Deviation of Age.

sumber gbr:
https://www.kdnuggets.com/2020/04/data-transformation-standardization-normalization.html

Rekayasa Fitur untuk Data Tidak Terstruktur

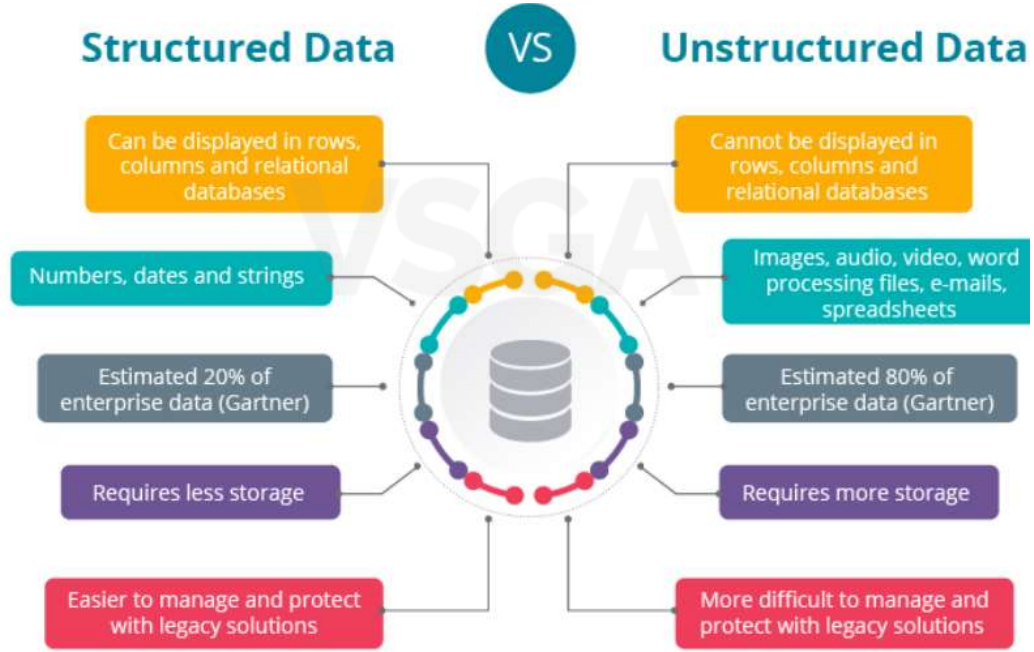
- Jenis *Unstructured Data*:
  - Teks
  - Grafik
  - Video
  - Audio

- Manfaat utk Bisnis:
  - Meningkatkan pengalaman pelanggan.
  - Menemukan celah di pasar & berinovasi.
  - 'Mendengarkan' pelanggan Anda.



sumber:
https://monkeylearn.com/unstructured-data/

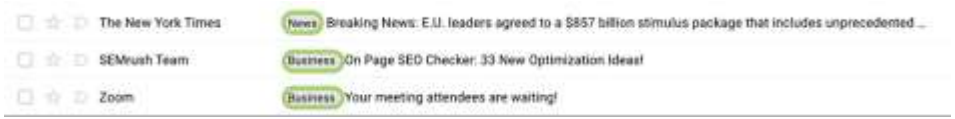
Structured vs Unstructured Data



sumber:
https://www.knowledgehut.com/blog/data-science/role-of-unstructured-data-in-data-science

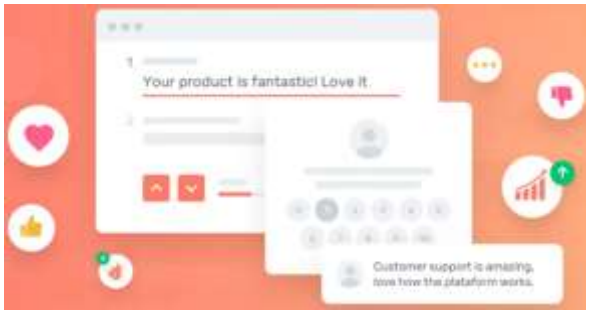
Contoh Unstructured Data

email

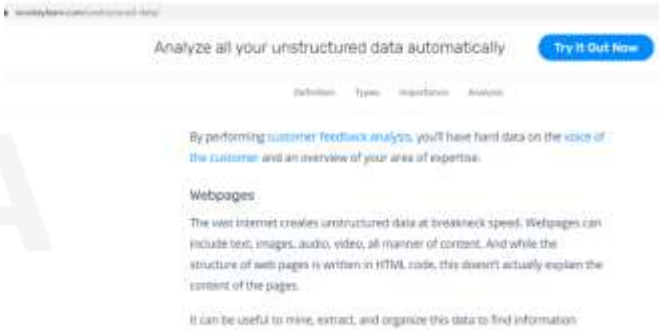


email: semi-structured data; isi email: unstructured data

review pelanggan



webpages

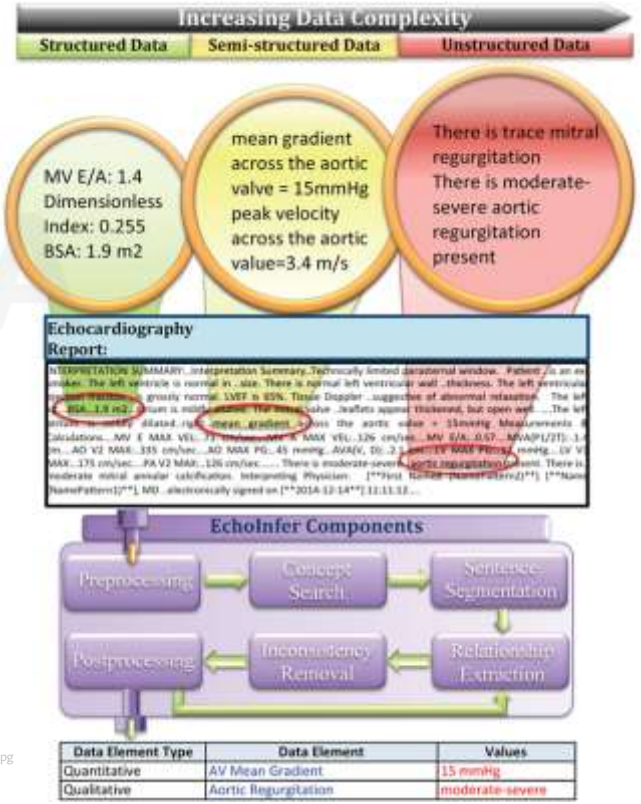


sosmed

sumber:
https://monkeylearn.com/unstructured-data/

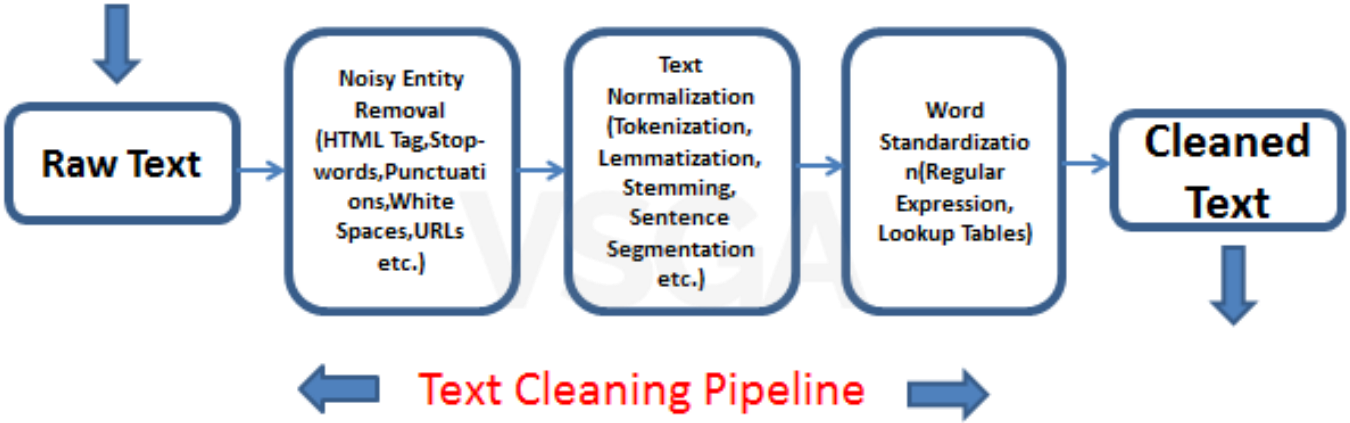
Rekayasa Fitur untuk Teks Tidak Terstruktur

