

Mata Kuliah Coding & Machine Learning

Laporan Tugas 3.8 Pertemuan 8

Dosen Pengampu: Sri Wulandari, S.Kom., M.Cs.



Disusun oleh:

Lathif Ramadhan (5231811022)

**PROGRAM STUDI SAINS DATA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS TEKNOLOGI YOGYAKARTA
YOGYAKARTA
2025**

Daftar Isi

Daftar Isi	II
Perintah Tugas/Soal:	1
Soal 1:Implementasi Unsupervised Learning Penjelasan Algoritma	2
A. Contoh Kasus: Identifikasi Aspek Produk dan Sentimen Pelanggan dari Ulasan	2
B. Algoritma yang Digunakan: Latent Dirichlet Allocation (LDA)	2
1. Penjelasan Algoritma LDA	2
2. Mengapa LDA Cocok untuk Kasus Ini	3
C. Implementasi di RapidMiner	3
D. Kesimpulan	4
Soal 2: Penjelasan Weight	6
A. Maksud dari Weight (Koefisien Korelasi)	6
B. Makna dari Weight (Koefisien Korelasi)	6
C. Implikasi Praktis dari Weight (Koefisien Korelasi)	7
Soal 3: Implementasi Analisis Korelasi	7
A. Deskripsi Dataset Diamonds Kaggle	8
1. Carat (Karat)	8
2. Cut (Potongan)	8
3. Color (Warna)	8
4. Clarity (Kejernihan)	8
5. Depth (Kedalaman)	8
6. Table (Meja)	8
7. Price (Harga)	8
8. x, y, z (Dimensi)	9
B. Langkah-Langkah Analisis Korelasi Dataset Diamonds di RapidMiner	9
1. Mempersiapkan dan Mengimpor Dataset Diamonds:	9
2. Membuat Proses Baru dan Memuat Data	13
3. Memilih Atribut (Variabel) untuk Analisis Korelasi	14
4. Melakukan Analisis Korelasi	15
5. Menjalankan Proses dan Melihat Hasil	16
6. Menginterpretasikan Hasil Analisis Korelasi	16
C. Penjelasan Hasil Analisis Korelasi Dataset Diamonds	17
1. Matriks Korelasi (Correlation Matrix)	17
2. Bobot Atribut (AttributeWeights) (Terkait Korelasi dengan price)	18
3. Kesimpulan Umum dari Analisis Korelasi	20

Soal 4: Penjelasan Support dan Confidence	21
A. Penjelasan dan Makna Support:	21
B. Penjelasan dan Makna Confidence:	22
C. Hubungan antara Support dan Confidence:	22

Perintah Tugas/Soal:

1. Berikan contoh implementasi unsupervised learning sesuai dengan program studi saudara (sains data) dan sebutkan serta beri penjelasan algoritma yang harus saudara gunakan.
2. Salah satu keluaran dari operator korelasi adalah weight, berikan penjelasan maksud dan maknanya.
3. Lakukan analisis korelasi terhadap data yang saudara temukan sesuai dengan program studi saudara, berikan penjelasan hasil analisis korelasi tersebut.
4. Dalam Association Analysis dikenal istilah support dan confidence, berikan penjelasan dan maknanya.

Soal 1: Implementasi Unsupervised Learning Penjelasan Algoritma

Berikan contoh implementasi unsupervised learning sesuai dengan program studi saudara (sains data) dan sebutkan serta beri penjelasan algoritma yang harus saudara gunakan.

Dalam ranah Sains Data, *unsupervised learning* memegang peranan krusial dalam mengeksplorasi dan memahami struktur tersembunyi di dalam data yang belum memiliki label atau kategori yang ditentukan sebelumnya. Salah satu contoh implementasi yang menarik dan memiliki aplikasi luas adalah dalam bidang **analisis sentimen berbasis aspek (Aspect-Based Sentiment Analysis - ABSA)** pada data teks ulasan pelanggan.

A. Contoh Kasus: Identifikasi Aspek Produk dan Sentimen Pelanggan dari Ulasan

Misalnya terdapat sebuah perusahaan yang bergerak di bidang penjualan perangkat elektronik menerima ribuan ulasan pelanggan setiap harinya. Ulasan-ulasan ini mengandung berbagai informasi mengenai aspek-aspek produk yang berbeda, seperti kualitas layar, performa baterai, kecepatan prosesor, desain, dan harga. Setiap aspek ini dapat memiliki sentimen yang berbeda dari pelanggan (positif, negatif, atau netral).

Tujuan dari implementasi *unsupervised learning* dalam kasus ini adalah untuk secara otomatis mengidentifikasi aspek-aspek produk yang dibicarakan dalam ulasan dan mengelompokkan ulasan berdasarkan sentimen yang diungkapkan terhadap aspek-aspek tersebut, tanpa adanya label sentimen atau aspek yang diberikan sebelumnya pada data ulasan.

B. Algoritma yang Digunakan: Latent Dirichlet Allocation (LDA)

Salah satu algoritma *unsupervised learning* yang sangat sesuai untuk tugas ini adalah **Latent Dirichlet Allocation (LDA)**.

1. Penjelasan Algoritma LDA

LDA adalah sebuah model probabilistik generatif yang digunakan untuk menemukan struktur topik tersembunyi dalam sekumpulan dokumen (dalam kasus ini, setiap ulasan dianggap sebagai sebuah dokumen). Algoritma ini bekerja berdasarkan asumsi bahwa setiap dokumen merupakan campuran dari sejumlah kecil topik, dan setiap kata dalam dokumen dapat diatribusikan ke salah satu topik tersebut. Topik itu sendiri direpresentasikan sebagai distribusi probabilitas atas sekumpulan kata.

Secara konseptual, LDA mencoba untuk memecah setiap ulasan menjadi kombinasi dari beberapa "topik sentimen berbasis aspek". Misalnya, sebuah ulasan tentang sebuah *smartphone* mungkin mengandung topik "kualitas layar - positif", "performa baterai - negatif", dan "harga - netral". LDA akan berusaha untuk mengidentifikasi topik-topik laten ini berdasarkan pola ko-kemunculan kata-kata dalam seluruh kumpulan ulasan.

Proses kerja LDA secara garis besar adalah sebagai berikut:

1. **Inisialisasi Acak:** Algoritma memulai dengan penugasan topik secara acak ke setiap kata dalam setiap ulasan.
2. **Iterasi Inferensi:** Kemudian, secara iteratif, LDA akan memperbaiki penugasan topik ini berdasarkan dua probabilitas:
 - Probabilitas suatu kata muncul dalam suatu topik.
 - Probabilitas suatu topik ada dalam suatu dokumen (ulasan). Proses ini dilakukan dengan menghitung probabilitas bahwa sebuah kata dalam sebuah ulasan seharusnya termasuk ke dalam suatu topik tertentu, dengan mempertimbangkan topik dari kata-kata lain dalam ulasan yang sama dan distribusi kata secara keseluruhan dalam topik tersebut.
3. **Konvergensi:** Iterasi ini terus berlanjut hingga model mencapai konvergensi, yaitu ketika penugasan topik ke kata-kata menjadi stabil.

2. Mengapa LDA Cocok untuk Kasus Ini

a. Kemampuan Mengidentifikasi Topik Laten

LDA secara efektif dapat menemukan tema atau topik tersembunyi (dalam hal ini, aspek-aspek produk dan sentimen terkait) berdasarkan pola penggunaan kata dalam data teks.

b. Probabilistic Output

Output dari LDA berupa distribusi probabilitas topik untuk setiap dokumen (ulasan) dan distribusi probabilitas kata untuk setiap topik. Ini memungkinkan kita untuk memahami komposisi topik dalam setiap ulasan dan interpretasi dari setiap topik berdasarkan kata-kata yang paling mungkin muncul di dalamnya.

c. Fleksibilitas

Jumlah topik yang ingin ditemukan dapat ditentukan sebagai parameter awal dalam algoritma LDA, memungkinkan kita untuk mengontrol granularitas analisis aspek dan sentimen.

