umpr
Universitas Muhammadiyah Palangkaraya
Publishing

SMART ANALYTICS

Panduan Visual Regresi dan Klasifikasi dengan Orange Data Mining di Era Data Digital

Dr. Irwani, S.Sos., M.A.P.
Ir. Doddy Teguh Yuwono, S.Kom., M.Kom.
Muhammad Anzarach Pratama, S.A.N., M.P.A

SMART ANALYTICS :
Panduan Visual Regresi dan Klasifikasi dengan Orange Data Mining di Era Data Digital

Dr. Irwani, S.Sos., M.A.P.
Ir. Doddy Teguh Yuwono, S.Kom., M.Kom.
Muhammad Anzarach Pratama, S.A.N., M.P.A

UMPR Publishing
2025

Smart Analytics: Panduan Visual Regresi dan Klasifikasi dengan Orange Data Mining di Era Data Digital

Penulis:
Dr. Irwani, S.Sos., M.A.P.
Ir. Doddy Teguh Yuwono, S.Kom., M.Kom.
Muhammad Anzarach Pratama, S.A.N., M.P.A

Penyunting:
UMPR Publishing

Setting:
UMPR Publishing

Penata Isi:
UMPR Publishing

Desain Sampul:
Hafi Mahyudi

Diterbitkan pertama kali oleh:
UMPR Publishing
Lembaga Penelitian dan Pengabdian kepada Masyarakat
Universitas Muhammadiyah Palangkaraya
Jl. RTA Milono KM 1,5 Palangka Raya, Kalimantan Tengah, Indonesia
Website : https://omp.umpr.ac.id/index.php/umprpublishing/
Email : lp2m@umpr.ac.id
2025

Anggota IKAPI : 004/Anggota Luar Biasa/Kalteng/2024
Anggota APPTI : 004.148.1.11.2021
Anggota APPTIMA : 23/B/ANGGOTA APPTIMA/2023.

ISBN: xxx-xxx-xxxxx-x-x

Kata Pengantar

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

Segala puji bagi Allah Subhanahu wa Ta'ala, Tuhan semesta alam. Shalawat dan salam semoga senantiasa tercurah kepada Nabi Muhammad Shallallahu 'alaihi wa sallam, suri teladan sepanjang zaman, beserta keluarga, sahabat, dan seluruh umat yang setia menapaki jejak perjuangan beliau hingga akhir zaman.

Atas limpahan rahmat, taufik, dan inayah-Nya, penulis dapat menyelesaikan penyusunan buku berjudul ***Smart Analytics: Panduan Visual Regresi dan Klasifikasi dengan Orange Data Mining di Era Data Digital untuk Pemula, Mahasiswa, dan Profesional Data***. Buku ini disusun sebagai bagian dari kontribusi edukatif dalam menjawab tantangan transformasi digital, khususnya dalam bidang analisis data dan pembelajaran mesin (machine learning) secara praktis dan inklusif.

Buku ini menyajikan pendekatan pembelajaran yang visual, terstruktur, dan bebas dari kompleksitas pemrograman. Dengan memanfaatkan platform Orange Data Mining, pembaca diajak untuk memahami konsep regresi dan klasifikasi melalui alur kerja yang mudah diikuti, dilengkapi ilustrasi dan studi kasus nyata. Pendekatan ini memungkinkan siapa saja baik pelajar, dosen, peneliti, maupun professional untuk mengembangkan keterampilan analitik data tanpa harus menguasai bahasa pemrograman terlebih dahulu.

Penulis menyadari bahwa buku ini tentu belum sepenuhnya sempurna. Oleh karena itu, dengan lapang dada, penulis membuka diri terhadap berbagai masukan, kritik, dan saran yang membangun demi penyempurnaan di masa mendatang. Harapan besar penulis, semoga buku ini dapat menjadi salah satu referensi yang bermanfaat dalam memperkuat kompetensi data *science* di kalangan pembelajar dari berbagai latar belakang, sejalan dengan semangat dakwah pencerahan dan kemajuan ilmu pengetahuan yang diusung oleh Persyarikatan Muhammadiyah.

Semoga setiap ilmu yang tersampaikan dalam buku ini menjadi amal jariyah dan memberikan manfaat seluas-luasnya untuk umat dan bangsa.

Palangka Raya, Juli 2025

Tim Penulis

Materi Pokok

Pembelajaran mengenai regresi dan klasifikasi dengan Orange Data Mining dalam konteks perguruan tinggi tidak hanya bertujuan membekali mahasiswa dengan keterampilan teknis, tetapi juga menumbuhkan pola pikir analitis dan tanggung jawab etis dalam menghadapi tantangan transformasi digital. Mahasiswa tidak sekadar diajarkan memahami algoritma statistik, tetapi juga didorong untuk berpikir kritis, mengevaluasi informasi melalui pendekatan visual, dan merancang model prediksi yang aplikatif serta berdampak pada lingkungan sosial sekitarnya.

Dengan pendekatan *project-based learning* dan eksplorasi data nyata dari platform seperti Kaggle dan UCI Repository, mahasiswa diajak mengalami proses pembelajaran yang kontekstual. Mereka tidak hanya "mempelajari metode" tetapi juga "memecahkan masalah" yang mencerminkan persoalan dunia nyata. Model visualisasi interaktif yang disediakan oleh Orange

Data Mining memungkinkan mahasiswa tanpa latar belakang pemrograman sekalipun untuk terlibat secara aktif dalam proses pengolahan dan interpretasi data.

Tujuan pembelajaran ini selaras dengan Capaian Pembelajaran Lulusan (CPL) dan Capaian Pembelajaran Mata Kuliah (CPMK) dengan fokus pada penguasaan pengetahuan data science, keterampilan berpikir sistematis, kemampuan menyampaikan hasil analisis secara efektif, serta keterampilan teknis membangun dan mengevaluasi model regresi dan klasifikasi. Lebih jauh lagi, pembelajaran ini memperkuat pencapaian IKU kampus, seperti IKU 2 (pengalaman praktis mahasiswa), IKU 4 (dosen praktisi mengajar), dan IKU 7 (kelas berbasis proyek dan kolaboratif).

Sebagai bagian dari kontribusi terhadap pembangunan global, pembelajaran ini juga berkorelasi dengan beberapa *Sustainable Development Goals* (SDGs):

- **SDG 4** – mendorong pendidikan berkualitas dan inklusif melalui pendekatan tanpa coding,

- **SDG 8** – mendukung kesiapan mahasiswa dalam menghadapi tantangan pekerjaan digital,
- **SDG 9** – menumbuhkan inovasi berbasis data,
- **SDG 3** – meningkatkan kesadaran terhadap isu kesehatan melalui klasifikasi penyakit menggunakan data terbuka.

Dengan demikian, secara khusus mahasiswa diharapkan mampu:

1. Melakukan analisis eksploratif terhadap data dengan menggunakan visualisasi yang sederhana dan informatif.
2. Mengidentifikasi serta memahami keterkaitan antar variabel dalam dataset.
3. Menginterpretasikan sebaran data terhadap target untuk mendukung pengambilan keputusan berbasis data.
4. Membangun model prediktif regresi dan klasifikasi menggunakan Orange Data Mining.

5. Mengevaluasi performa model dengan metrik evaluasi standar seperti MSE, MAE, R^2, akurasi, dan F1-score.
6. Mengaplikasikan model yang telah dievaluasi untuk membuat prediksi pada data baru secara relevan dan kontekstual.