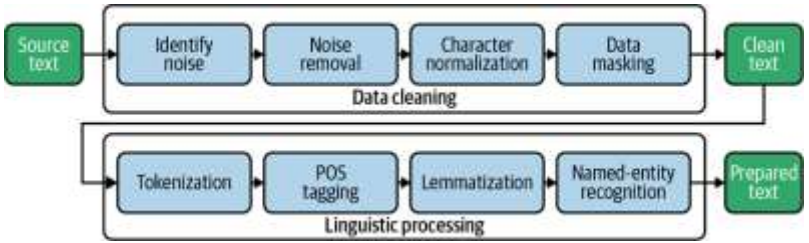
- *Text Pre-processing*:
  - Removing special characters
  - Removing tags
  - Converting to lowercase
  - Contraction expansion
  - Removing stopwords
  - Correcting spellings
  - Stemming
  - Lemmatization
  - Dll.



sumber: ...ues-into-structured-format-from-structured_W640.jpg

Tahapan Text Processing





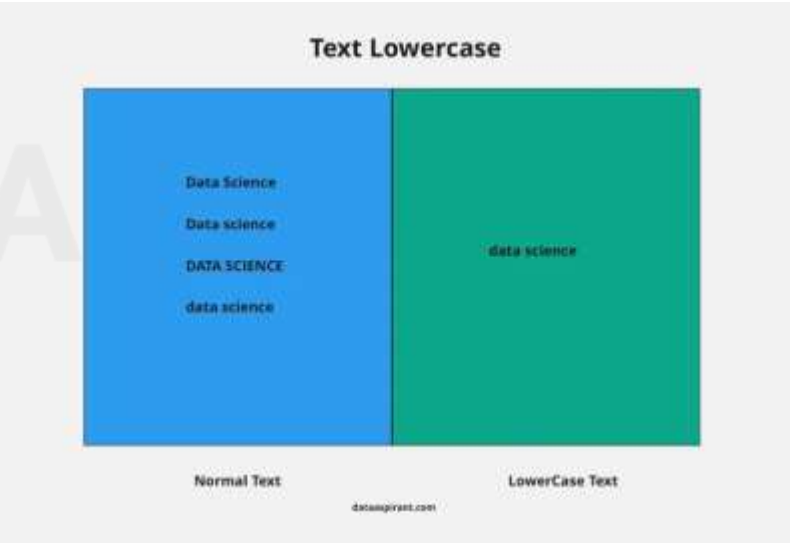
sumber:
https://miro.medium.com/max/748/1*zfGwHSTGqbaFwb_O8KQrnQ.png
https://www.oreilly.com/library/view/blueprints-for-text/9781492074076/assets/btap_0401.png

Aplikasi untuk Unstructured Data

- Tableau
- MonkeyLearn
- Apache Spark
- Amazon AWS
- SAS
- Python
- Microsoft Azure
- IBM Cloud
- IBM Cloud
- RapidMiner
- KNIME
- QlikView
- R programming
- MS. Excel

Converting to Lowercase

- Definisi: Mengubah semua teks ke dalam huruf kecil (lowercase).
- Pro:
  - Simple dan paling efektif.
  - Efektif untuk kasus yang bergantung pada frekuensi kata, mis. klasifikasi dokumen.
- Kontra:
  - Sebaiknya dilakukan di awal tahapan.
  - Tidak cocok untuk pengenalan *Parts-Of-Speech* (POS) tag atau *Dependency Parsing*.



sumber: https://dataaspirant.com/nlp-text-preprocessing-techniques-implementation-python/

Removal of HTML tags

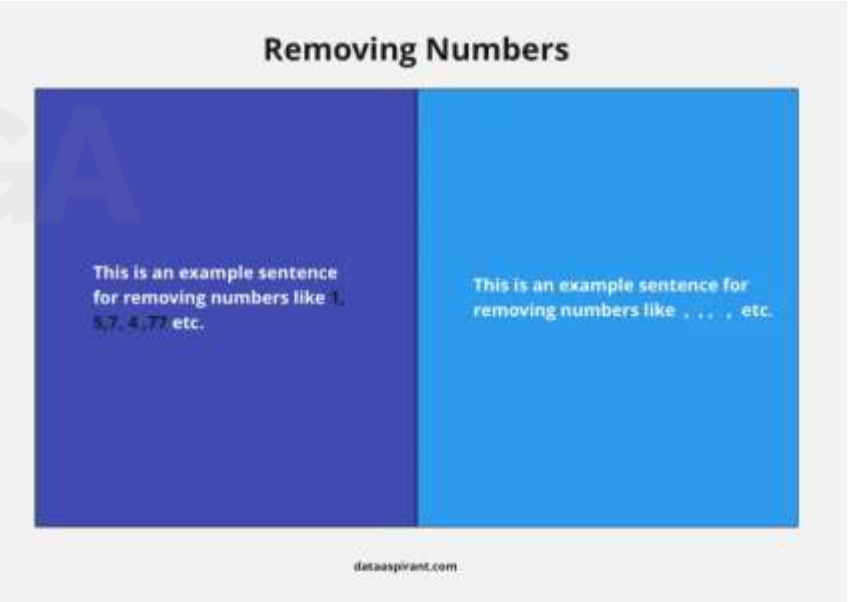
- Pro:
  - Tag html tidak memberikan informasi berharga utk analisis DS.
  - Menghilangkan data text dengan RegEx atau modul BeautifulSoup (library bs4).

sumber: https://dataaspirant.com/nlp-text-preprocessing-techniques-implementation-python/

Removal of URLs and Numbers

Removal of URLs

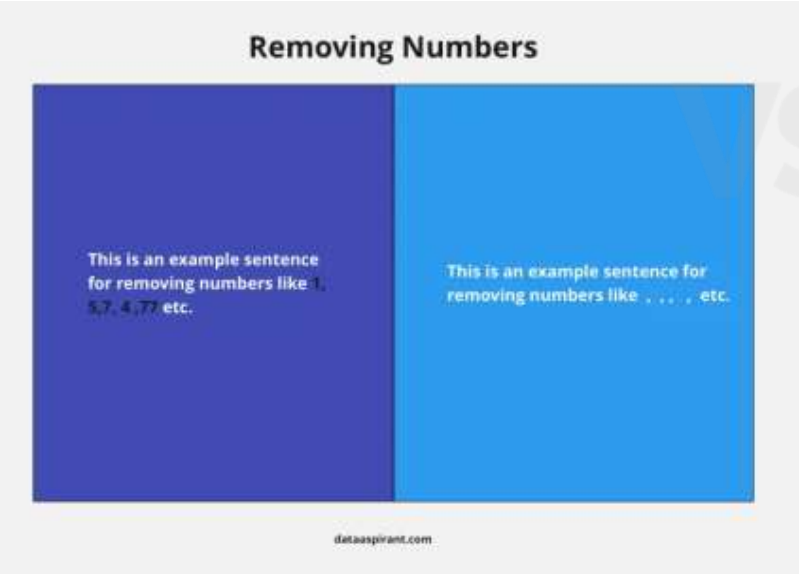
Removing Numbers



sumber:
https://dataaspirant.com/nlp-text-preprocessing-techniques-implementation-python/

Converting Number and Spelling Check

Converting numbers to words



Spelling correction

sumber:
https://dataaspirant.com/nlp-text-preprocessing-techniques-implementation-python/

Convert Chat and Accented Characters

Converting chat conversion words to normal words

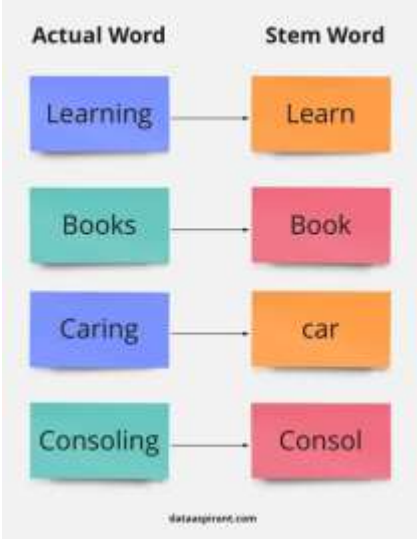
Convert accented characters to ASCII characters

sumber:
https://dataaspirant.com/nlp-text-preprocessing-techniques-implementation-python/

Expanding Contraction and Stemming

Expanding Contraction

Stemming



sumber:
https://dataaspirant.com/nlp-text-preprocessing-techniques-implementation-python/