C. Implementasi di RapidMiner

Untuk mengimplementasikan LDA dalam RapidMiner, langkah-langkah umum yang dapat dilakukan adalah:

1. Pengambilan dan Pra-pemrosesan Data Teks

Data ulasan pelanggan perlu diimport ke RapidMiner. Langkah-langkah pra-pemrosesan teks seperti *tokenization* (pemecahan teks menjadi kata-kata), *stop word removal* (menghilangkan kata-kata umum yang tidak informatif), dan *stemming/lemmatization* (mengubah kata-kata ke bentuk dasarnya) sangat penting untuk meningkatkan kualitas hasil LDA. Operator-operator seperti `Process Documents from Data`, `Tokenize`, `Remove Stopwords (Dictionary)`, dan `Stem` atau `Lemmatize` dapat digunakan.

2. Aplikasi Algoritma LDA

Operator `LDA` dapat diaplikasikan pada data teks yang telah dipra-proses. Parameter penting yang perlu diatur adalah jumlah topik (`number of topics`) yang ingin ditemukan. Pemilihan jumlah topik yang tepat seringkali memerlukan eksperimen dan evaluasi lebih lanjut.

3. Analisis Hasil

Output dari operator `LDA` meliputi model topik dan penugasan topik untuk setiap dokumen.

○ Model Topik

Menampilkan distribusi kata untuk setiap topik. Dengan melihat kata-kata dengan probabilitas tertinggi dalam suatu topik, kita dapat menginterpretasikan aspek produk dan sentimen yang mungkin terkandung dalam topik tersebut. Misalnya, sebuah topik dengan kata-kata "layar", "jernih", "tajam", "bagus" mungkin mengindikasikan aspek "kualitas layar" dengan sentimen positif. Sebaliknya, topik dengan kata-kata "baterai", "cepat habis", "boros", "buruk" mungkin mengindikasikan aspek "performa baterai" dengan sentimen negatif.

○ Penugasan Topik

Menunjukkan distribusi probabilitas topik untuk setiap ulasan. Ini memungkinkan kita untuk melihat topik-topik mana yang paling dominan dalam setiap ulasan.

D. Kesimpulan

Implementasi *unsupervised learning* menggunakan algoritma Latent Dirichlet Allocation (LDA) pada data ulasan pelanggan memberikan kemampuan untuk secara otomatis mengidentifikasi aspek-aspek produk yang dibicarakan dan sentimen yang terkait dengan aspek-aspek tersebut. Analisis ini sangat berharga bagi perusahaan untuk memahami umpan balik pelanggan secara mendalam, mengidentifikasi area produk yang perlu ditingkatkan, dan merancang strategi pemasaran yang lebih efektif berdasarkan sentimen pelanggan terhadap fitur-fitur spesifik produk. Pendekatan ini memanfaatkan kekuatan *unsupervised learning* untuk menemukan pola dan informasi berharga dari data teks tanpa memerlukan pelabelan manual yang mahal dan memakan waktu.

Soal 2: Penjelasan Weight

Salah satu keluaran dari operator korelasi adalah weight, berikan penjelasan maksud dan maknanya.

Dalam konteks analisis data, khususnya ketika kita menggunakan operator korelasi (seperti operator *Correlation Matrix* dalam perangkat lunak RapidMiner), salah satu hasil yang paling signifikan dan informatif adalah nilai yang sering disebut sebagai *weight*. Secara fundamental, *weight* dalam keluaran operator korelasi merepresentasikan **koefisien korelasi** antara dua variabel yang sedang dianalisis.

A. Maksud dari Weight (Koefisien Korelasi)

Maksud utama dari *weight* atau koefisien korelasi adalah untuk **mengukur dan mengkuantifikasi kekuatan serta arah hubungan linear antara dua variabel numerik**. Dengan kata lain, nilai ini memberikan indikasi seberapa erat perubahan pada satu variabel terkait dengan perubahan pada variabel lainnya, dan apakah hubungan tersebut bersifat positif (kedua variabel bergerak ke arah yang sama) atau negatif (satu variabel meningkat ketika variabel lain menurun).

Operator korelasi akan menghitung koefisien korelasi untuk setiap pasangan variabel dalam dataset yang diberikan. Hasilnya biasanya disajikan dalam bentuk matriks korelasi, di mana setiap sel pada baris *i* dan kolom *j* berisi nilai koefisien korelasi antara variabel ke-*i* dan variabel ke-*j*. Nilai *weight* inilah yang mengisi setiap sel di dalam matriks tersebut.

B. Makna dari Weight (Koefisien Korelasi)

Nilai koefisien korelasi, atau *weight*, memiliki rentang nilai antara **-1 hingga +1**, dan setiap nilai dalam rentang ini memiliki makna yang spesifik dalam menginterpretasikan hubungan antara dua variabel:

- **Nilai +1:** Menunjukkan **korelasi positif sempurna**. Ini berarti bahwa terdapat hubungan linear yang sangat kuat dan searah antara kedua variabel. Ketika satu variabel meningkat, variabel lainnya juga akan meningkat secara proporsional, dan sebaliknya. Jarang sekali ditemukan korelasi sempurna seperti ini pada data riil, namun nilai yang sangat mendekati +1 mengindikasikan hubungan positif yang sangat signifikan.
- **Nilai -1:** Menunjukkan **korelasi negatif sempurna**. Ini mengindikasikan hubungan linear yang sangat kuat namun berlawanan arah. Ketika satu variabel meningkat, variabel lainnya akan menurun secara proporsional, dan sebaliknya. Sama seperti korelasi positif sempurna, korelasi negatif sempurna juga jarang terjadi dalam praktik. Nilai yang sangat mendekati -1 menandakan hubungan negatif yang sangat signifikan.

- **Nilai 0:** Menunjukkan **tidak ada korelasi linear** atau hubungan linear yang sangat lemah antara kedua variabel. Perubahan pada satu variabel tidak secara sistematis terkait dengan perubahan pada variabel lainnya. Namun, penting untuk dicatat bahwa korelasi nol tidak berarti tidak ada hubungan sama sekali antara kedua variabel; mungkin saja terdapat hubungan non-linear yang tidak terdeteksi oleh koefisien korelasi linear.
- **Nilai antara 0 dan +1 (Korelasi Positif Tidak Sempurna):** Nilai positif antara 0 dan 1 mengindikasikan adanya korelasi positif, namun tidak sempurna. Semakin mendekati +1, semakin kuat hubungan positif tersebut. Misalnya, nilai +0.7 menunjukkan korelasi positif yang cukup kuat, di mana kenaikan pada satu variabel cenderung diikuti oleh kenaikan pada variabel lainnya, meskipun tidak selalu dalam proporsi yang tetap.
- **Nilai antara -1 dan 0 (Korelasi Negatif Tidak Sempurna):** Nilai negatif antara -1 dan 0 mengindikasikan adanya korelasi negatif, namun tidak sempurna. Semakin mendekati -1, semakin kuat hubungan negatif tersebut. Sebagai contoh, nilai -0.5 menunjukkan korelasi negatif yang moderat, di mana kenaikan pada satu variabel cenderung diikuti oleh penurunan pada variabel lainnya, meskipun hubungannya tidak sepenuhnya linear atau proporsional.

C. Implikasi Praktis dari Weight (Koefisien Korelasi)

Pemahaman mengenai nilai *weight* atau koefisien korelasi sangat penting dalam analisis data karena memungkinkan kita untuk:

- Mengetahui variabel mana saja yang memiliki hubungan linear yang signifikan satu sama lain.
- Menentukan apakah hubungan antar variabel bersifat positif atau negatif.
- Menilai seberapa kuat keterkaitan linear antara dua variabel. Korelasi yang kuat dapat memberikan wawasan yang lebih mendalam tentang bagaimana variabel-variabel tersebut berinteraksi.
- Hasil korelasi dapat menjadi landasan untuk teknik analisis yang lebih lanjut, seperti pemodelan regresi, di mana kita mencoba untuk memprediksi nilai satu variabel berdasarkan nilai variabel lain yang berkorelasi.
- Dalam pemodelan multivariat, korelasi yang tinggi antar variabel prediktor (multikolinearitas) dapat menjadi masalah dan perlu diatasi.

Dengan demikian, *weight* yang dihasilkan oleh operator korelasi bukan hanya sekadar angka, melainkan sebuah ukuran yang kaya akan makna mengenai hubungan linear antara variabel-variabel dalam dataset. Interpretasi yang tepat dari nilai ini merupakan langkah krusial dalam proses analisis data untuk mendapatkan pemahaman yang lebih baik tentang pola dan keterkaitan yang ada.