Tabel 1. Integrasi Pembelajaran dengan CPL, CPMK, IKU, dan SDGs

Komponen	Deskripsi Keterkaitan
CPL – Sikap	Mahasiswa menunjukkan sikap bertanggung jawab dalam menerapkan teknologi prediktif, serta menjunjung etika penggunaan data.
CPL – Pengetahuan	Mahasiswa memahami dasar-dasar regresi, klasifikasi, statistik deskriptif, dan machine learning.
CPL – Keterampilan Umum	Mahasiswa mampu menyajikan hasil analisis data secara visual dan logis, serta menjelaskan model yang dibangun kepada pihak lain.
CPL – Keterampilan Khusus	Mahasiswa mampu membangun dan mengevaluasi model prediktif berbasis data,

Komponen	Deskripsi Keterkaitan
	menggunakan Orange Data Mining secara praktis.
CPMK 1	Mahasiswa memahami prinsip regresi dan klasifikasi serta peranannya dalam analisis data.
CPMK 2	Mahasiswa dapat membangun dan mengevaluasi model regresi dan klasifikasi dengan Orange Data Mining.
CPMK 3	Mahasiswa mampu memvisualisasikan data, menganalisis relasi antar variabel, dan menginterpretasikan hasilnya.
IKU 2	Mahasiswa mendapatkan pengalaman belajar dari data dunia nyata (Kaggle, UCI), mendorong keterampilan praktis.
IKU 4	Kegiatan belajar memungkinkan integrasi praktisi data science untuk mengajar/mendampingi kelas.
IKU 7	Modul dirancang dengan model kelas kolaboratif, eksploratif, dan berbasis proyek data.

Komponen	Deskripsi Keterkaitan
SDG 4 – Pendidikan Berkualitas	Meningkatkan literasi data dan keterampilan teknologi yang inklusif dan aplikatif tanpa coding.
SDG 8 – Pekerjaan Layak dan Pertumbuhan Ekonomi	Mempersiapkan mahasiswa untuk terjun ke dunia kerja berbasis ekonomi digital.
SDG 9 – Inovasi dan Infrastruktur Industri	Menumbuhkan kemampuan membangun solusi data untuk mendukung inovasi sektor industri.
SDG 3 – Kehidupan Sehat dan Sejahtera	Studi kasus (prediksi diabetes, kanker) memberikan kontribusi pada solusi berbasis data di bidang kesehatan.

DAFTAR ISI

DAFTAR GAMBAR

BAB I
PENDAHULUAN

Perkembangan teknologi informasi dan komunikasi saat ini telah mengubah cara manusia berinteraksi dengan data dalam berbagai aspek kehidupan. Di era digital yang semakin kompleks ini, data tidak lagi dipandang sebagai sekadar catatan administratif, melainkan sebagai aset strategis yang dapat dimanfaatkan untuk menciptakan nilai dan inovasi. Berbagai sektor seperti kesehatan, keuangan, pendidikan, pertanian, hingga pemerintahan kini memanfaatkan data sebagai fondasi utama dalam pengambilan keputusan. Hal ini sejalan dengan meningkatnya jumlah data yang tersedia secara masif melalui internet, sensor digital, transaksi elektronik, dan interaksi media sosial yang membentuk fenomena yang dikenal sebagai *Big Data*. Ledakan data ini menuntut kemampuan untuk menggali makna dari data dalam skala besar, yang secara alami melahirkan cabang keilmuan baru yang disebut *Data Science*, atau ilmu data.

Data Science merupakan suatu pendekatan interdisipliner yang menggabungkan ilmu komputer, statistik, matematika, serta domain knowledge tertentu untuk mengekstraksi wawasan atau pengetahuan dari data dalam jumlah besar dan beragam. Dhar (2013) menyebut bahwa tujuan utama dari data science adalah untuk melakukan generalisasi dari data melalui proses ekstraksi pengetahuan yang sistematis dan terstruktur. Pendekatan ini memberikan nilai tambah dalam proses pengambilan keputusan karena tidak hanya memotret kondisi masa lalu (descriptive dan diagnostic analytics), tetapi juga mampu memprediksi kondisi di masa depan (predictive analytics) dan bahkan memberikan saran keputusan terbaik (prescriptive analytics). Dengan demikian, data science menjadi pondasi penting bagi organisasi modern yang ingin bergerak secara adaptif dan berbasis bukti (evidence-based).

Dalam konteks predictive analytics, terdapat dua pendekatan utama dalam data science, yakni regresi dan klasifikasi. Regresi digunakan untuk memprediksi nilai

numerik atau kontinyu, seperti misalnya memproyeksikan biaya kesehatan, jumlah panen, atau nilai properti berdasarkan sejumlah fitur yang diketahui. Sebaliknya, klasifikasi digunakan untuk memprediksi label atau kategori dari suatu entitas, misalnya menentukan apakah seseorang mengidap penyakit tertentu, mengklasifikasikan dokumen berita, atau mengidentifikasi kelayakan pemberian kredit berdasarkan atribut demografis dan finansial. Kedua metode ini sering digunakan dalam berbagai disiplin ilmu dan profesi karena fleksibilitasnya dalam menganalisis pola serta kecenderungan dalam data yang sebelumnya tidak tampak secara eksplisit.

Penerapan regresi dan klasifikasi telah meluas ke berbagai domain kehidupan nyata dan didukung oleh banyak penelitian kontemporer. Misalnya, Khan et al. (2022) memanfaatkan algoritma machine learning untuk memprediksi berat badan bayi yang akan lahir di Uni Emirat Arab dengan akurasi tinggi, yang berguna dalam praktik medis prenatal. Sementara itu, Mateo et al.

(2021) mengembangkan sistem klasifikasi madu berdasarkan jenis bunga asalnya untuk mendukung sektor teknologi pangan dan sertifikasi mutu. Di bidang pendidikan, Wiradinata et al. (2021) menggunakan algoritma *Support Vector Machine* (SVM) untuk memprediksi kelulusan peserta seleksi dalam sebuah *developer academy*, yang memberikan gambaran bagaimana data dapat dimanfaatkan untuk mendukung proses seleksi yang lebih objektif dan efisien. Berbagai kasus tersebut menunjukkan bahwa penguasaan regresi dan klasifikasi bukan hanya bernilai akademik, tetapi juga memiliki dampak nyata dalam kehidupan masyarakat.

Untuk memastikan penerapan data science yang sistematis, dibutuhkan suatu kerangka kerja yang dapat memandu proses analisis data dari awal hingga implementasi. Salah satu standar yang paling umum digunakan di industri dan akademisi adalah *CRISP-DM* (Cross-Industry Standard Process for Data Mining) yang diperkenalkan oleh Wirth dan Hipp (2000). Model ini terdiri dari enam fase utama, yaitu pemahaman bisnis,

pemahaman data, persiapan data, pemodelan, evaluasi, dan implementasi. CRISP-DM memberikan struktur logis yang memungkinkan pengguna data untuk tetap fokus pada tujuan bisnis atau kebijakan, sambil memastikan kualitas teknis dari proses data mining yang dilakukan. Kerangka ini juga bersifat fleksibel dan dapat disesuaikan dengan konteks serta tingkat kompleksitas proyek data yang dikerjakan.

Buku ini dirancang untuk memfasilitasi pemahaman dan penerapan task machine learning, khususnya yang termasuk dalam kategori *supervised learning*. Dalam pendekatan ini, model belajar dari dataset yang telah memiliki label atau target output yang telah ditentukan sebelumnya. Fokus utama dari buku ini adalah pada dua metode supervised learning yang paling umum: regresi linier (*Linear Regression*) dan klasifikasi (*Classification*). Keduanya akan dipelajari secara mendalam melalui pendekatan berbasis visualisasi dan studi kasus, sehingga mahasiswa atau pembaca dapat memahami bagaimana

teori diterapkan secara konkret dalam konteks pemecahan masalah riil.