Lemmatization and Removing Emoji/Emoticon

Lemmatization

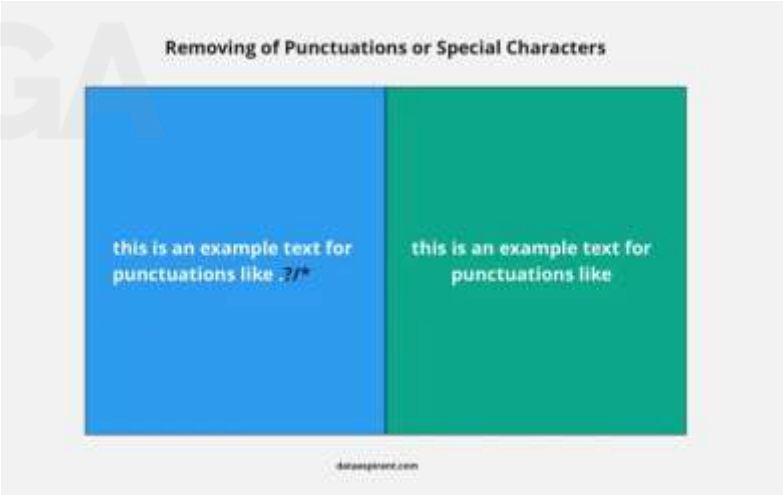
Removing Emoji

sumber:
https://dataaspirant.com/nlp-text-preprocessing-techniques-implementation-python/

Removing Stopwords and Punctuation

Removing Stopwords

Removing of Punctuations or Special Characters



sumber:
https://dataaspirant.com/nlp-text-preprocessing-techniques-implementation-python/

Teknik Rekayasa Fitur (Transformasi) utk Text Unstructured

- Setelah text preprocessing (text cleaning), lanjutkan dengan feature engineering untuk transformasi data teks tersebut.
- Teknik:
  - Model Bag of Words
  - Model Bag of N-Grams
  - Model TF-IDF
  - Document Similarity
  - Topic Models
  - Dll.

Model Bag of Words

- Komputer hanya mengenal angka bukan text, mk
- perlu konversi text ke vektor
- Operasi matematika BoW mengubah text ke vector.
- Limitasi:
  - BoW tidak mempertimbangkan semantik. Misal, *laki-laki* dan *pria*
  - Tidak bisa membedakan pernyataan vs pertanyaan, mis. *Barang ini bagus* vs *Baguskah barang ini?*

contoh ilustrasi *BoW*

dokumen review

1

```
'Movie is good and movie is worth watch'
'Movie is average but story is really good'
'I like the movie and the fight'
```

setelah di text preprocessing

2

```
'movie good movie worth watch'
'movie average story really good'
'i like movie fight'
```

koleksi kata-kata (words) dari review

3

```
['movie', 'good', 'movie', 'worth', 'watch']
['movie', 'average', 'story', 'really', 'good']
['I', 'like', 'movie', 'fight']
```

Model Bag of Words

Perlakukan setiap kalimat sebagai dokumen terpisah dan buat daftar semua kata dari keempat dokumen (tidak termasuk punctuation dan pengulangan

4

```
'movie', 'good', 'worth', 'watch','average', 'story', 'really', 'I',
'like', 'fight'
```

Buat vektor untuk review pertama:
”movie is good and movie is worth watching”

5

```
'movie' : 2
'good' : 1
'worth': 1
'watch' : 1
'average' : 0
'story': 0
'really': 0
'I': 0
'like': 0
'fight' : 0
```

Vektor untuk setiap review

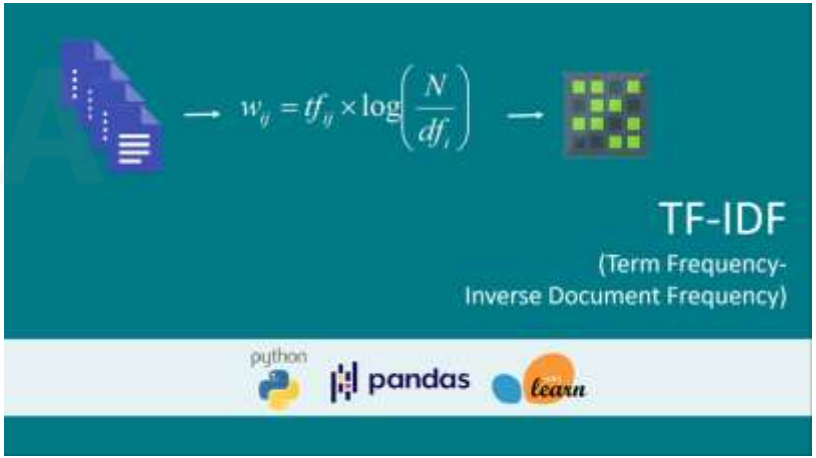
6

```
'movie good worth watch'' = [2,1,1,1,0,0,0,0,0,0]

'movie average story really good' =  [1,1,0,0,1,1,1,0,0,0]

'I like movie fight' = [1,0,0,0,0,0,0,1,1,1]
```

Setiap kata atau token dalam vektor di atas disebut “gram”

Model TF-IDF

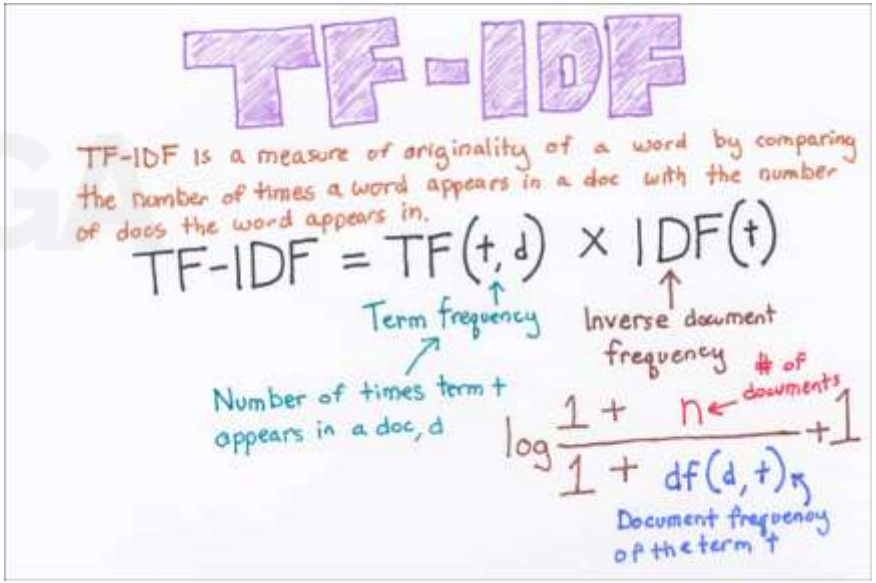
- Singkatan dari: *Term Frequency-Inverse Document Frequency.*
- Definisi: Ukuran statistik yang mengevaluasi seberapa relevan sebuah kata dengan dokumen dalam kumpulan dokumen.
- Perkalian dua metrik: *Berapa kali sebuah kata muncul dalam sebuah dokumen x frekuensi dokumen terbalik (invers) dari kata tersebut di seluruh kumpulan dokumen.*



sumber:
https://miro.medium.com/max/1400/1*qgta3eBYLAkTzmW6b-17sQ.png

Model TF-IDF

- TF dari kata: Frekuensi kata dalam dokumen.
  - Cara perhitungan: hitung langsung kata yang sering muncul atau berdasarkan panjang dokumen.
- IDF: Seberapa umum atau jarang sebuah kata di seluruh kumpulan dokumen.
  - Cara perhitungan: mengambil jumlah total dokumen, membaginya dengan jumlah dokumen yang berisi kata, dan menghitung logaritma.
  - Jika kata biasa dan sering muncul maka nilainya 0, jika tidak bernilai 1.



sumber:
https://towardsdatascience.com/tf-term-frequency-idf-inverse-document-frequency-from-scratch-in-python-6c2b61b78558

Referensi

- https://www.ucl.ac.uk/population-health-sciences/sites/population-health-sciences/files/quartagno_1.pdf
- https://rianneschouten.github.io/missing_data_science/assets/blogpost/blogpost.html
- https://towardsdatascience.com/tf-term-frequency-idf-inverse-document-frequency-from-scratch-in-python-6c2b61b78558
- https://dataaspirant.com/nlp-text-preprocessing-techniques-implementation-python
- https://www.oreilly.com/library/view/blueprints-for-text/9781492074076/assets/btap_0401.png
- https://monkeylearn.com/unstructured-data
- https://medium.com/machine-learning-id/melakukan-feature-scaling-pada-dataset-229531bb08de
- https://protobi.com/post/extreme-values-winsorize-trim-or-retain
- https://heartbeat.fritz.ai/hands-on-with-feature-engineering-techniques-dealing-with-outliers-fcc9f57cb63b
- https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/