Soal 3: Implementasi Analisis Korelasi

Lakukan analisis korelasi terhadap data yang saudara temukan sesuai dengan program studi saudara, berikan penjelasan hasil analisis korelasi tersebut.

A. Deskripsi Dataset Diamonds Kaggle

Dataset "Diamonds" yang saya temukan di Kaggle (Link:

<https://www.kaggle.com/datasets/joebeachcapital/diamonds?select=diamonds.csv>)

merupakan kumpulan data yang sangat populer dan sering digunakan untuk berbagai keperluan analisis data, termasuk analisis korelasi seperti yang akan kita lakukan. Dataset ini berisi informasi detail mengenai karakteristik fisik dan harga dari puluhan ribu berlian.

Bayangkan kita sedang berada di sebuah toko perhiasan besar yang memiliki ribuan berlian. Setiap berlian tentu memiliki ciri khasnya masing-masing yang mempengaruhi keindahan dan nilainya. Nah, dataset ini mencoba menangkap informasi tersebut secara terstruktur.

Secara umum, dataset ini memuat berbagai atribut atau variabel yang menjelaskan setiap berlian. Beberapa atribut utama yang akan kita temukan antara lain:

1. **Carat (Karat):** Ini adalah ukuran berat berlian. Semakin tinggi angka karatnya, umumnya semakin berat dan besar berlian tersebut. Ini adalah salah satu faktor paling signifikan yang mempengaruhi harga berlian.
2. **Cut (Potongan):** Kualitas potongan berlian sangat mempengaruhi bagaimana berlian tersebut memantulkan cahaya atau kilau. Potongan yang baik akan membuat berlian terlihat lebih berkilau dan menarik. Dalam dataset ini, kualitas potongan biasanya dikategorikan secara berurutan mulai dari *Fair* (kurang baik), *Good* (baik), *Very Good* (sangat baik), *Premium* (premium), hingga *Ideal* (ideal/sempurna).
3. **Color (Warna):** Warna berlian diukur berdasarkan seberapa bening atau tidak berwarnanya berlian tersebut. Skala warna umumnya dimulai dari huruf D (paling tidak berwarna dan paling mahal) hingga Z (memiliki rona kuning atau coklat yang jelas). Semakin tidak berwarna sebuah berlian, biasanya semakin tinggi nilainya.
4. **Clarity (Kejernihan):** Kejernihan mengacu pada ada atau tidaknya inklusi (cacat internal) atau noda (cacat eksternal) pada berlian. Skala kejernihan juga berurutan, mulai dari yang memiliki banyak inklusi seperti *I1* (Included 1), *S12* (Slightly Included 2), *S11* (Slightly Included 1), *VS2* (Very Slightly Included 2), *VS1* (Very Slightly Included 1), *VVS2* (Very Very Slightly Included 2), *VVS1* (Very Very Slightly Included 1), hingga yang paling jernih yaitu *IF* (Internally Flawless) dan *FL* (Flawless).
5. **Depth (Kedalaman):** Ini adalah persentase kedalaman total berlian relatif terhadap diameter rata-ratanya. Angka ini mempengaruhi bagaimana cahaya bergerak di dalam berlian dan kembali ke mata pengamat.
6. **Table (Meja):** Ini adalah aspek terbesar dari sebuah berlian, yaitu permukaan datar di bagian atas berlian. Ukuran meja juga mempengaruhi bagaimana cahaya berinteraksi dengan berlian.
7. **Price (Harga):** Ini adalah variabel target yang paling sering dianalisis, yaitu harga berlian dalam mata uang dolar Amerika Serikat.

8. **x, y, z (Dimensi):** Ini adalah ukuran fisik berlian dalam milimeter, yaitu panjang (x), lebar (y), dan kedalaman (z).

Dataset ini sangat berguna untuk memahami bagaimana berbagai karakteristik fisik berlian saling berhubungan dan bagaimana kombinasi dari karakteristik tersebut pada akhirnya mempengaruhi harga. Dengan data ini, kita dapat mengeksplorasi, misalnya, apakah berlian dengan karat lebih tinggi selalu memiliki harga yang jauh lebih mahal, atau bagaimana pengaruh kualitas potongan terhadap harga dibandingkan dengan warna atau kejernihan.

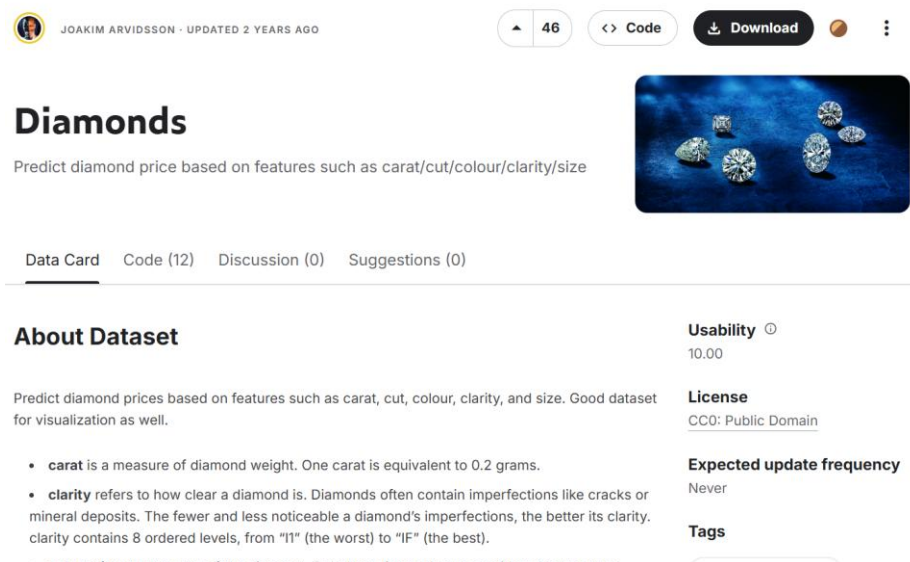
Kumpulan data ini cukup bersih dan terstruktur dengan baik, menjadikannya pilihan yang sangat baik untuk latihan analisis data menggunakan perangkat lunak seperti RapidMiner. Saya akan dapat dengan mudah mengimpor data ini dan mulai melakukan berbagai analisis, termasuk analisis korelasi untuk melihat seberapa kuat hubungan antar variabel yang ada.

B. Langkah-Langkah Analisis Korelasi Dataset Diamonds di RapidMiner

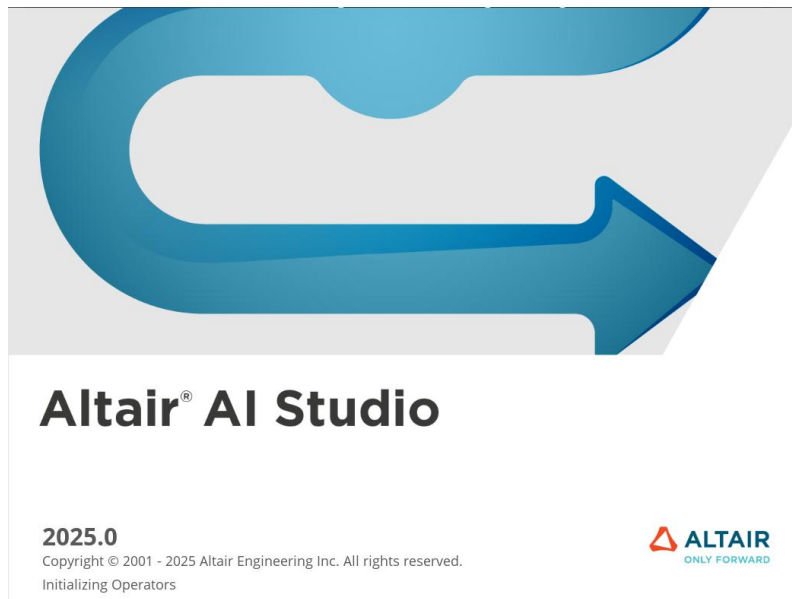
RapidMiner adalah platform yang sangat intuitif untuk melakukan analisis data tanpa perlu menulis kode pemrograman yang rumit. Berikut adalah tahapan umum yang akan kita lakukan untuk menganalisis korelasi pada dataset "Diamonds":