Namun demikian, realita di lapangan menunjukkan bahwa tidak semua individu, terutama yang berasal dari latar belakang non-teknis atau non-programmer, mampu dengan mudah menerapkan konsep data science. Hambatan utama sering kali muncul pada aspek teknis pemrograman, pengolahan data, dan implementasi algoritma machine learning. Oleh karena itu, buku ini hadir dengan pendekatan yang lebih inklusif: memanfaatkan Orange Data Mining sebagai alat bantu visual yang intuitif dan ramah pengguna. Dengan Orange, proses data science tidak lagi terhalang oleh kemampuan coding, sehingga memungkinkan siapa pun, termasuk pendidik, pelajar, analis bisnis, dan peneliti sosial, untuk ikut serta dalam proses eksplorasi dan analisis data.

Orange Data Mining merupakan perangkat lunak *open-source* yang dikembangkan oleh University of Ljubljana sejak tahun 1996. Aplikasi ini dibangun dengan menggunakan bahasa pemrograman Python dan

didukung oleh modul-modul grafis yang memungkinkan pengguna untuk melakukan *drag-and-drop* dalam menyusun alur kerja analisis data. Orange tidak hanya mendukung pemodelan regresi dan klasifikasi, tetapi juga menyediakan fitur-fitur penting seperti eksplorasi data, visualisasi interaktif, evaluasi model, dan prediksi terhadap data baru. Keunggulan ini menjadikan Orange sebagai salah satu alternatif paling efektif untuk mengenalkan konsep-konsep data mining secara visual dan intuitif kepada pemula maupun praktisi.

Melalui buku ini, pembaca akan diajak untuk menelusuri langkah demi langkah dalam membangun siklus data science, dimulai dari proses akuisisi data, eksplorasi dan visualisasi, pembangunan model prediktif, evaluasi akurasi model, hingga penerapan hasil prediksi dalam skenario dunia nyata. Penyajian materi dikemas dalam bentuk praktis, studi kasus yang relevan, dan ilustrasi yang membantu pemahaman. Dengan demikian, buku ini bukan hanya berfungsi sebagai panduan teknis, tetapi juga sebagai sarana refleksi kritis tentang bagaimana

data dapat digunakan untuk menghasilkan keputusan yang lebih cerdas, adil, dan berdampak positif bagi masyarakat luas.

BAB II
REGRESI BIAYA ASURANSI KESEHATAN

Pada penerapan ilmu data, regresi menjadi salah satu pendekatan utama yang digunakan untuk memperkirakan nilai kontinu berdasarkan variabel-variabel input yang telah diketahui sebelumnya. Salah satu contoh aplikatifnya adalah estimasi biaya asuransi kesehatan seseorang berdasarkan informasi medis dan demografis yang dimilikinya. Topik ini relevan, tidak hanya dari sisi teknis, tetapi juga secara sosial, karena memungkinkan sektor asuransi memberikan pelayanan yang lebih terukur, transparan, dan adil. Pada bab ini, pembaca akan diajak untuk memahami bagaimana membangun model regresi, mulai dari tahap pengumpulan data hingga tahap prediksi akhir dengan bantuan Orange Data Mining.

A. Akuisisi Data

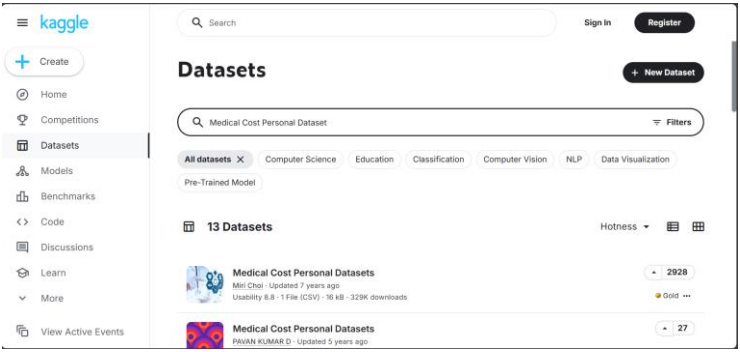
Langkah pertama yang perlu dilakukan dalam membangun sebuah model prediktif adalah memperoleh data yang relevan. Proses ini dikenal dengan istilah *data acquisition* atau akuisisi data. Dalam konteks ini, kita akan menggunakan data terbuka (open-source) yang dapat diakses publik untuk tujuan pembelajaran. Dataset yang digunakan berjudul *Medical Cost Personal Dataset* yang tersedia secara gratis di platform *Kaggle*, sebuah komunitas global bagi praktisi dan penggiat data science.

1. Pengambilan Data dari Kaggle

Untuk mengakses dataset, pengguna terlebih dahulu perlu membuka situs resmi Kaggle di alamat https://kaggle.com. Setelah masuk ke laman tersebut, pengguna dianjurkan untuk membuat akun atau melakukan login jika sudah memiliki akun aktif.

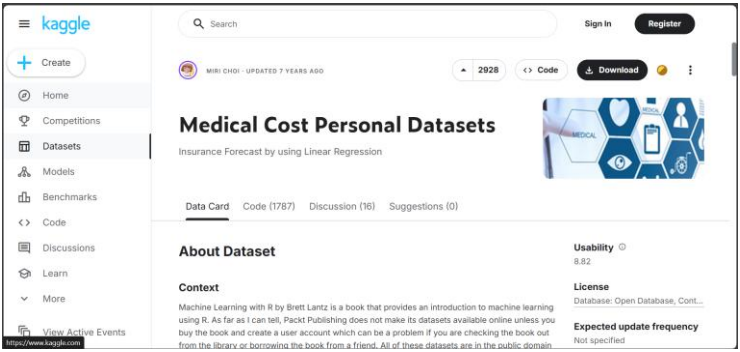
Gambar 1. Laman Registrasi Dataset di Kaggle

Langkah selanjutnya adalah melakukan pencarian dataset dengan mengetikkan kata kunci "Medical Cost Personal Dataset" pada kolom pencarian. Pilih dataset yang diterbitkan oleh pengguna bernama "Miri Choi", karena dataset ini telah banyak digunakan untuk analisis regresi dalam konteks biaya kesehatan.



Gambar 2. Laman Pencarian Dataset di Kaggle

Setelah dataset ditemukan, pengguna dapat mengunduhnya dengan menekan tombol **Download**. File hasil unduhan berbentuk .zip dan perlu diekstraksi terlebih dahulu untuk memperoleh file utama dalam format .csv (Comma-Separated Values) yang akan kita gunakan dalam proses analisis.

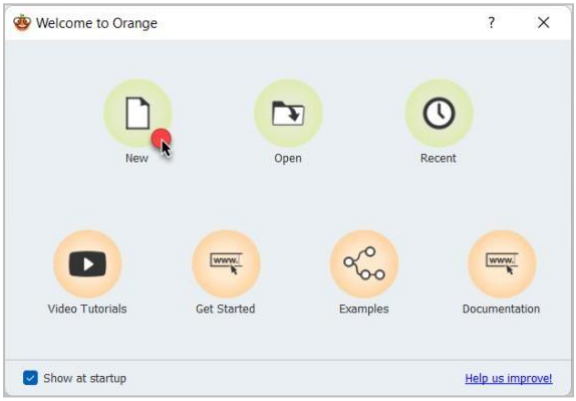


Gambar 3. Laman Dataset Medical Cost di Kaggle

2. Set-up Data di Orange

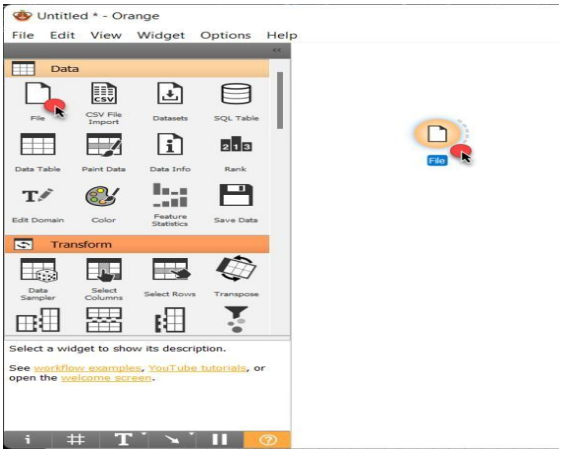
Tahap berikutnya adalah menyiapkan data tersebut di dalam aplikasi Orange Data Mining. Proses ini dikenal sebagai *data setup*, yaitu mengatur dataset agar dapat dikenali dan diproses oleh perangkat lunak Orange.