Tools / Lab Online

- Jupyter Notebook
- Google Collabs

Summary

- Transformasi Data adalah bagian dari Data Preparation
- Membutuhkan pengetahuan dasar dan detail serta waktu yang mayoritas untuk menjamin data yang akan dianalisis sebersih mungkin
- Transformasi data dapat menggunakan beberapa teknik rekayasa fitur (feature engineering)
- Normalisasi, Standardisasi adalah bagian proses atau tahapan yang diperlukan untuk mentransformasi data
- Selain data terstruktur, transformasi data juga krusial dilakukan untuk data yang semi terstruktur dan tidak terstruktur (unstructured) seperti teks, image, audio dan video
- Data tidak terstruktur menjadi mayoritas dan tantangan sendiri dalam proses transformasi data di Data Science

Contoh Penerapan Transformasi Data - Binning

Menggunakan Rapid Miner

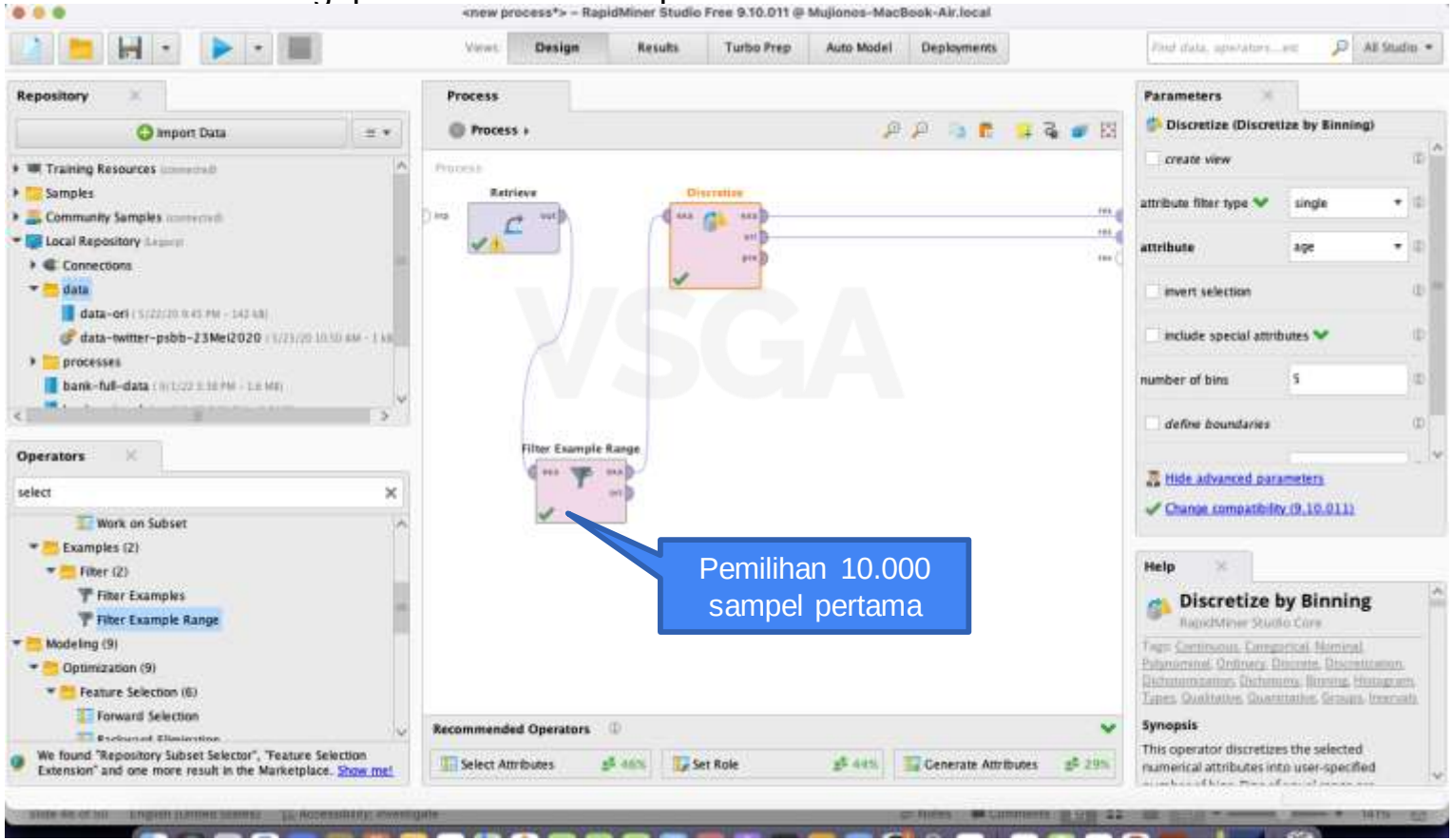
Smoothing: Binning berdasar “banyaknya kategori”

Pada studi kasus ini kita akan melakukan binning atas fitur “age” pada dataset bank-full-data-CSV seperti pada modul sebelumnya. Ikut cara sbb. :

1. Buka proses baru, Load Data / Import Dataset yang sudah disediakan nama file : *bank-full-data.CSV*
2. Gunakan / Pilih proses Discritization Binning (Anda bisa melakukan pencarian di menu Operator). Lengkapi parameter yang diperlukan untuk proses binning :
    1. Attribut file type : ***single***
    2. Attribute: ***age***
    3. Number of bin : ***5 (*Ini adalah banyaknya kategori yang dihasilkan dari proses binning)***
3. Jalankan proses, perhatikan hasilnya seperti berikut

Catatan*: Proses ditengah adalah “select”, untuk memilih instan data (baris) yang akan dilakukan binning. Pada proses ini dari 45.000 lebih data, dipilih 10.000 yang pertama

Konfigurasi Proses ”Binning” pada antar muka rapid miner



Hasil Proses Binning

Data asli

Data hasil binning

Data
Statistics
Visualizations
Annotations

Open in Turbo Prep

Row No.	age	job	marital	education
1	58	management	married	tertiary
2	44	technician	single	secondary
3	33	entrepreneur	married	secondary
4	47	blue-collar	married	unknown
5	33	unknown	single	unknown
6	35	management	married	tertiary
7	28	management	single	tertiary
8	42	entrepreneur	divorced	tertiary
9	58	retired	married	primary
10	43	technician	single	secondary
11	41	admin.	divorced	secondary
12	29	admin.	single	secondary
13	53	technician	married	secondary
14	58	technician	married	unknown

Data
Statistics
Visualizations
Annotations

Open in Turbo Prep
Auto Model

Row No.	age
1	range5 [52.800 – ∞]
2	range3 [36.400 – 44.600]
3	range2 [28.200 – 36.400]
4	range4 [44.600 – 52.800]
5	range2 [28.200 – 36.400]
6	range2 [28.200 – 36.400]
7	range1 [-∞ – 28.200]
8	range3 [36.400 – 44.600]
9	range5 [52.800 – ∞]
10	range3 [36.400 – 44.600]
11	range3 [36.400 – 44.600]
12	range2 [28.200 – 36.400]
13	range5 [52.800 – ∞]