1. Mempersiapkan dan Mengimpor Dataset Diamonds:

- a. **Unduh Dataset:** Pastikan Anda sudah mengunduh file diamonds.csv dari Kaggle dan menyimpannya di lokasi yang mudah Anda akses di komputer Anda. (Link: <https://www.kaggle.com/datasets/joebeachcapital/diamonds?select=diamonds.csv>)

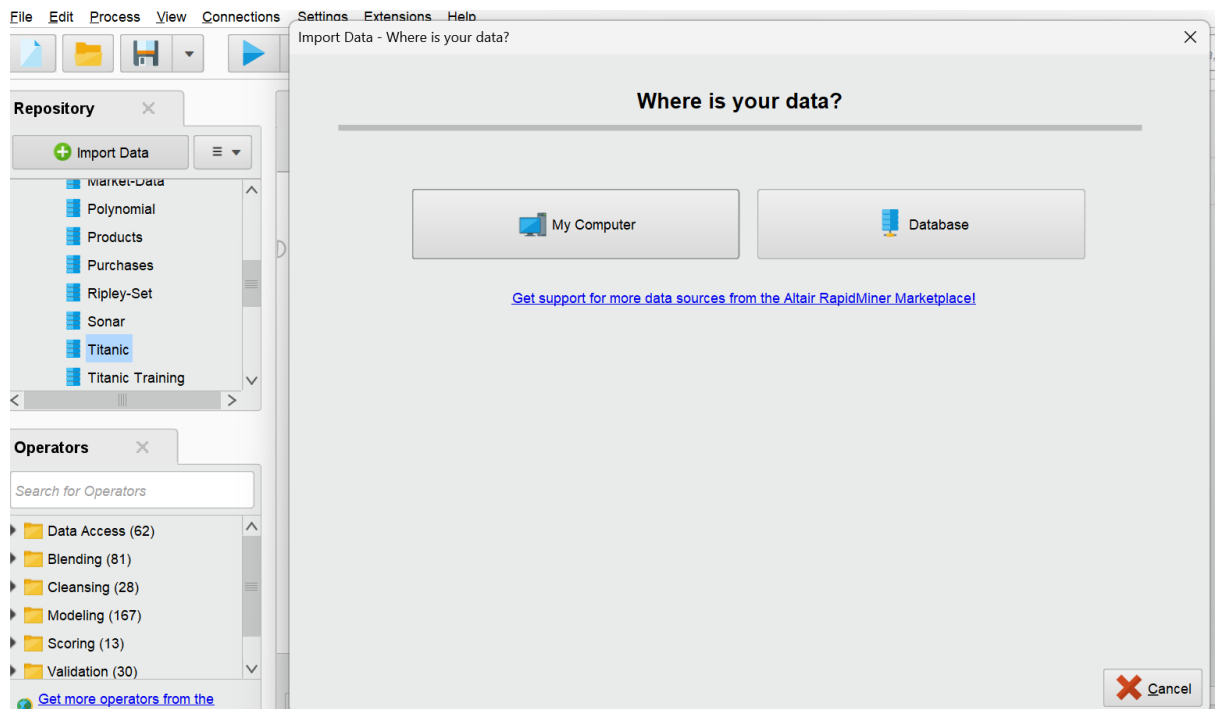


- b. **Buka RapidMiner Studio:** Jalankan aplikasi RapidMiner Studio.

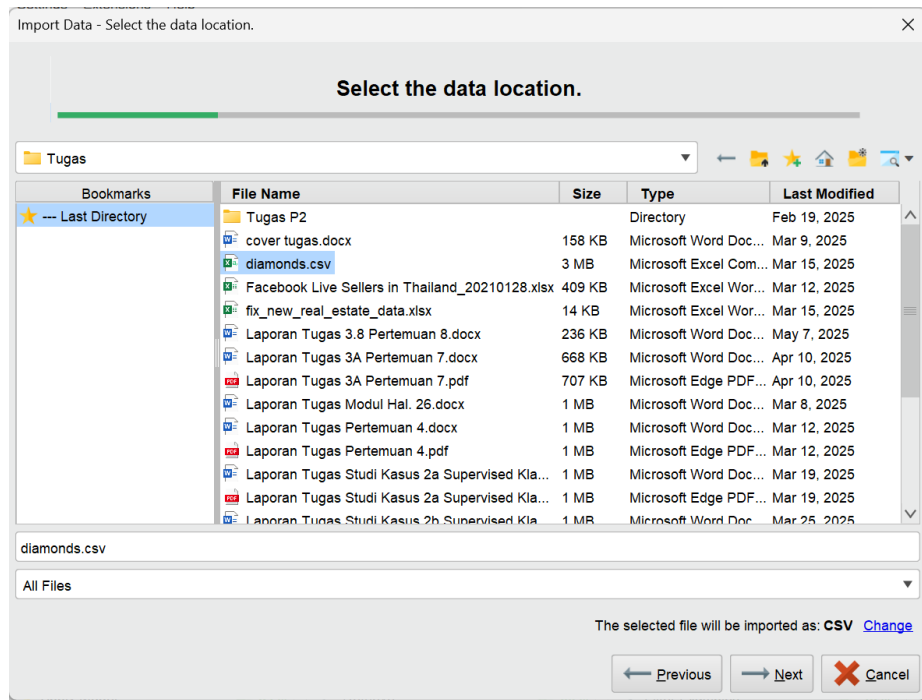


c. Impor Data:

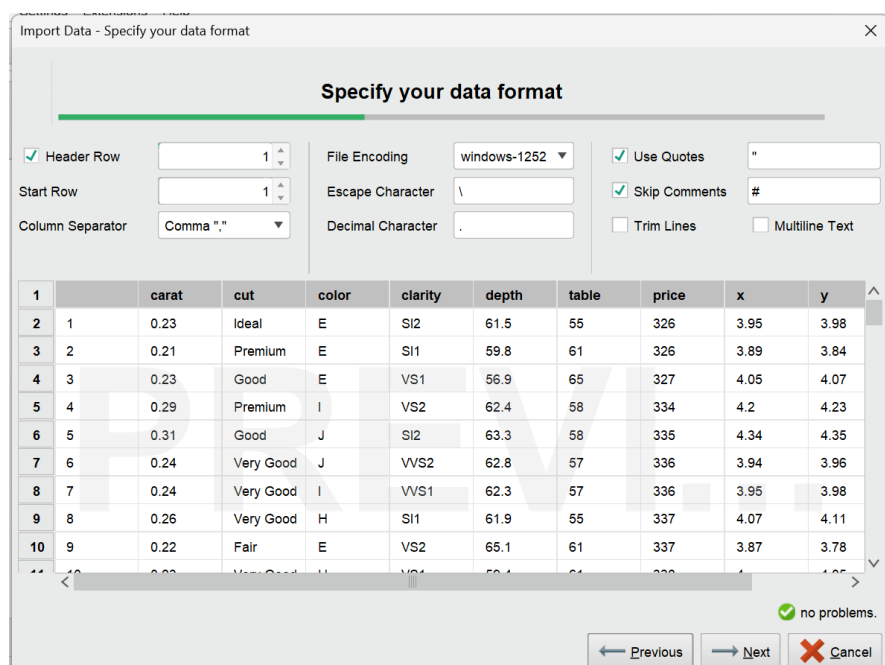
- i. Di panel "Repository" (biasanya di sisi kiri), Anda bisa membuat folder baru untuk proyek ini agar lebih terorganisir. Klik kanan pada "Local Repository" atau folder pilihan Anda, lalu pilih "Create Folder".
- ii. Klik tombol "Import Data" yang biasanya terletak di bagian atas antarmuka, atau klik kanan pada folder yang baru dibuat (atau folder "data" di dalam repository loka)) dan pilih "Import Data".



- iii. Pilih "My Computer" (atau lokasi tempat Anda menyimpan file CSV), lalu navigasikan ke file diamonds.csv dan klik "Next".



- iv. RapidMiner akan menampilkan pratinjau data ini. Periksa apakah data terbaca dengan benar, termasuk pemisah kolom (biasanya koma untuk CSV) dan apakah baris pertama dianggap sebagai nama kolom (header). Biasanya RapidMiner cukup pintar untuk mendeteksinya secara otomatis. Klik "Next".