Langkah pertama, buka aplikasi Orange dan pilih tombol **New** untuk memulai proyek baru.



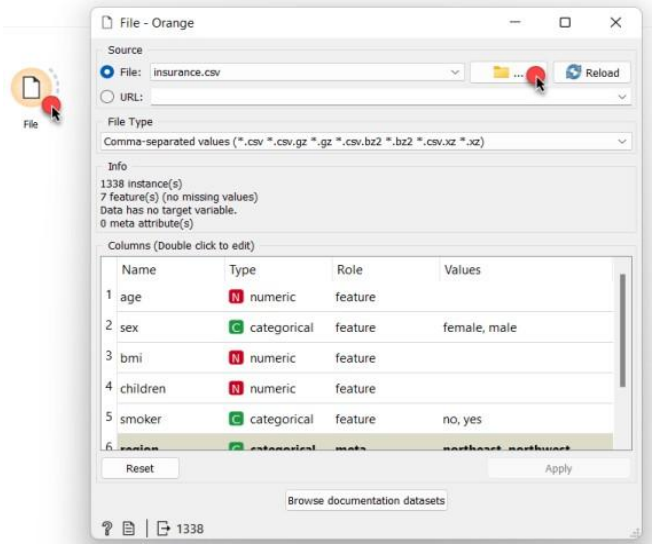
Gambar 4. Tampilan Depan dari Aplikasi Orange

Setelah itu, seret ikon **File** dari panel *Data* ke area kerja (*worksheet*). Ikon ini akan digunakan untuk memuat file .csv yang telah kita unduh sebelumnya.



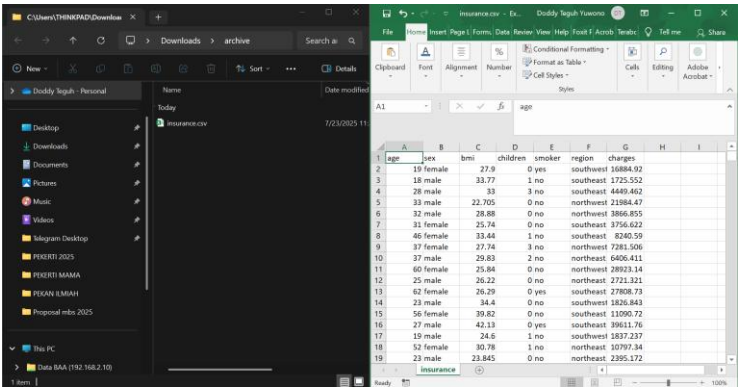
Gambar 5. Memasukkan Ikon File pada Worksheet

Klik dua kali ikon File untuk membuka jendela pemilihan file. Di sini kita akan menentukan lokasi dan file dataset yang ingin digunakan.



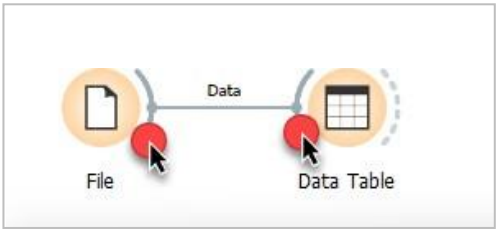
Gambar 6. Tampilan Set-up File di Orange

Setelah memilih file, aplikasi akan menampilkan pratinjau isi file tersebut. Langkah selanjutnya adalah memilih dataset yang akan diolah. Cukup klik file yang diinginkan dari komputer.

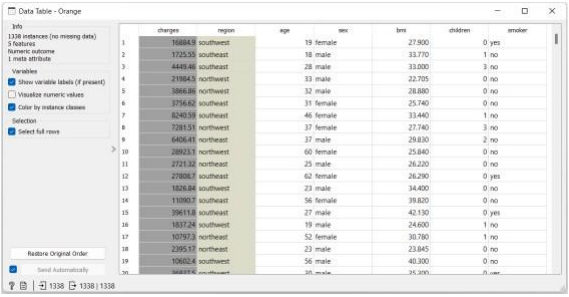


Gambar 7. Tampilan Pencarian File dalam Komputer

Untuk melihat dan memastikan isi dataset, tambahkan ikon **Data Table** dari panel *Data* ke worksheet dan hubungkan dengan ikon File. Dengan mengklik dua kali ikon Data Table, pengguna dapat meninjau semua entri dan atribut dalam dataset.



Gambar 8. Pembuatan Link pada File dan Data Table



Gambar 9. List Data Medical Cost Insurance

Dari tabel yang muncul, terlihat bahwa dataset memiliki 1.338 baris data dan enam kolom, terdiri atas lima atribut (Age, Sex, BMI, Children, Smoker) dan satu atribut target bernama Charges. Seluruh data tersedia secara lengkap tanpa nilai kosong, sehingga proses pembersihan data (data cleaning) tidak diperlukan untuk kasus ini.

B. Pemahaman Data

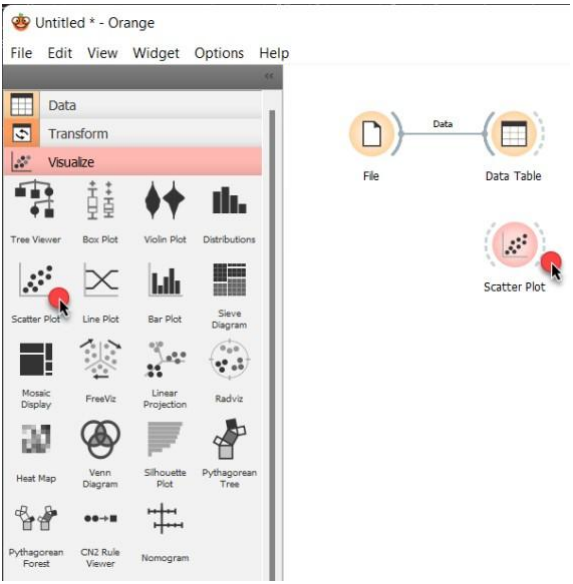
Memahami struktur dan isi data merupakan langkah penting sebelum masuk ke proses pemodelan. Pemahaman ini akan membantu kita mengenali pola, distribusi, serta hubungan antarvariabel dalam dataset, yang menjadi dasar dalam membangun model regresi yang akurat dan valid.

1. Visualisasi Data dengan Scatter Chart

Salah satu metode paling efektif dalam memahami hubungan antarvariabel adalah melalui visualisasi scatter

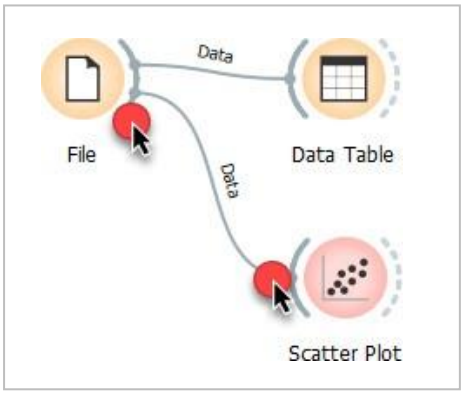
chart atau diagram pencar. Scatter chart dapat menunjukkan korelasi antara dua atribut secara visual dan intuitif.