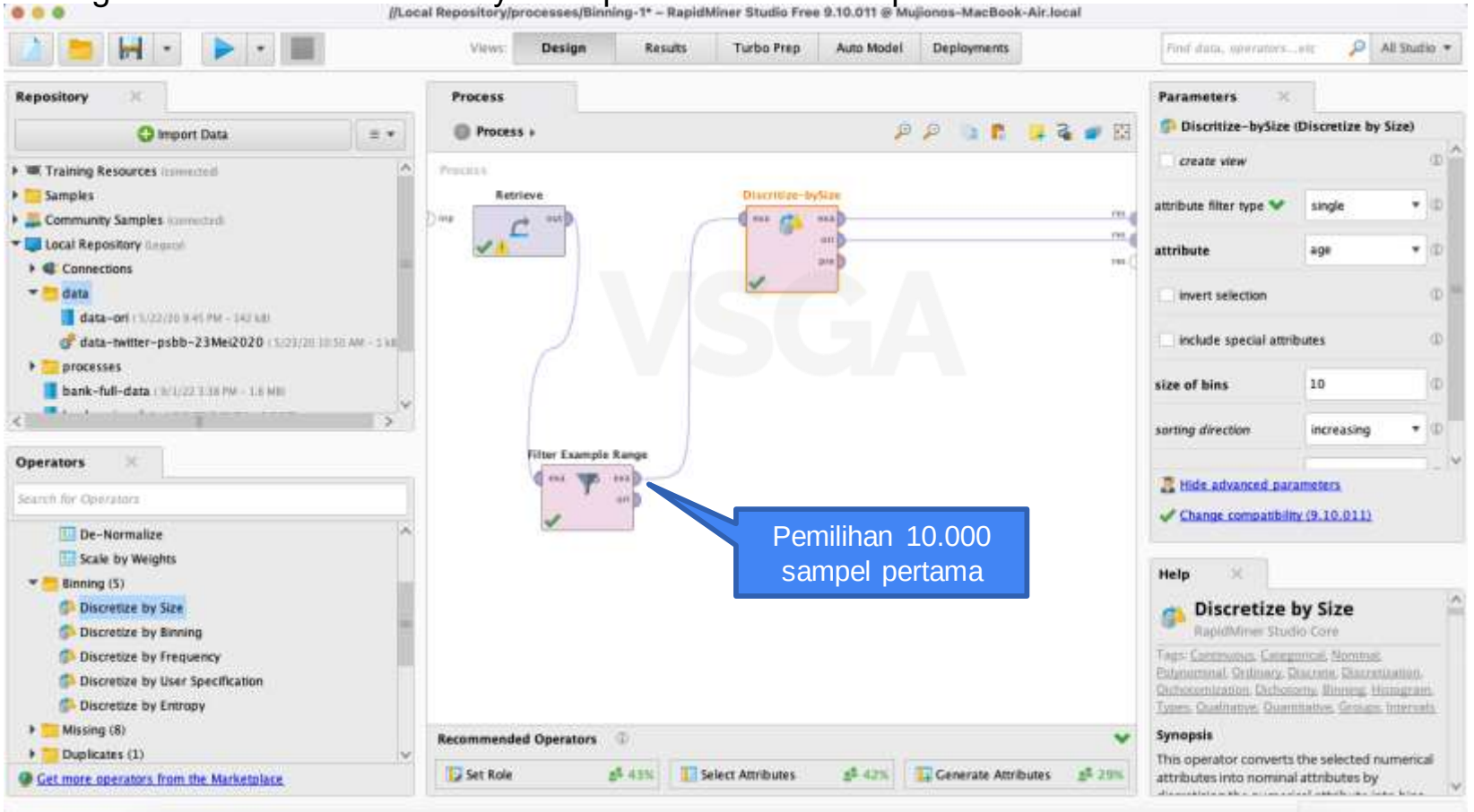
Smoothing: Binning berdasar “ukuran kategori”

Pada studi kasus ini kita akan melakukan binning atas fitur “age” pada dataset bank-full-data-CSV seperti pada modul sebelumnya. Ikut cara sbb. :

1. Buka proses baru, Load Data / Import Dataset yang sudah disediakan nama file : *bank-full-data.CSV*
2. Gunakan / Pilih proses Discritization **Discretize by Size** (Anda bisa melakukan pencarian di menu Operator). Lengkapi parameter yang diperlukan untuk proses binning :
    1. Attribut file type : ***single***
    2. Attribute: ***age***
    3. Size of bin : ***5 (*Ini adalah ukuran kategori yang dihasilkan dari proses binning)***
3. Jalankan proses, perhatikan hasilnya seperti berikut

Catatan*: Proses ditengah adalah “select”, untuk memilih instan data (baris) yang akan dilakukan binning. Pada proses ini dari 45.000 lebih data, dipilih 10.000 yang pertama

Konfigurasi Proses ”Discretize-by Size” pada antar muka rapid miner



Hasil Proses Binning

Data asli

Data hasil binning

Data
Statistics
Visualizations
Annotations

Open in Turbo Prep

Row No.	age	job	marital	education
1	58	management	married	tertiary
2	44	technician	single	secondary
3	33	entrepreneur	married	secondary
4	47	blue-collar	married	unknown
5	33	unknown	single	unknown
6	35	management	married	tertiary
7	28	management	single	tertiary
8	42	entrepreneur	divorced	tertiary
9	58	retired	married	primary
10	43	technician	single	secondary
11	41	admin.	divorced	secondary
12	29	admin.	single	secondary
13	53	technician	married	secondary
14	58	technician	married	unknown

Data
Statistics
Visualizations
Annotations

et (Discritize-bySize)
ory

Turbo Prep
Auto Model

Row No.		job
1	range960 [57.500 - 58.500]	management
2	range670 [43.500 - 44.500]	technician
3	range262 [32.500 - 33.500]	entrepreneur
4	range754 [46.500 - 47.500]	blue-collar
5	range262 [32.500 - 33.500]	unknown
6	range345 [34.500 - 35.500]	management
7	range73 [27.500 - 28.500]	management
8	range611 [41.500 - 42.500]	entrepreneur
9	range960 [57.500 - 58.500]	retired
10	range641 [42.500 - 43.500]	technician
11	range579 [40.500 - 41.500]	admin.
12	range99 [28.500 - 29.500]	admin.
13	range876 [52.500 - 53.500]	technician
14	range960 [57.500 - 58.500]	technician
15	range943 [56.500 - 57.500]	services

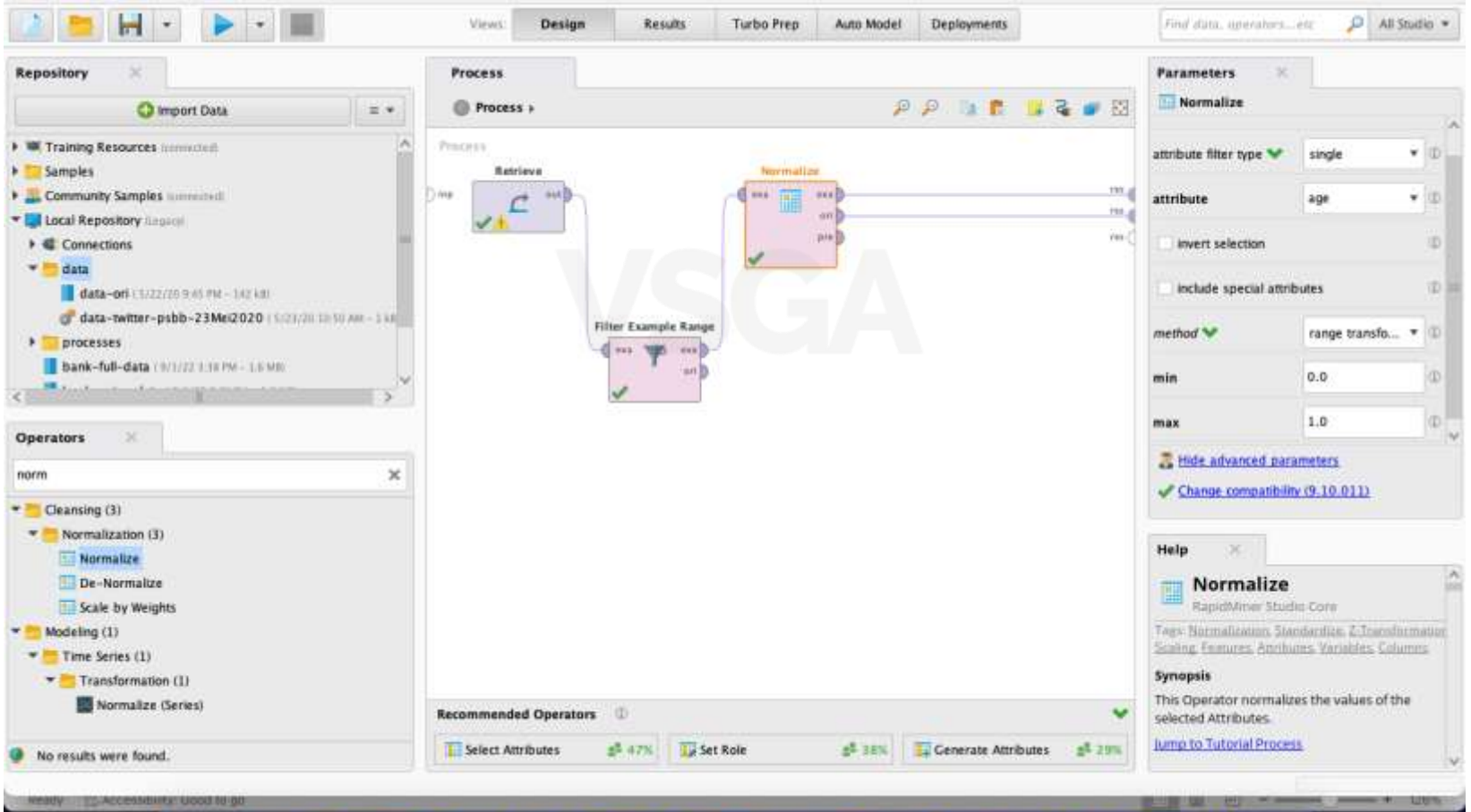
Normalisasi: Scaling

Pada studi kasus ini kita akan melakukan scaling atas fitur “age” pada dataset bank-full-data-CSV seperti pada modul sebelumnya. Ikut cara sbb. :