- v. Pada langkah berikutnya, kita akan melihat format kolom. RapidMiner akan mencoba menebak tipe data untuk setiap kolom (misalnya,

numerik, nominal/kategorikal). Periksa kembali apakah tipe data ini sudah sesuai.

1. Kolom seperti carat, depth, table, price, x, y, dan z seharusnya bertipe numerik (integer atau real).
2. Kolom cut, color, dan clarity adalah data kategorikal atau nominal. Penting untuk memastikan tipe data ini benar karena analisis korelasi biasanya dilakukan pada data numerik. Jika kita ingin melihat hubungan yang melibatkan data kategorikal, kita mungkin perlu melakukan beberapa pra-pemrosesan terlebih dahulu (seperti mengubahnya menjadi numerik dengan teknik tertentu, namun untuk analisis korelasi standar, kita fokus pada numerik).

Import Data - Format your columns.

Format your columns.

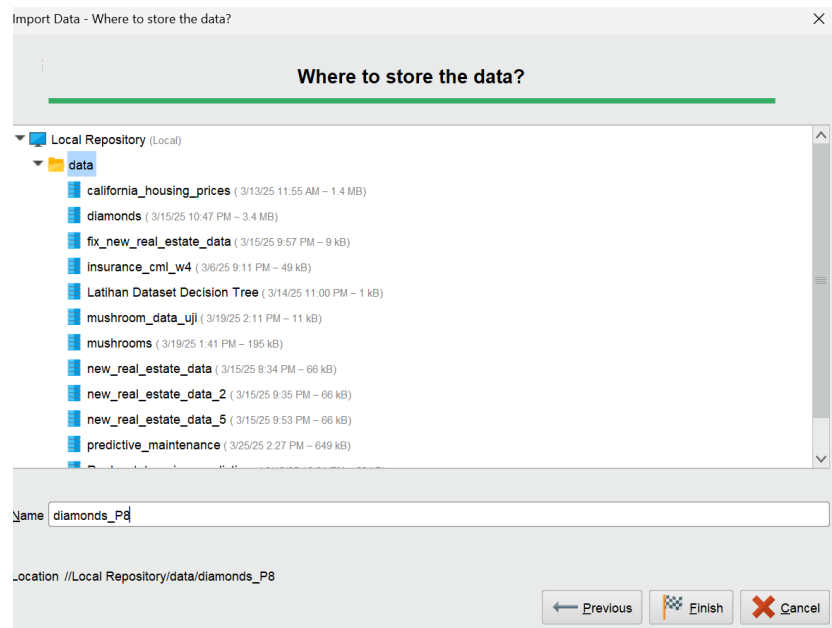
Date format: ☐ Replace errors with missing values ⓘ

	att1 <i>Integer</i>	carat <i>real</i>	cut <i>polynominal</i>	color <i>polynominal</i>	clarity <i>polynominal</i>	depth <i>real</i>
1	1	0.230	Ideal	E	SI2	61.500
2	2	0.210	Premium	E	SI1	59.800
3	3	0.230	Good	E	VS1	56.900
4	4	0.290	Premium	I	VS2	62.400
5	5	0.310	Good	J	SI2	63.300
6	6	0.240	Very Good	J	VVS2	62.800
7	7	0.240	Very Good	I	VVS1	62.300
8	8	0.260	Very Good	H	SI1	61.900
9	9	0.220	Fair	E	VS2	65.100
10	10	0.230	Very Good	H	VS1	59.400
11	11	0.300	Good	J	SI1	64.000

no problems.

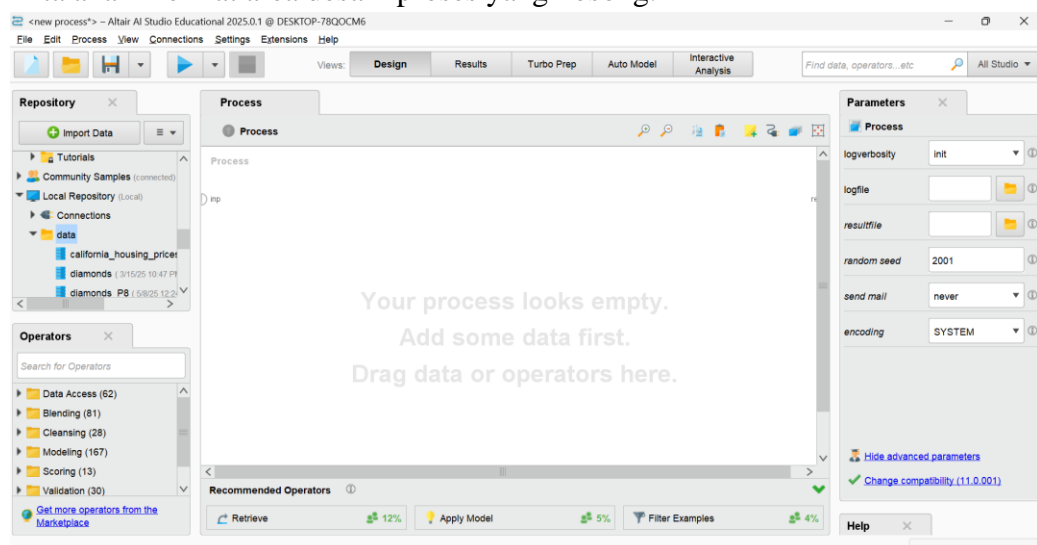
Previous Next Cancel

- vi. Setelah memastikan tipe data sudah benar, klik "Next".
- vii. Pilih lokasi di repository RapidMiner untuk menyimpan data yang sudah diimpor ini. Beri nama yang deskriptif (misalnya, "diamonds_P8"). Klik "Finish". Kini dataset sudah siap digunakan di RapidMiner.

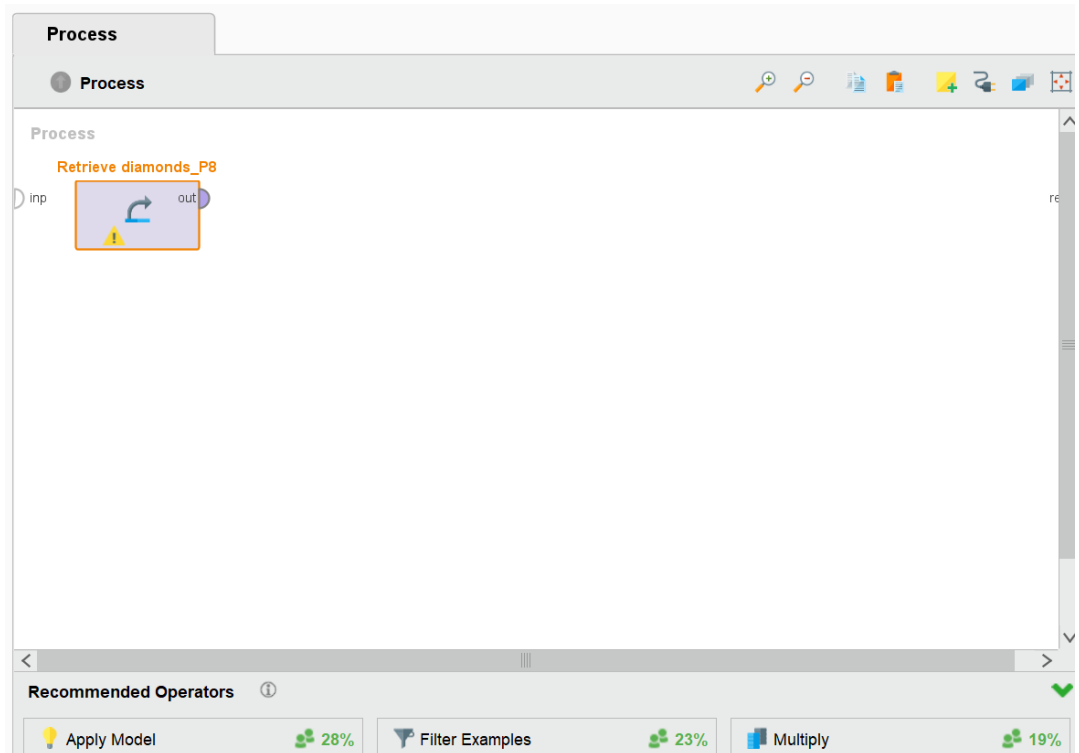


2. Membuat Proses Baru dan Memuat Data:

- Di RapidMiner, analisis dilakukan dalam sebuah "Process". Klik tombol "Blank Process" untuk memulai atau pergi ke "File" > "New Process".
- Kita akan melihat area desain proses yang kosong.