Untuk memulai, tambahkan ikon **Scatter Plot** dari menu *Visualize* ke worksheet Orange.



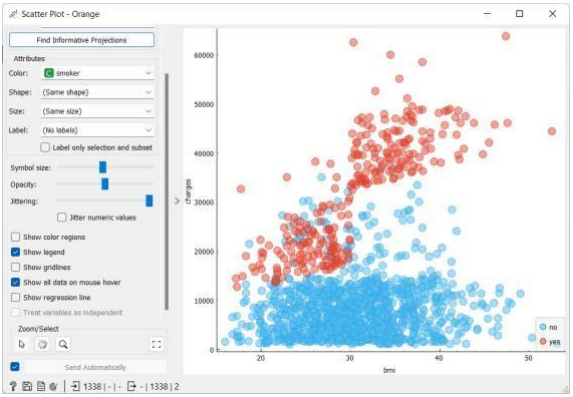
Gambar 10. Pemilihan Scatter Chart untuk Visualisasi Data

Hubungkan ikon dataset (File) dengan ikon Scatter Plot untuk mengalirkan data ke dalam tampilan visual.



Gambar 11. Koneksi antara File dan Scatter Chart

Setelah koneksi dibuat, klik dua kali Scatter Plot untuk membuka tampilan grafik. Di sini pengguna bisa memilih variabel X dan Y secara bebas.

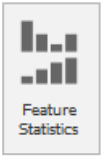


Gambar 12. Scatter Plot untuk Data Medical Cost Insurance

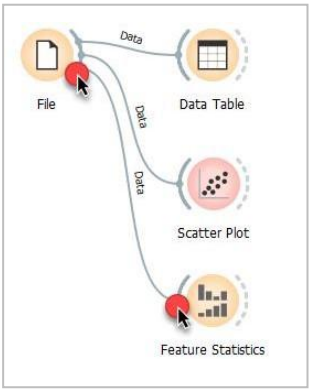
Sebagai contoh, jika kita memilih Smoker sebagai sumbu X dan BMI sebagai sumbu Y, kita dapat melihat bahwa individu dengan BMI tinggi dan kebiasaan merokok (Smoker = Yes) cenderung memiliki biaya medis (Charges) yang lebih tinggi dibandingkan mereka yang tidak merokok.

2. Statistik Setiap Atribut

Selain visualisasi, kita juga dapat memahami data melalui perhitungan statistik deskriptif. Orange menyediakan fitur **Feature Statistics** yang menampilkan ringkasan statistik dari setiap atribut.

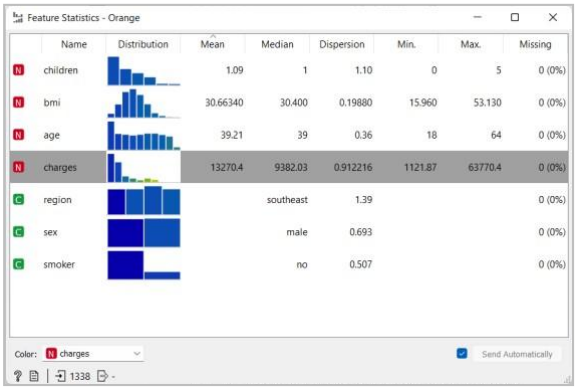


Tambahkan ikon Feature Statistics dari menu *Data* ke worksheet, lalu hubungkan dengan ikon File.



Gambar 13. Koneksi File dengan Feature Statistics

Klik dua kali ikon tersebut untuk melihat grafik batang serta nilai statistik seperti mean, median, minimum, maksimum, dan range dari masing-masing fitur.



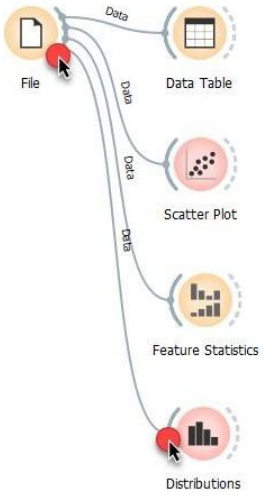
Gambar 14. Hasil Statistik Setiap Atribut

Dengan informasi ini, pengguna bisa melihat variasi data dan memutuskan apakah perlu dilakukan normalisasi atau transformasi lebih lanjut sebelum membuat model.

3. Distribusi Data

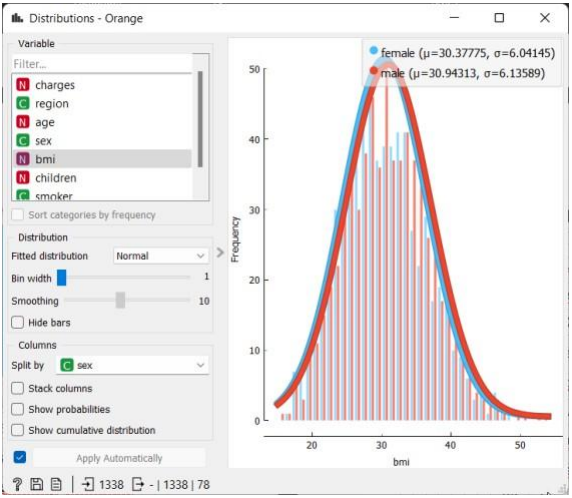
Untuk memperdalam analisis, kita dapat memanfaatkan fitur **Distributions** yang menyajikan sebaran nilai dari setiap atribut dalam bentuk diagram batang.

Tambahkan ikon Distribution dari menu *Visualize ke worksheet* dan hubungkan dengan ikon File dataset.



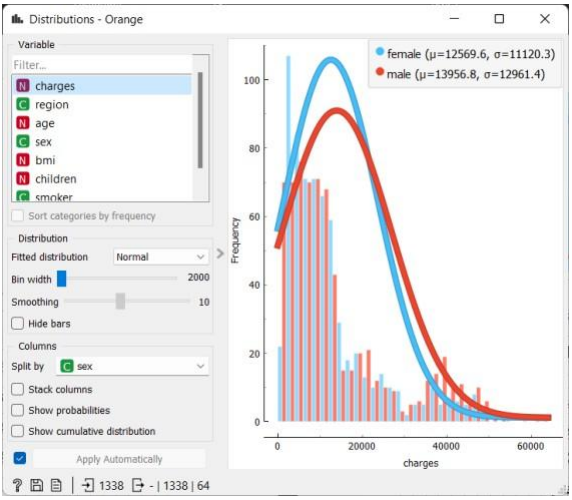
Gambar 15. Koneksi File dengan *Distributions*

Klik dua kali pada Distribution untuk menampilkan diagram sebaran nilai atribut seperti BMI berdasarkan kategori seperti jenis kelamin.



Gambar 16. Distribusi Data BMI berdasarkan Jenis Kelamin

Pengguna juga dapat menyaring dan memilih variabel yang ingin dianalisis lebih lanjut. Misalnya, kita dapat melihat bahwa distribusi biaya medis cenderung lebih tinggi pada laki-laki dengan BMI ekstrem.