1. Buka proses baru, Load Data / Import Dataset yang sudah disediakan nama file : *bank-full-data.CSV*
2. Gunakan / Pilih proses Cleansing **Normalization** (Anda bisa melakukan pencarian di menu Operator). Lengkapi parameter yang diperlukan untuk proses Normalisasi :
    1. Attribut file type : ***single***
    2. Attribute: ***age***
    3. Method : ***range transformation***
    4. Min : **nilai minimal attribute hasil (kasus ini diset 0/nol)**
    5. Max: **nilai maksimal attribute hasil (kasus ini diset 1)**
3. Jalankan proses, perhatikan hasilnya seperti berikut

Catatan*: Proses ditengah adalah “select”, untuk memilih instan data (baris) yang akan dilakukan binning. Pada proses ini dari 45.000 lebih data, dipilih 10.000 yang pertama

Konfigurasi Proses ”Scaling-Normalize” pada antar muka rapid miner



Hasil Proses Normalisasi

Data asli

Data hasil normalisasi

Result History

Data

Statistics

Visualizations

Annotations

Open in Turbo Prep | Auto Model

Row No.	age	job	marital	e
1	58	management	married	te
2	44	technician	single	se
3	33	entrepreneur	married	se
4	47	blue-collar	married	u
5	33	unknown	single	u
6	35	management	married	te
7	28	management	single	te
8	42	entrepreneur	divorced	te
9	58	retired	married	p
10	43	technician	single	se
11	41	admin.	divorced	se
12	29	admin.	single	se
13	53	technician	married	se
14	58	technician	married	u
15	57	services	married	se

Data

Statistics

Visualizations

Annotations

Open in Turbo Prep | Auto Model

Row No.	age	job	marital	education	d
1	0.927	management	married	tertiary	no
2	0.585	technician	single	secondary	no
3	0.317	entrepreneur	married	secondary	no
4	0.659	blue-collar	married	unknown	no
5	0.317	unknown	single	unknown	no
6	0.366	management	married	tertiary	no
7	0.195	management	single	tertiary	no
8	0.537	entrepreneur	divorced	tertiary	ye
9	0.927	retired	married	primary	no
10	0.561	technician	single	secondary	no
11	0.512	admin.	divorced	secondary	no
12	0.220	admin.	single	secondary	no
13	0.805	technician	married	secondary	no
14	0.927	technician	married	unknown	no

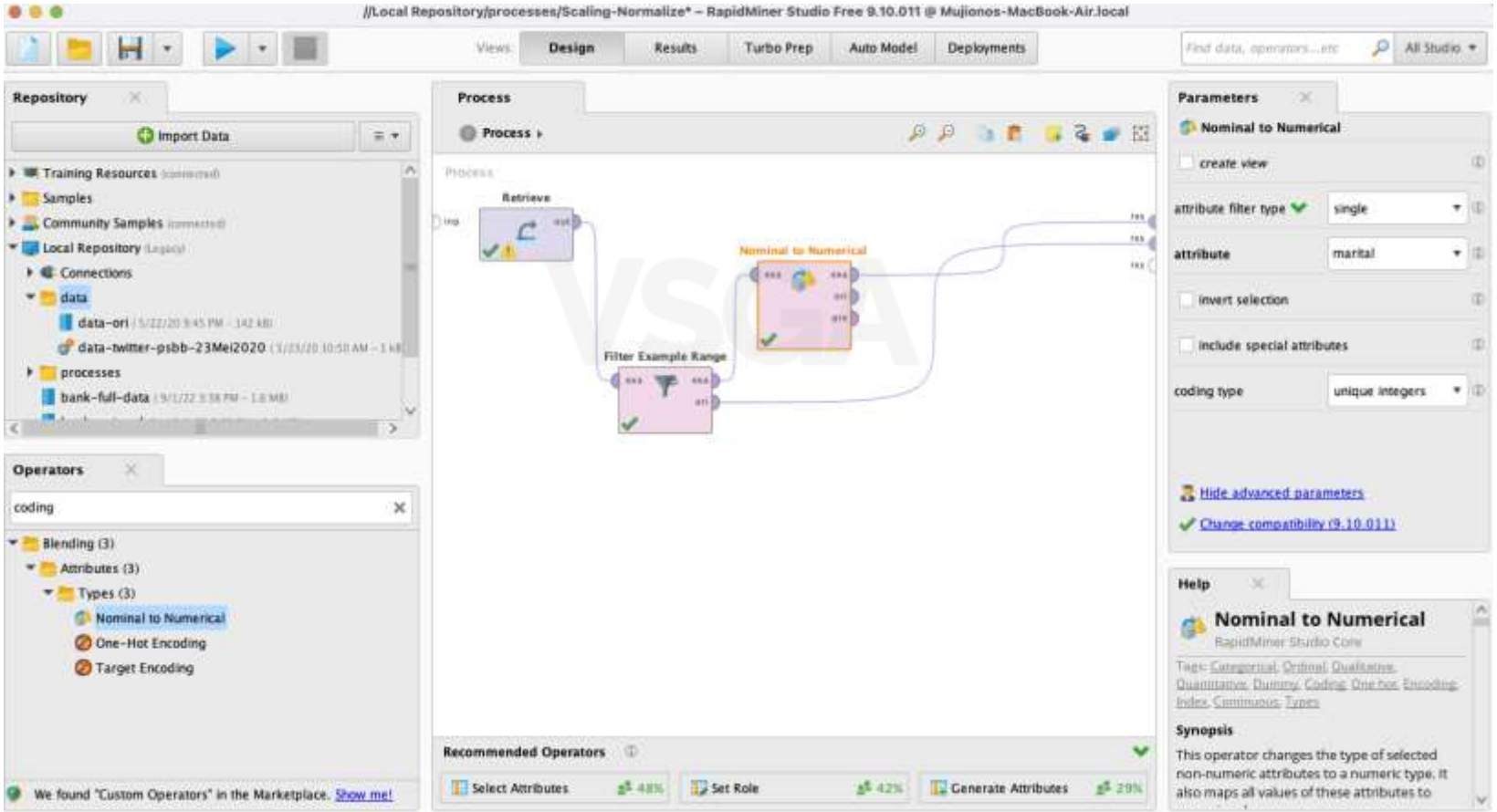
Pengkodean Label atau Pengkodean Ordinal

Pada studi kasus ini kita akan melakukan Pengkoden atas fitur “marital” pada dataset bank-full-data-CSV seperti pada modul sebelumnya. Ikut cara sbb. :

1. Buka proses baru, Load Data / Import Dataset yang sudah disediakan nama file : *bank-full-data.CSV*
2. Gunakan / Pilih proses Type **Nominal-to-numerical** (Anda bisa melakukan pencarian di menu Operator). Lengkapi parameter yang diperlukan untuk proses Normalisasi :
    1. Attribut file type : ***single***
    2. Attribute: ***marital***
    3. Coding type : ***unique integer***
3. Jalankan proses, perhatikan hasilnya seperti berikut

Catatan*: Proses ditengah adalah “select”, untuk memilih instan data (baris) yang akan dilakukan binning. Pada proses ini dari 45.000 lebih data, dipilih 10.000 yang pertama

Konfigurasi Proses ”Nominal-to-numerical” pada antar muka rapid miner



Hasil Proses Pengkodean

Data asli

Data hasil pengkodean

Your license only supports

Open in Turbo Prep Auto Model

Data

Statistics

Visualizations

Annotations

Row No.	age	job	marital	education
1	58	management	married	tertiary
2	44	technician	single	secondary
3	33	entrepreneur	married	secondary
4	47	blue-collar	married	unknown
5	33	unknown	single	unknown
6	35	management	married	tertiary
7	28	management	single	tertiary
8	42	entrepreneur	divorced	tertiary
9	58	retired	married	primary
10	43	technician	single	secondary
11	41	admin.	divorced	secondary
12	29	admin.	single	secondary