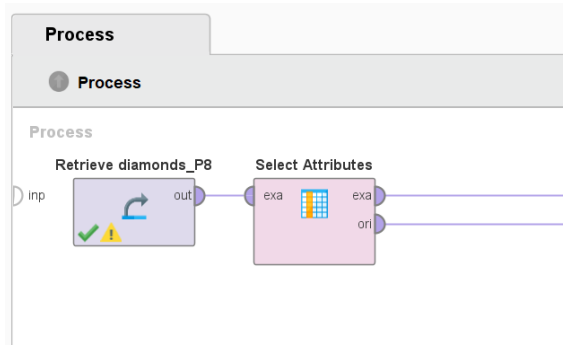


- Dari panel "Repository" tempat menyimpan data tadi (misalnya, "diamonds_P8"), seret (drag) dataset tersebut ke area desain proses. Ini akan memunculkan operator "Retrieve diamonds_P8" (atau nama apa pun yang Anda berikan).

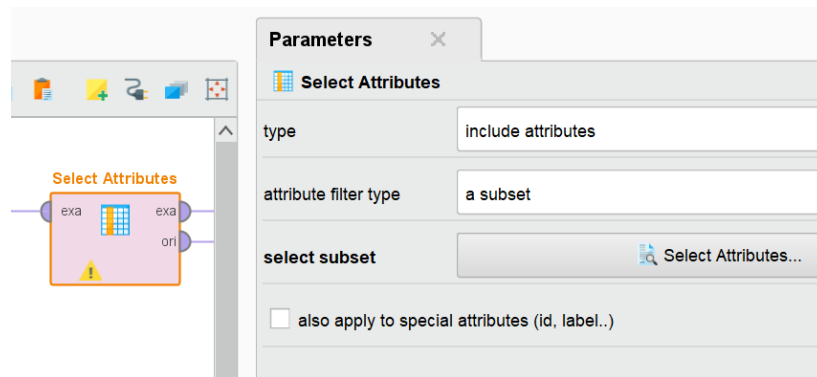


3. Memilih Atribut (Variabel) untuk Analisis Korelasi:

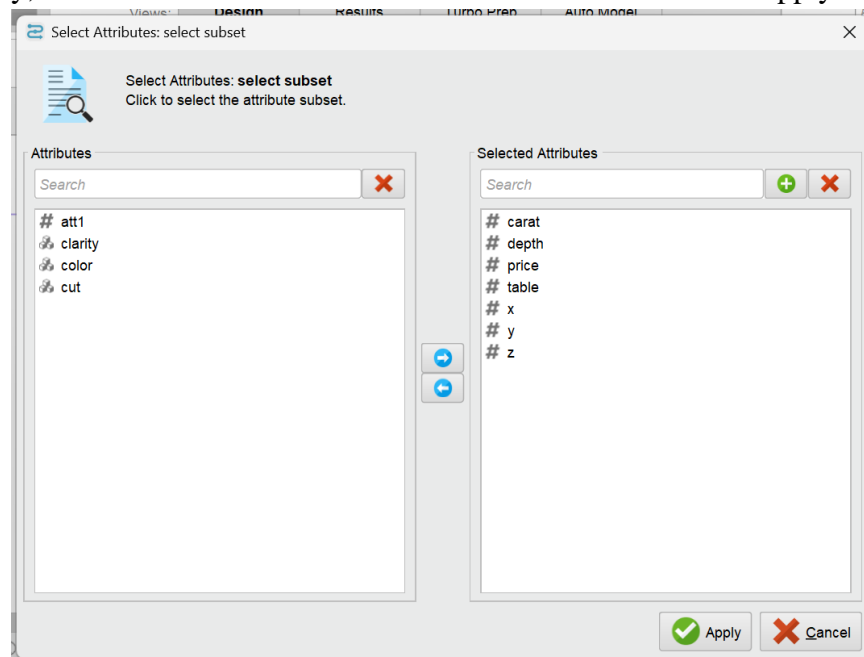
- Analisis korelasi paling umum dilakukan antar variabel numerik untuk melihat seberapa kuat hubungan linear di antara mereka.
- Dataset ini memiliki banyak kolom dan kita hanya ingin menganalisis korelasi antar kolom numerik tertentu, kita bisa menggunakan operator "Select Attributes".
- Cari operator "Select Attributes" di panel "Operators" (biasanya di sisi kiri bawah). Seret operator ini ke area desain proses dan letakkan setelah operator "Retrieve".
- Hubungkan *output port* (port di sisi kanan) dari operator "Retrieve diamonds_P8" ke *input port* (port di sisi kiri) dari operator "Select Attributes".



- Klik pada operator "Select Attributes" di area desain. Di panel "Parameters" (biasanya di sisi kanan atas), kita bisa mengatur atribut mana yang ingin disertakan.
 - Set "attribute filter type" ke "subset".

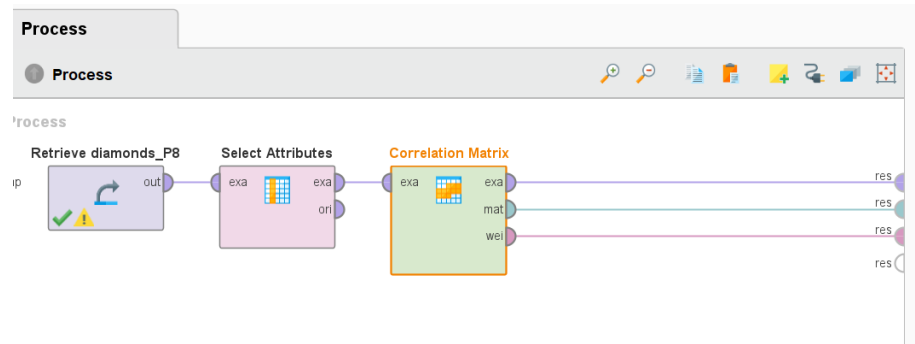


- ii. Klik tombol "Select Attributes". Pilih kolom-kolom numerik yang relevan untuk analisis korelasi kita, seperti carat, depth, table, price, x, y, dan z. Pindahkan kolom-kolom ini ke sisi kanan. Klik "Apply".



4. Melakukan Analisis Korelasi:

- a. Sekarang kita akan menambahkan operator inti untuk analisis korelasi. Cari operator "Correlation Matrix" di panel "Operators".
- b. Seret operator "Correlation Matrix" ke area desain proses dan letakkan setelah operator "Select Attributes" (atau setelah "Retrieve" jika Anda tidak menggunakan "Select Attributes").
- c. Hubungkan *output port* dari "Select Attributes" (atau "Retrieve") ke *input port* "exa" (example set) dari operator "Correlation Matrix".
- d. Operator "Correlation Matrix" akan menghitung koefisien korelasi (biasanya Pearson) antara semua pasangan atribut numerik yang masuk.



5. Menjalankan Proses dan Melihat Hasil:

- Hubungkan *output port* "mat" (matrix) dari operator "Correlation Matrix" ke *port* "res" (result) di sisi kanan area desain proses. Ini akan memberitahu RapidMiner untuk menampilkan matriks korelasi sebagai hasil akhir.
- Kita juga bisa menghubungkan *output port* "exa" (example set) dari "Correlation Matrix" ke "res" jika kita ingin melihat data asli bersamaan dengan hasil korelasi (opsional, tergantung apa yang ingin kita tampilkan).
- Klik tombol "Run" (biasanya ikon play berwarna biru) di bagian atas untuk menjalankan proses.
- RapidMiner akan memproses data dan kemudian menampilkan hasilnya di tab "Results".