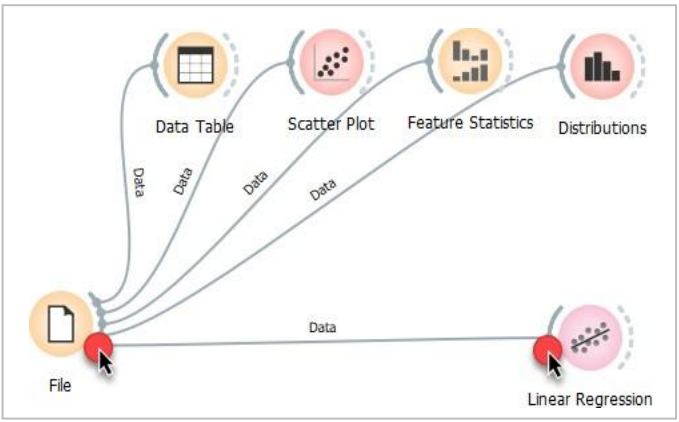


Gambar 17. Kemiringan *Data Medical Cost* Berdasar Jenis Kelamin

C. Pembuatan Model

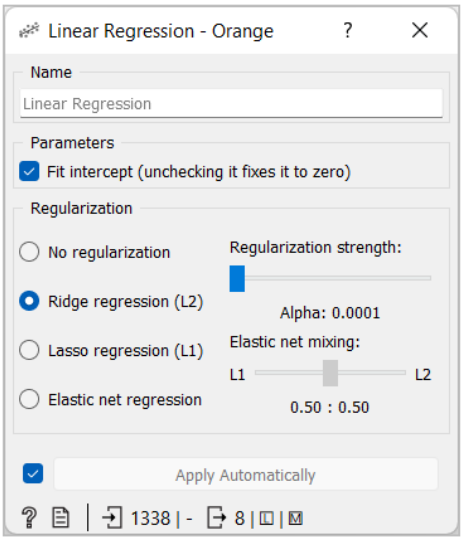
Langkah selanjutnya adalah membangun model regresi menggunakan Orange. Untuk kasus ini, kita menggunakan algoritma *Linear Regression*, yang sangat cocok digunakan untuk data dengan hubungan linier antara input dan output.

Dari menu *Model*, tarik ikon **Linear Regression** ke worksheet.



Gambar 18. Koneksi File Dataset dengan Linear Regression Model

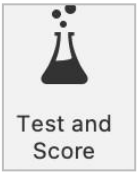
Hubungkan ikon File ke ikon Linear Regression untuk mengalirkan data ke model. Klik dua kali ikon model jika ingin mengatur parameter, meskipun konfigurasi default sudah cukup untuk pembelajaran dasar.



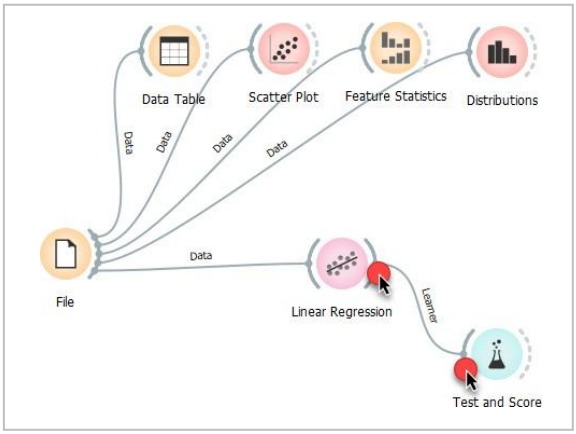
Gambar 19. Setting Parameter untuk Linear Regression

D. Evaluasi

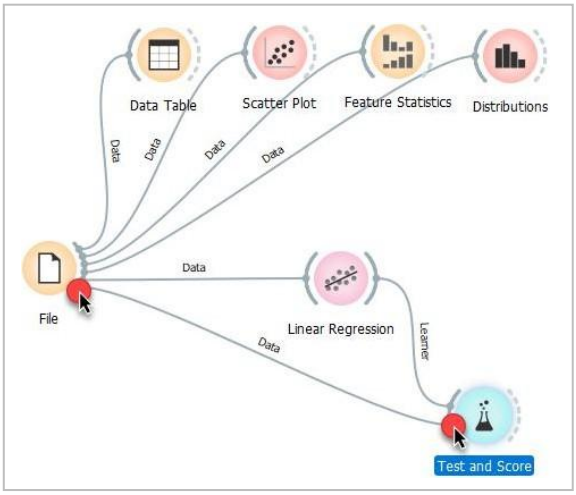
Setelah model dibuat, kita perlu mengetahui seberapa baik performa model tersebut dalam membuat prediksi. Orange menyediakan widget **Test and Score** untuk tujuan evaluasi ini.



Tambahkan ikon Test and Score dari menu *Evaluate* ke worksheet, lalu hubungkan dengan model Linear Regression dan dataset File.



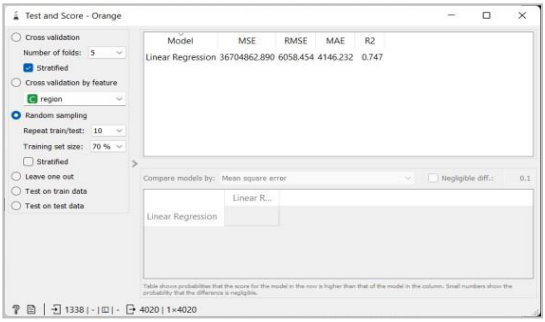
Gambar 20. Hubungan antara Model Prediksi dengan Evaluasi Test and Score



Gambar 21. Hubungan Dataset dengan Test and Score

Klik dua kali ikon Test and Score untuk melihat hasil evaluasi. Metrik evaluasi yang tersedia meliputi MAE

(*Mean Absolute Error*), MSE (*Mean Squared Error*), RMSE (*Root Mean Squared Error*), dan R^2 (*R-Squared*).



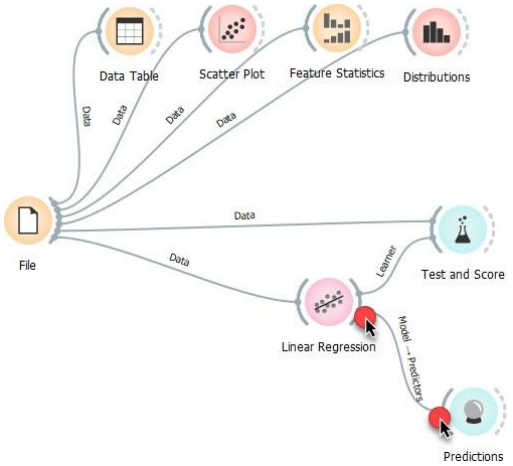
Gambar 22. Hasil Evaluasi Model untuk Linear Regression

Nilai-nilai ini membantu kita menilai sejauh mana model regresi berhasil meniru data aktual dan memberikan gambaran objektif tentang akurasi prediksi.

E. Prediksi

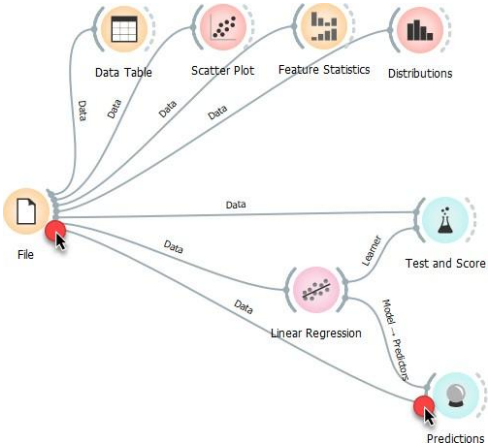
Langkah terakhir dalam siklus regresi adalah membuat prediksi berdasarkan model yang sudah dibuat dan diuji. Prediksi ini dilakukan pada data yang belum digunakan selama pelatihan.

Tambahkan ikon **Predictions** dari menu *Evaluate* ke worksheet.



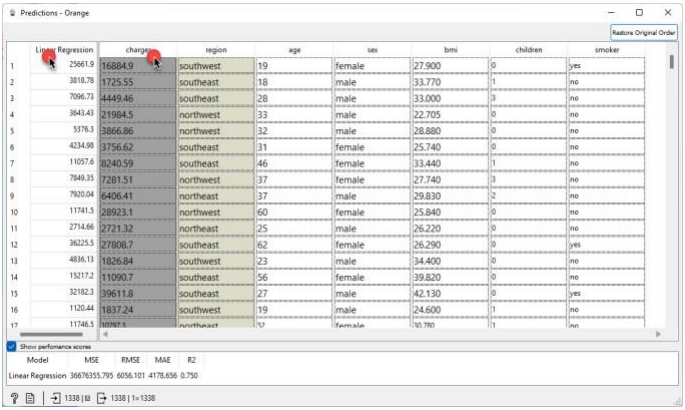
Gambar 23. Link Model dan Prediction

Hubungkan ikon model dan ikon File ke ikon Predictions.