ExampleSet (Nominal to Numerical)

Result History

Turbo Prep Auto Model

Data

Statistics

Visualizations

Annotations

Row No.	marital	age	job	edu
1	0	58	management	terti
2	1	44	technician	seco
3	0	33	entrepreneur	seco
4	0	47	blue-collar	unkr
5	1	33	unknown	unkr
6	0	35	management	terti
7	1	28	management	terti
8	2	42	entrepreneur	terti
9	0	58	retired	prim
10	1	43	technician	seco
11	2	41	admin.	seco
12	1	29	admin.	seco
13	0	53	technician	seco
14	0	58	technician	unkr
15	0	57	services	seco
16	0	51	retired	prim
17	1	45	admin.	unkr

Meninggalkan fitur-fitur yang tidak berpengaruh

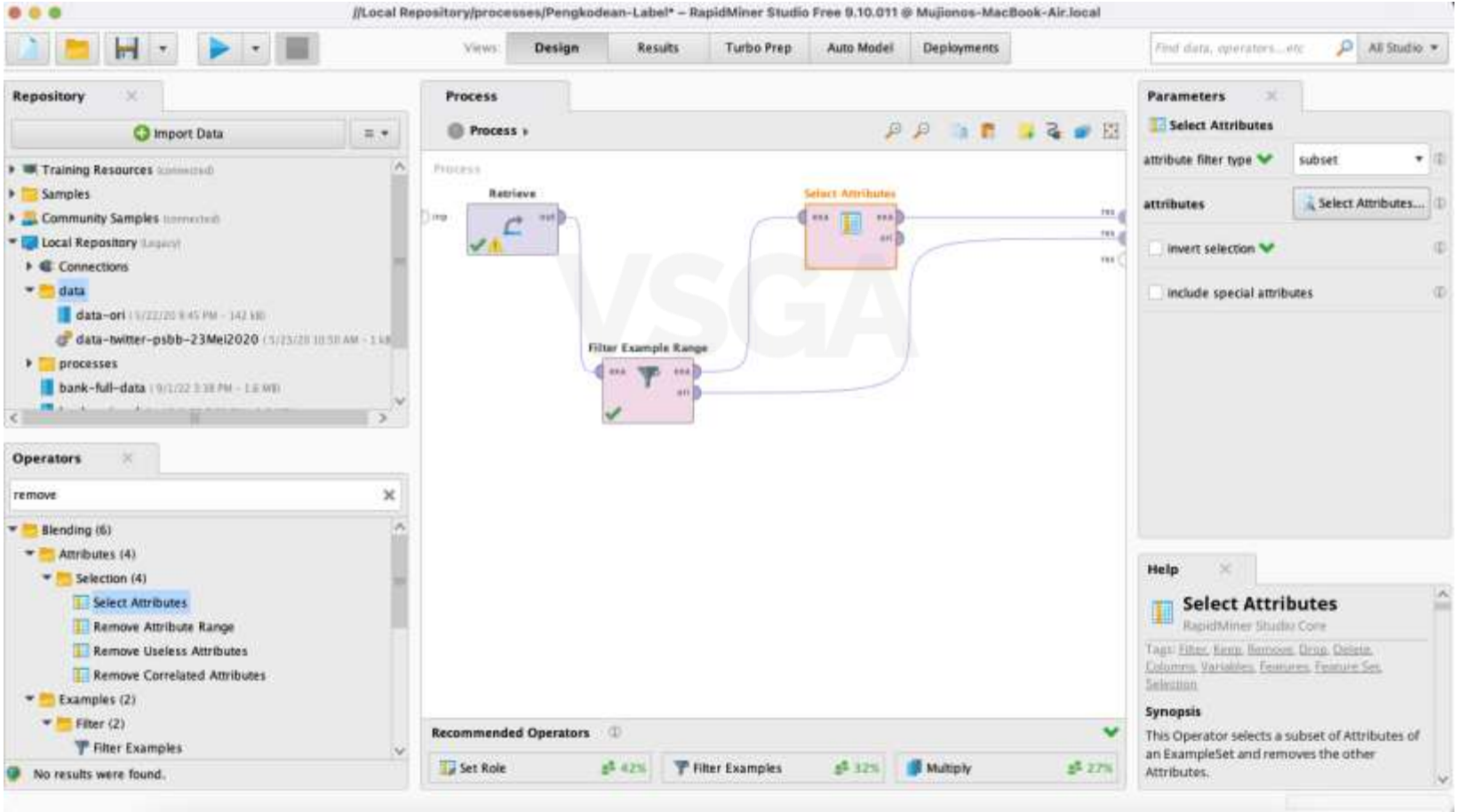
Jika diperhatikan dataset *bank-full-data.CSV,* maka akan terlihat beberapa attribute/fitur tidak berguna untuk proses mining (klasifikasi, regresi, clustering) karena nilainya semua sama untuk semua instan (individu) data. Attribut-2 seperti ini perlu dihilangkan sehingga proses tidak mubazir.

Pada studi kasus ini kita akan melakukan meninggalkan fitur-fitur yang tidak berguna, dalam hal ini “contact” dan “pdays” pada dataset bank-full-data-CSV seperti pada modul sebelumnya. Ikut cara sbb. :

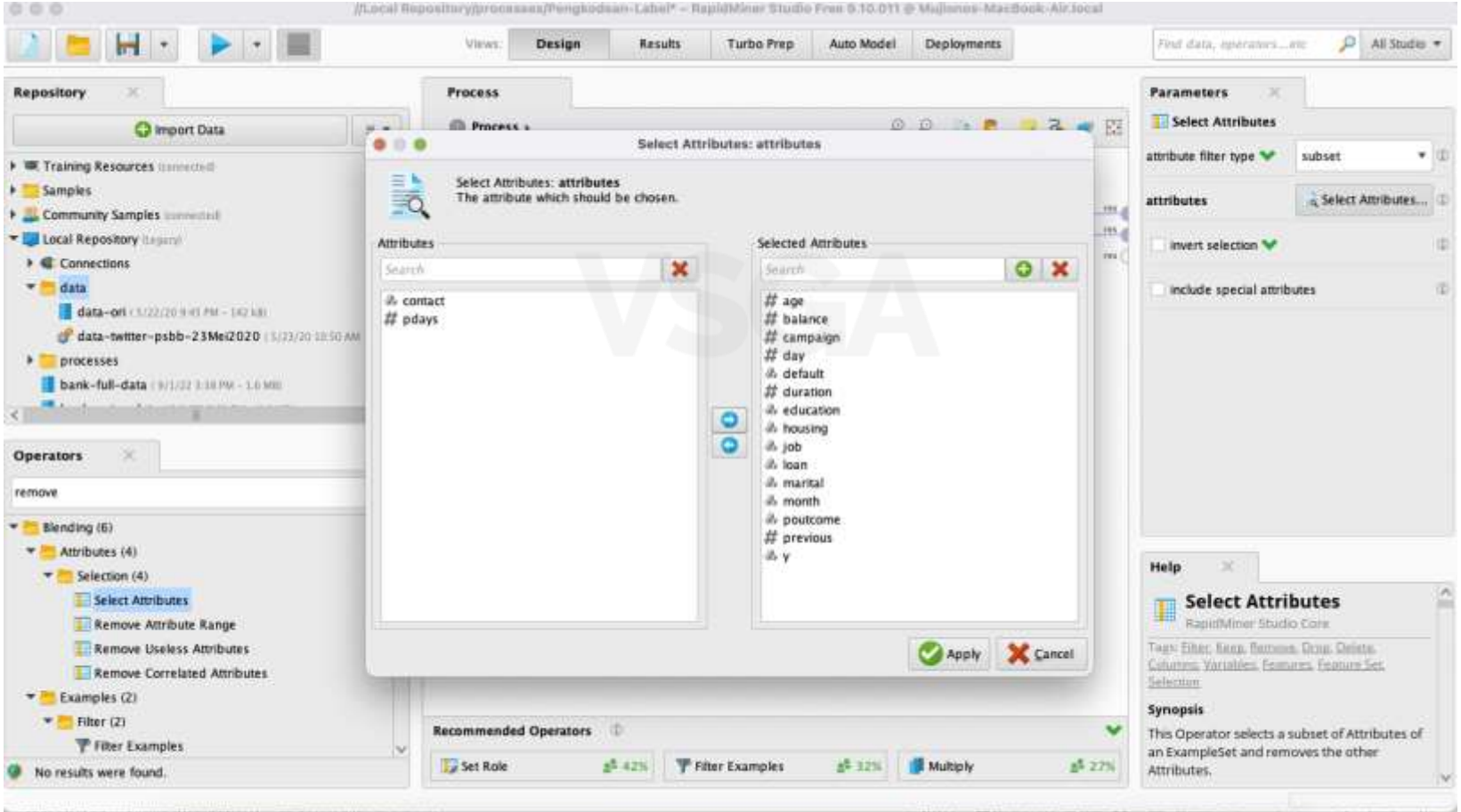
1. Buka proses baru, Load Data / Import Dataset yang sudah disediakan nama file : *bank-full-data.CSV*
2. Gunakan / Pilih proses Attribute Selection **Select Attribute** (Anda bisa melakukan pencarian di menu Operator). Lengkapi parameter yang diperlukan untuk proses Normalisasi :
    1. Attribut file type : ***subset (bisa lebih dari satu attribute)***
    2. Attribut: ***select attribute (akan muncul pop up windows pemilihan attribute yg akan kita gunakan untuk proses selanjutnya)***
    3. Coding type : ***unique integer***
3. Jalankan proses, perhatikan hasilnya seperti berikut