6. Menginterpretasikan Hasil Analisis Korelasi:


- Di tab "Results", kita akan melihat sebuah matriks. Baris dan kolom matriks ini adalah nama-nama atribut numerik yang kita analisis.
- Setiap sel dalam matriks menunjukkan koefisien korelasi antara dua atribut. Nilai koefisien korelasi berkisar antara -1 dan +1:
 - +1:** Menunjukkan korelasi positif sempurna. Artinya, jika satu variabel meningkat, variabel lainnya juga meningkat secara proporsional.
 - 1:** Menunjukkan korelasi negatif sempurna. Artinya, jika satu variabel meningkat, variabel lainnya menurun secara proporsional.
 - 0:** Menunjukkan tidak ada korelasi linear antara dua variabel.
 - Nilai antara 0 dan +1 menunjukkan korelasi positif dengan berbagai tingkat kekuatan (semakin dekat ke +1, semakin kuat).
 - Nilai antara 0 dan -1 menunjukkan korelasi negatif dengan berbagai tingkat kekuatan (semakin dekat ke -1, semakin kuat).
- Perhatikan nilai-nilai korelasi yang tinggi (mendekati +1 atau -1) karena ini menunjukkan hubungan yang kuat. Misalnya, kita mungkin akan menemukan korelasi positif yang kuat antara carat dan price.
- RapidMiner juga sering memberikan visualisasi dari matriks korelasi, terkadang dengan warna untuk membantu mengidentifikasi korelasi yang kuat secara visual.


Tentu, mari kita bahas hasil analisis korelasi dari dataset "Diamonds" yang telah kita peroleh menggunakan RapidMiner. Berdasarkan gambar hasil yang telah ada, kita bisa menginterpretasikan matriks korelasi dan bobot atribut (AttributeWeights) yang dihasilkan.


C. Penjelasan Hasil Analisis Korelasi Dataset Diamonds

Hasil analisis korelasi yang kita dapatkan disajikan dalam dua bagian utama: Matriks Korelasi (Correlation Matrix) dan Bobot Atribut (AttributeWeights). Mari kita bedah satu per satu:

1. Matriks Korelasi (Correlation Matrix)

 Correlation Matrix (Correlation Matrix)



 AttributeWeights (Correlation Matrix)

Attribut...	carat	depth	table	price	x	y	z
carat	1	0.001	0.033	0.849	0.951	0.906	0.909
depth	0.001	1	0.087	0.000	0.001	0.001	0.009
table	0.033	0.087	1	0.016	0.038	0.034	0.023
price	0.849	0.000	0.016	1	0.782	0.749	0.742
x	0.951	0.001	0.038	0.782	1	0.950	0.942
y	0.906	0.001	0.034	0.749	0.950	1	0.906
z	0.909	0.009	0.023	0.742	0.942	0.906	1

Matriks ini menunjukkan seberapa kuat hubungan linear antara setiap pasang atribut numerik dalam dataset "Diamonds" ini. Nilai dalam matriks ini adalah koefisien korelasi Pearson, yang berkisar dari -1 hingga +1.

Mari kita lihat beberapa poin penting dari matriks korelasi ini:

- **Korelasi antara carat dan price:** Terdapat nilai korelasi sebesar **0.849**. Angka ini menunjukkan adanya **korelasi positif yang sangat kuat** antara berat berlian (carat) dan harganya (price). Ini adalah temuan yang sangat intuitif; semakin berat sebuah berlian (semakin besar karatnya), maka harganya cenderung semakin tinggi. Kekuatan hubungan ini cukup signifikan.
- **Korelasi antara carat dan dimensi fisik (x, y, z):**
 - carat dan x (panjang): Korelasi **0.951** (sangat kuat positif).
 - carat dan y (lebar): Korelasi **0.906** (sangat kuat positif).
 - carat dan z (kedalaman): Korelasi **0.909** (sangat kuat positif). Ini menunjukkan bahwa seiring bertambahnya berat (carat) berlian, dimensi fisiknya (panjang, lebar, dan kedalaman) juga cenderung meningkat secara signifikan. Hal ini logis karena berlian yang lebih berat umumnya memiliki ukuran fisik yang lebih besar.
- **Korelasi antara dimensi fisik (x, y, z) dan price:**
 - x dan price: Korelasi **0.782** (kuat positif).
 - y dan price: Korelasi **0.749** (kuat positif).

- z dan price: Korelasi **0.742** (kuat positif). Dimensi fisik berlian juga memiliki hubungan positif yang kuat dengan harga. Ini memperkuat ide bahwa ukuran berlian (tidak hanya beratnya) adalah faktor penting dalam penentuan harga.
- **Korelasi antara dimensi fisik (x, y, z):**
 - x dan y: Korelasi **0.950** (sangat kuat positif).
 - x dan z: Korelasi **0.942** (sangat kuat positif).
 - y dan z: Korelasi **0.906** (sangat kuat positif). Ini menunjukkan bahwa ketiga dimensi fisik berlian sangat erat kaitannya satu sama lain. Jika panjangnya bertambah, lebar dan kedalamannya pun cenderung bertambah, yang masuk akal untuk bentuk geometris sebuah berlian.
- **Korelasi yang melibatkan depth (kedalaman persentase) dan table (meja berlian):**
 - depth dengan price: Korelasi **0.000** (hampir tidak ada korelasi linear).
 - table dengan price: Korelasi **0.016** (korelasi positif yang sangat lemah, hampir bisa diabaikan).
 - depth dengan carat: Korelasi **0.001** (hampir tidak ada korelasi linear).
 - table dengan carat: Korelasi **0.033** (korelasi positif yang sangat lemah).

Temuan ini menarik. Berbeda dengan carat dan dimensi fisik, persentase kedalaman (depth) dan ukuran meja (table) tampaknya memiliki hubungan linear yang sangat lemah atau bahkan hampir tidak ada dengan harga dan karat berlian. Ini tidak berarti depth dan table tidak penting, namun hubungan linear langsungnya dengan harga atau karat tidak sekuat variabel lain. Mungkin pengaruhnya lebih kompleks atau non-linear, atau lebih terkait dengan aspek kualitas potongan yang tidak secara langsung terukur hanya dari angka persentase depth dan table saja dalam konteks korelasi linear sederhana.
- **Diagonal Matriks:** Nilai 1 di sepanjang diagonal matriks menunjukkan korelasi sempurna suatu atribut dengan dirinya sendiri, yang memang sudah seharusnya demikian.

2. Bobot Atribut (AttributeWeights) (Terkait Korelasi dengan price)

AttributeWeights (Correlation Matrix)	
attribute	weight
carat	0.004
depth	1
table	0.963
price	0.147
x	0
y	0.033
z	0.037

Tabel "AttributeWeights" ini tampaknya menunjukkan bobot atau pentingnya setiap atribut dalam hubungannya dengan variabel target tertentu (kemungkinan besar price, meskipun tidak secara eksplisit disebutkan di judul tabel ini dalam screenshot, namun ini adalah interpretasi yang umum dalam konteks analisis seperti ini, terutama jika tujuannya adalah memprediksi harga). Bobot ini seringkali berasal dari koefisien korelasi (nilai absolutnya) atau metode seleksi fitur lainnya.

Jika kita mengasumsikan bobot ini merefleksikan kekuatan korelasi dengan price:

- **table:** Memiliki bobot tertinggi (**0.963**). Ini sedikit kontras dengan nilai korelasi table vs price yang rendah (0.016) di matriks korelasi. *Penting untuk diperiksa kembali bagaimana bobot ini dihitung oleh RapidMiner dalam konteks operator yang kita gunakan setelah "Correlation Matrix". Mungkin ada langkah atau operator tambahan yang menghasilkan bobot ini, atau ini adalah output dari operator "Correlation Matrix" itu sendiri yang menghitung bobot berdasarkan kriteria tertentu terhadap satu variabel target (jika dikonfigurasi demikian). Jika ini adalah output langsung dari "Correlation Matrix" tanpa penentuan variabel target secara spesifik untuk pembobotan, maka interpretasi ini perlu dikaji ulang.*
- **price:** Memiliki bobot **0.147**.
- Atribut lain seperti y (0.033), z (0.037), carat (0.004), dan depth (1) serta x (0) memiliki bobot yang bervariasi. Angka 1 untuk depth dan 0 untuk x juga perlu dicermati lebih lanjut mengenai bagaimana bobot ini dihasilkan.

Penting untuk diperhatikan mengenai "AttributeWeights":

Tanpa mengetahui secara pasti bagaimana tabel "AttributeWeights" ini dihasilkan (misalnya, apakah ada pemilihan variabel target tertentu untuk pembobotan), interpretasinya bisa jadi kurang akurat. Namun, jika kita fokus pada **Matriks Korelasi** yang merupakan output standar, interpretasi di atas sudah cukup solid. Untuk bagian "AttributeWeights", kita

mungkin perlu merujuk kembali ke konfigurasi operator di RapidMiner untuk memahami dasar perhitungannya secara lebih presisi.