Gambar 24. Koneksi antara File dengan Prediction

Klik dua kali ikon Predictions untuk melihat hasil output berupa estimasi biaya asuransi kesehatan (Charges) untuk setiap baris data.



	Linear Regression	charges	region	age	sex	bmi	children	smoker
1	25661.9	16884.9	southwest	19	female	27.900	0	yes
2	3818.78	1725.55	southeast	18	male	33.770	1	no
3	7096.73	4449.46	southeast	28	male	33.000	3	no
4	3643.43	21984.5	northwest	33	male	22.705	0	no
5	5376.3	3866.86	northwest	32	male	28.880	0	no
6	4234.98	3756.62	southeast	31	female	25.740	0	no
7	11057.6	8240.59	southeast	46	female	33.440	1	no
8	7849.35	7281.51	northwest	37	female	27.740	3	no
9	7920.04	6406.41	northeast	37	male	29.830	2	no
10	11741.5	28923.1	northwest	60	female	25.840	0	no
11	2714.66	2721.32	northeast	25	male	26.220	0	no
12	36225.5	27808.7	southeast	62	female	26.290	0	yes
13	4836.13	1826.84	southwest	23	male	34.400	0	no
14	15217.2	11090.7	southeast	56	female	39.820	0	no
15	32182.3	39611.8	southeast	27	male	42.130	0	yes
16	1120.44	1837.24	southwest	19	male	24.600	1	no
17	11746.5	10767.1	northeast	52	female	30.780	1	no

Gambar 25. Hasil Prediksi Charges dengan Linear Regression

Prediksi ini bisa dimanfaatkan oleh praktisi atau pemangku kebijakan untuk menyusun strategi pelayanan atau penyesuaian premi asuransi yang lebih tepat sasaran berdasarkan profil risiko individu.

Bab 3:
Regresi Harga Rumah

Perkiraan harga rumah menjadi salah satu kebutuhan penting dalam perencanaan keuangan, pengambilan keputusan investasi, hingga penyusunan kebijakan publik di sektor perumahan. Dalam dunia nyata, harga rumah ditentukan oleh berbagai faktor seperti ukuran bangunan, jumlah kamar, lokasi terhadap fasilitas umum, hingga kualitas lingkungan. Namun, sering kali keputusan mengenai harga rumah menjadi bias atau kurang objektif apabila hanya didasarkan pada intuisi atau pendekatan konvensional. Oleh karena itu, diperlukan model prediktif berbasis data yang dapat memberikan estimasi harga secara lebih terukur dan berbasis bukti. Dalam bab ini, akan diuraikan bagaimana pendekatan regresi linier dapat digunakan untuk memprediksi harga rumah melalui studi kasus pada dataset Boston House Prices, dengan bantuan Orange Data Mining sebagai alat visualisasi dan pemodelan.

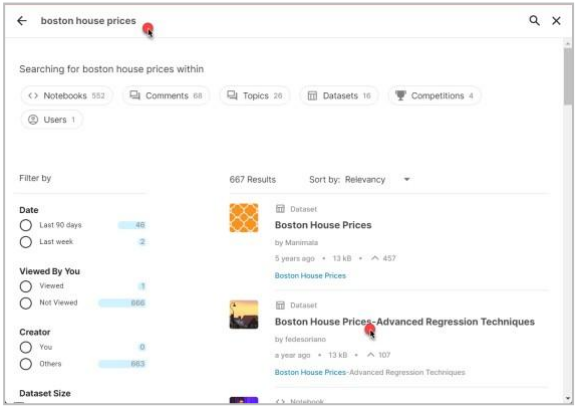
A. Akuisisi Data

Tahap pertama dalam analisis regresi adalah mengumpulkan data yang relevan dan dapat dipercaya. Proses ini bukan sekadar mengunduh data, tetapi juga memastikan bahwa data tersebut berasal dari sumber yang kredibel dan sesuai dengan tujuan analisis. Dalam konteks akademik dan pembelajaran, sumber data terbuka (*open source*) menjadi pilihan utama karena

dapat diakses secara gratis dan transparan oleh siapa saja.

1. Pengambilan Data dari Kaggle

Kaggle merupakan platform global yang banyak digunakan oleh komunitas data scientist dan peneliti untuk berbagi dataset, berpartisipasi dalam kompetisi analisis data, serta berdiskusi mengenai teknik machine learning. Untuk kasus regresi harga rumah, kita akan menggunakan dataset klasik *Boston House Prices* yang telah diolah dengan teknik regresi lanjutan. Langkah awal adalah membuka situs https://kaggle.com, lalu melakukan pendaftaran atau login bagi pengguna yang sudah memiliki akun.

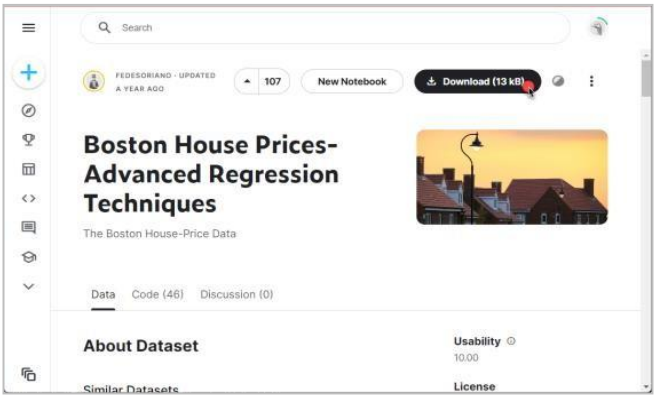


Gambar 26. Laman Pencarian Dataset di Kaggle

Setelah berhasil masuk ke platform, pengguna dapat melakukan pencarian dengan mengetikkan kata kunci

"*Boston House Prices – Advanced Regression Techniques*". Pilih dataset yang dipublikasikan oleh "*Fedesoriano*" karena versi ini telah dirapikan dan siap untuk dianalisis. Dataset ini menjadi sumber belajar yang sangat kaya karena mencakup banyak variabel yang berpotensi memengaruhi harga rumah, seperti jumlah kamar, tingkat kejahatan di lingkungan, akses ke jalan utama, dan lain sebagainya.

Langkah berikutnya adalah menekan tombol *Download* untuk mengunduh dataset. Setelah proses pengunduhan selesai, file yang diperoleh dalam format .zip perlu diekstrak di komputer. Proses ekstraksi ini akan menghasilkan file dengan format .csv (Comma-Separated Values) yang siap untuk diolah menggunakan perangkat lunak Orange.

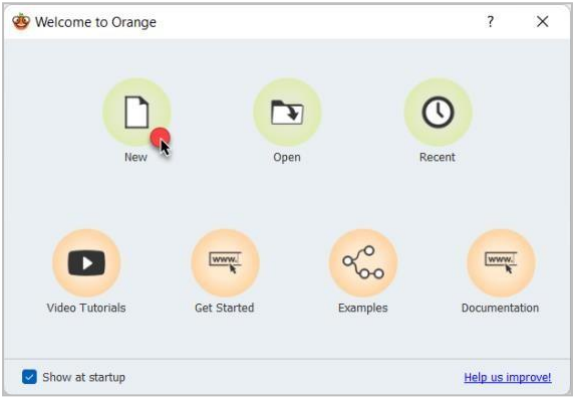


Gambar 27. Laman Dataset Boston House Prices

2. Set-up Data di Orange

Setelah data berhasil diunduh dan disiapkan, kini saatnya melakukan konfigurasi awal di dalam Orange Data Mining. *Proses set-up* ini sangat penting karena akan menentukan bagaimana data dikenali dan diolah lebih lanjut oleh sistem visual dalam Orange.