Catatan*: Proses ditengah adalah “select”, untuk memilih instan data (baris) yang akan dilakukan binning. Pada proses ini dari 45.000 lebih data, dipilih 10.000 yang pertama

Konfigurasi Proses ”Select Attribute” pada antar muka rapid miner



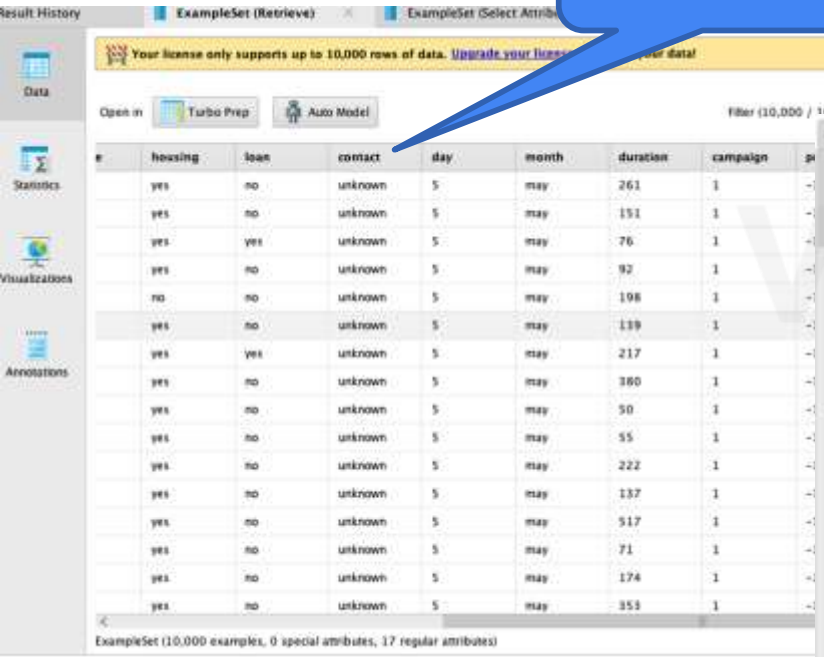
Pop up Windows Proses ”Select Attribute” pada antar muka rapid miner

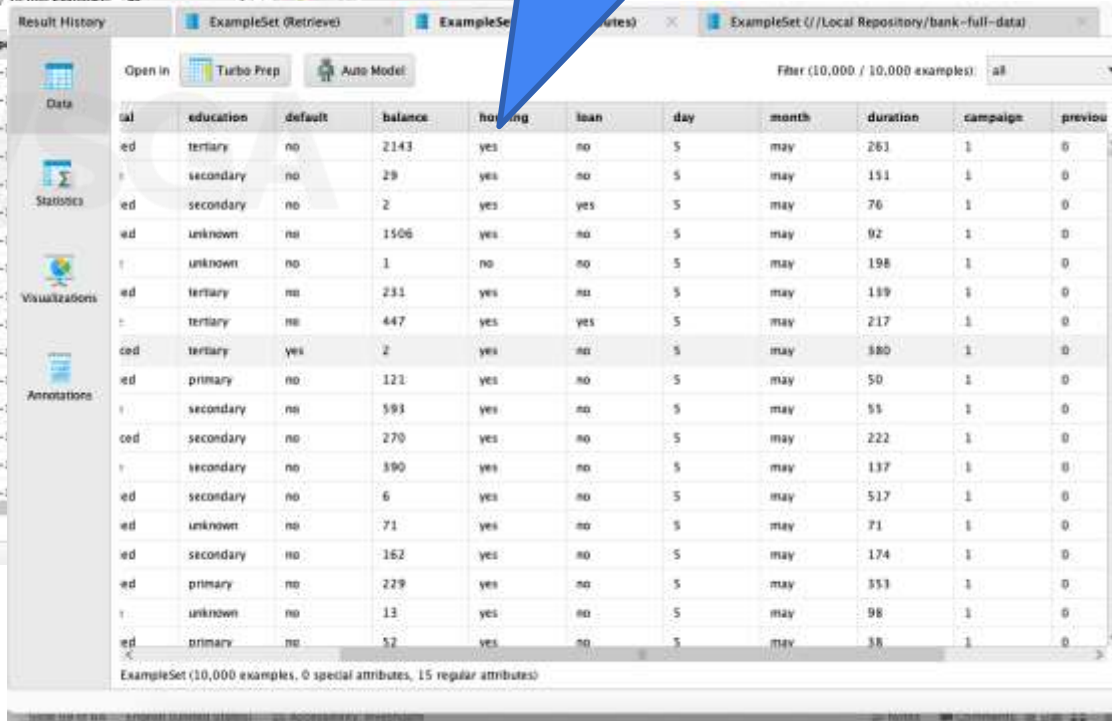


Hasil Proses Penyeleksian Fitur

Data asli

Data hasil pemilihan fitur





Tugas / Latihan

Dengan menggunakan dataset *bank-full-data-CSV* seperti pada modul sebelumnya, lakukan analisis dan pemahaman data / tiap atribut untuk kemudian dilakukan rekonstruksi data: 1. Binning, 2. Penskalaan, 3. Normalisasi, dan 4. Pemilihan attribute. Kerjakan Latihan/tugas berikut:

1. Berikan/uraikan hasil analisis anda dan alasan mengapa masing-masing atribut yang ditentukan cocok / sesuai untuk dilakukan proses rekonstruksi data tersebut.

1. Dari hasil setiap proses rekonstruksi data, lakukan analisis dan pemahaman data dan berikan uraian/penjelasan.

   Petunjuk : Anda bisa melihat/ mengelaborasi pola tiap – tiap attribute hasil secara sendiri – sendiri atau terpisah. Anda bisa melakukan analisis korelasi antar attribute, dll.

Summary

- Transformasi Data adalah bagian dari Data Preparation
- Membutuhkan pengetahuan dasar dan detail serta waktu yang mayoritas untuk menjamin data yang akan dianalisis sebersih mungkin
- Transformasi data dapat menggunakan beberapa teknik rekayasa fitur (feature engineering)
- Normalisasi, Standardisasi adalah bagian proses atau tahapan yang diperlukan untuk mentransformasi data
- Selain data terstruktur, transformasi data juga krusial dilakukan untuk data yang semi terstruktur dan tidak terstruktur (unstructured) seperti teks, image, audio dan video
- Data tidak terstruktur menjadi mayoritas dan tantangan sendiri dalam proses transformasi data di Data Science

Referensi

- https://www.ucl.ac.uk/population-health-sciences/sites/population-health-sciences/files/quartagno_1.pdf
- https://rianneschouten.github.io/missing_data_science/assets/blogpost/blogpost.html
- https://towardsdatascience.com/tf-term-frequency-idf-inverse-document-frequency-from-scratch-in-python-6c2b61b78558
- https://dataaspirant.com/nlp-text-preprocessing-techniques-implementation-python
- https://www.oreilly.com/library/view/blueprints-for-text/9781492074076/assets/btap_0401.png
- https://monkeylearn.com/unstructured-data
- https://medium.com/machine-learning-id/melakukan-feature-scaling-pada-dataset-229531bb08de
- https://protobi.com/post/extreme-values-winsorize-trim-or-retain
- https://heartbeat.fritz.ai/hands-on-with-feature-engineering-techniques-dealing-with-outliers-fcc9f57cb63b
- https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/

KOMDIGI

digitalent

Terima Kasih