3. Kesimpulan Umum dari Analisis Korelasi

Berdasarkan analisis matriks korelasi, dapat disimpulkan bahwa:

1. **Karat (carat) adalah faktor yang memiliki hubungan linear positif paling dominan dengan harga (price) berlian.** Semakin besar karat sebuah berlian, kecenderungan harganya untuk lebih tinggi sangatlah kuat.
2. **Dimensi fisik berlian (x, y, z) juga menunjukkan korelasi positif yang kuat dengan harga.** Ini berarti berlian yang lebih besar secara fisik (panjang, lebar, kedalaman) cenderung memiliki harga yang lebih tinggi. Selain itu, karat dan dimensi fisik juga saling berkorelasi sangat kuat, yang memang logis.
3. **Persentase kedalaman (depth) dan ukuran meja (table) berlian menunjukkan korelasi linear yang sangat lemah (hampir tidak ada) dengan harga (price) dan karat (carat) dalam dataset ini.** Ini mengindikasikan bahwa meskipun kedua aspek ini penting untuk kualitas visual berlian, hubungan linear langsungnya dengan harga dan berat tidak sekuat dimensi absolut atau berat itu sendiri. Pengaruh depth dan table terhadap harga mungkin lebih bersifat optimalisasi (ada rentang nilai ideal) daripada hubungan linear sederhana "semakin besar semakin mahal".
4. Atribut-atribut dimensi (x, y, z) saling berkorelasi sangat kuat satu sama lain, menandakan adanya konsistensi bentuk pada berlian.

Analisis korelasi ini memberikan pemahaman awal yang berharga mengenai hubungan antar variabel dalam dataset "Diamonds". Temuan ini bisa menjadi dasar untuk analisis lebih lanjut, misalnya dalam membangun model prediksi harga berlian, di mana variabel dengan korelasi tinggi terhadap harga (carat, x, y, z) kemungkinan akan menjadi prediktor yang penting.

Soal 4: Penjelasan Support dan Confidence

Dalam Association Analysis dikenal istilah support dan confidence, berikan penjelasan dan maknanya.

Dalam ranah penambangan data (data mining), khususnya dalam teknik *Association Analysis*, tujuan utamanya adalah untuk menemukan pola hubungan atau asosiasi yang menarik antara sekumpulan item dalam sebuah dataset transaksi. Dataset transaksi ini umumnya terdiri dari sejumlah transaksi, di mana setiap transaksi berisi sekumpulan item yang dibeli bersamaan (misalnya, dalam keranjang belanja di supermarket) atau terjadi bersamaan (misalnya, halaman web yang diakses oleh pengguna dalam satu sesi). Dua metrik kunci yang digunakan untuk mengevaluasi kekuatan dan signifikansi aturan asosiasi yang ditemukan adalah *support* dan *confidence*.

A. Penjelasan dan Makna Support:

Support, atau dukungan, merupakan ukuran frekuensi atau popularitas dari suatu itemset dalam keseluruhan dataset transaksi. Sebuah itemset adalah sekumpulan satu atau lebih item. Nilai *support* dari suatu itemset X dihitung sebagai proporsi transaksi dalam dataset yang mengandung itemset X . Secara matematis, *support* dapat dirumuskan sebagai berikut:

$$\text{Support}(X) = \frac{\text{Jumlah transaksi yang mengandung itemset } X}{\text{Total jumlah transaksi dalam dataset}}$$

Makna dari Support:

Nilai *support* memberikan indikasi seberapa sering suatu kombinasi item muncul bersamaan dalam dataset. Itemset dengan nilai *support* yang tinggi dianggap lebih signifikan dan berpotensi lebih menarik karena merepresentasikan pola pembelian atau kejadian yang umum terjadi. Dalam konteks bisnis ritel, misalnya, itemset dengan *support* tinggi menunjukkan produk-produk yang sering dibeli bersamaan oleh banyak pelanggan. Informasi ini dapat digunakan untuk berbagai keperluan, seperti penempatan produk di rak toko yang lebih strategis, perancangan promosi bersama (*bundling*), atau analisis segmentasi pelanggan.

Sebaliknya, itemset dengan nilai *support* yang rendah mungkin dianggap kurang menarik karena jarang terjadi, meskipun dalam beberapa kasus, itemset dengan *support* rendah namun *confidence* tinggi dapat mengungkapkan aturan asosiasi yang spesifik dan berharga untuk kelompok pelanggan tertentu.

Dalam proses penambangan aturan asosiasi, seringkali ditetapkan nilai minimum *support* (minimum support threshold) yang harus dipenuhi oleh suatu itemset agar dapat

dipertimbangkan dalam pembentukan aturan asosiasi. Itemset yang memenuhi ambang batas ini disebut sebagai *frequent itemset*.

B. Penjelasan dan Makna Confidence:

Confidence, atau kepercayaan, adalah ukuran Conditional probability dari kemunculan suatu itemset Y (consequent) given bahwa itemset X (antecedent) telah muncul dalam suatu transaksi. Untuk sebuah aturan asosiasi berbentuk "Jika X maka Y" ($X \rightarrow Y$), *confidence* dihitung sebagai proporsi transaksi yang mengandung X dan juga mengandung Y di antara semua transaksi yang mengandung X. Secara matematis, *confidence* dapat dirumuskan sebagai berikut:

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Jumlah transaksi yang mengandung itemset } X \cup Y}{\text{Jumlah transaksi yang mengandung itemset } X} = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$$

Di mana $X \cup Y$ merepresentasikan itemset yang berisi semua item dalam X dan semua item dalam Y.

Makna dari Confidence:

Nilai *confidence* memberikan indikasi seberapa besar kemungkinan itemset Y akan muncul dalam suatu transaksi jika itemset X sudah ada di dalamnya. Dengan kata lain, *confidence* mengukur kekuatan implikasi dari aturan asosiasi. Aturan dengan nilai *confidence* yang tinggi menunjukkan bahwa jika antecedent (X) terjadi, maka consequent (Y) juga kemungkinan besar akan terjadi.

Dalam konteks bisnis, aturan asosiasi dengan *confidence* tinggi dapat digunakan untuk membuat rekomendasi produk (misalnya, "pelanggan yang membeli roti juga cenderung membeli selai"), merancang kampanye pemasaran yang lebih efektif (misalnya, menargetkan pelanggan yang membeli produk A dengan promosi produk B), atau mengoptimalkan tata letak toko berdasarkan pola pembelian pelanggan.

Sama seperti *support*, dalam proses penambangan aturan asosiasi, seringkali ditetapkan nilai minimum *confidence* (minimum confidence threshold) yang harus dipenuhi oleh suatu aturan agar dianggap menarik dan signifikan.

C. Hubungan antara Support dan Confidence:

Support dan *confidence* adalah dua metrik yang saling melengkapi dalam mengevaluasi aturan asosiasi.

- *Support* mengukur seberapa umum atau sering suatu pola item muncul dalam dataset. Aturan dengan *support* rendah mungkin hanya berlaku untuk sebagian kecil transaksi dan mungkin kurang menarik dari perspektif bisnis secara keseluruhan.

- *Confidence* mengukur seberapa dapat diandalkan suatu aturan. Aturan dengan *confidence* tinggi menunjukkan probabilitas yang tinggi bahwa consequent akan terjadi jika antecedent terjadi.

Sebuah aturan asosiasi yang baik idealnya memiliki nilai *support* dan *confidence* yang tinggi. *Support* yang tinggi memastikan bahwa aturan tersebut berlaku untuk sebagian besar transaksi, sementara *confidence* yang tinggi memastikan bahwa aturan tersebut dapat diandalkan dalam memprediksi kemunculan consequent berdasarkan antecedent. Namun, dalam praktiknya, mungkin terdapat aturan dengan *support* rendah namun *confidence* tinggi yang tetap berharga untuk segmen pelanggan tertentu atau untuk produk-produk niche.

Dengan memahami konsep dan makna dari *support* dan *confidence*, para analis data dapat secara efektif mengidentifikasi dan mengevaluasi aturan-aturan asosiasi yang berharga dari dataset transaksi, yang pada gilirannya dapat memberikan wawasan yang berharga untuk pengambilan keputusan bisnis.