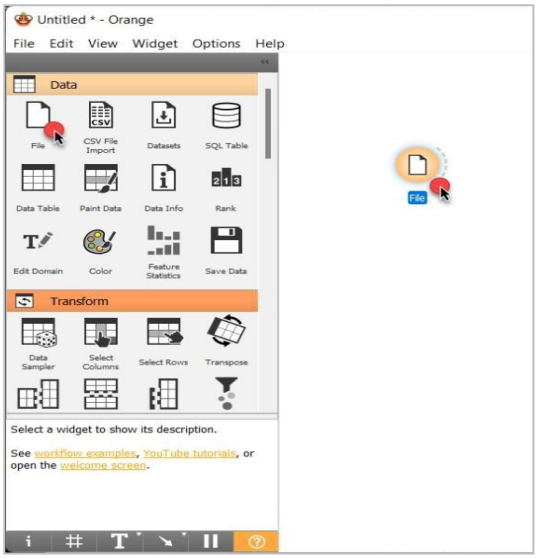
Langkah pertama adalah membuka aplikasi Orange dan memilih tombol **New** untuk memulai proyek analisis baru. Proyek ini akan menjadi tempat semua ikon (widget) saling terhubung dalam satu siklus kerja yang utuh.



Gambar 28. Tampilan Depan dari Aplikasi Orange

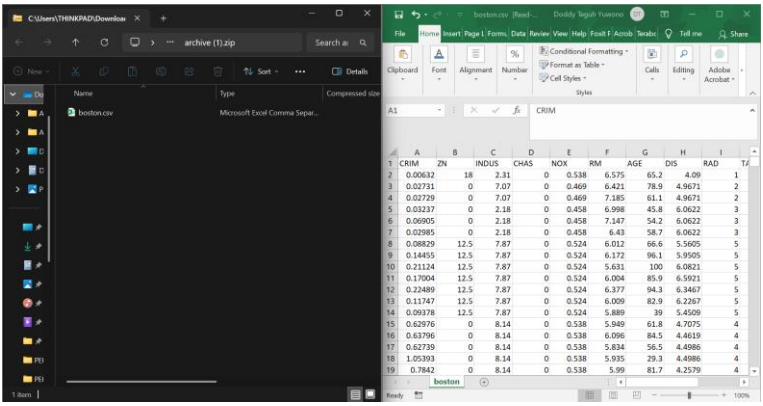
Selanjutnya, tarik ikon **File** dari menu *Data* ke worksheet. Ikon ini akan menjadi penghubung utama antara Orange dengan file dataset yang telah disiapkan. Klik dua kali

ikon tersebut untuk membuka jendela pemilihan file dari komputer.

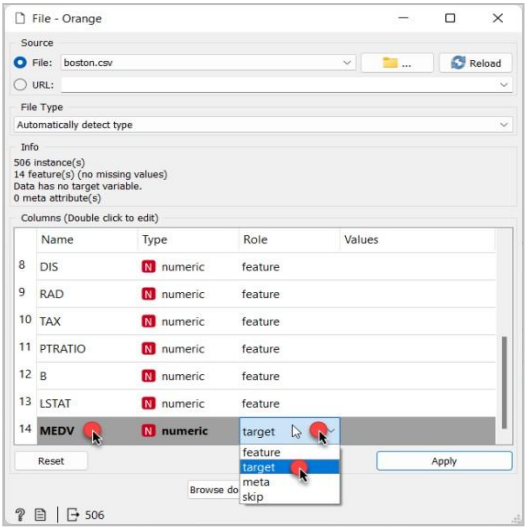


Gambar 29. Memasukkan Ikon File pada Worksheet

Setelah file dipilih, pengguna akan melihat tampilan pengaturan atribut. Klik ganda ikon file, dan pastikan bahwa atribut “MEDV” (Median value of owner-occupied homes) diset sebagai target variabel. Ini penting karena MEDV adalah variabel yang ingin kita prediksi.



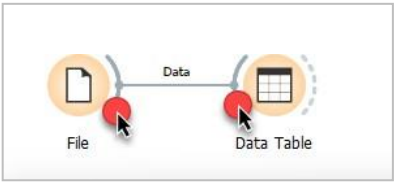
Gambar 30. Tampilan Pencarian File dalam Komputer



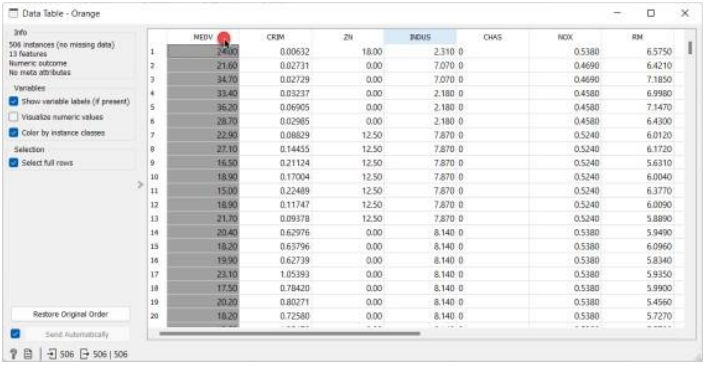
Gambar 31. *Set-up Atribut* Target dari Dataset

Untuk meninjau data secara menyeluruh, tambahkan ikon **Data Table** dari menu *Data* ke worksheet. Buat

koneksi dari ikon File ke Data Table. Setelah itu, klik dua kali ikon Data Table untuk melihat seluruh baris dan kolom data dalam format tabel.



Gambar 32. Pembuatan Link pada File dan Data Table



Gambar 33. List Data Boston House Prices

Dataset ini terdiri dari 506 entri dan 13 fitur atau atribut yang memengaruhi harga rumah. Tidak terdapat nilai kosong, sehingga kita bisa langsung melanjutkan ke tahap pemahaman data tanpa perlu proses imputasi.

B. Pemahaman Data

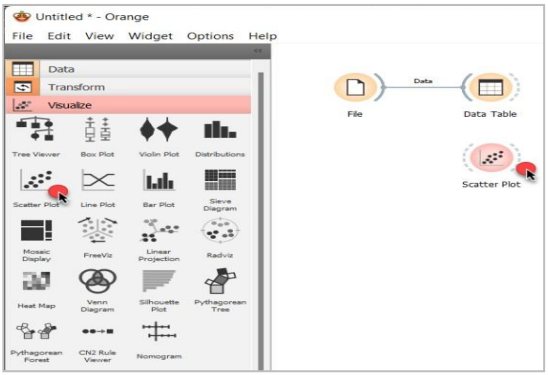
Sebelum melangkah ke tahap pemodelan, penting bagi analis untuk benar-benar memahami struktur dan makna

dari data yang akan digunakan. Pemahaman ini tidak hanya melalui deskripsi teks, tetapi juga melalui pendekatan visual dan statistik.

1. Visualisasi Data dengan Scatter Chart

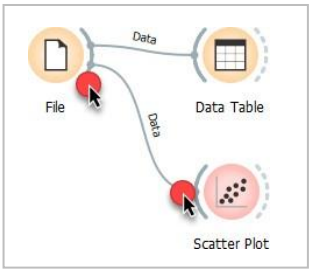
Scatter chart merupakan alat visual yang sangat efektif untuk melihat hubungan linier atau non-linier antara dua variabel. Dengan scatter chart, kita dapat mengenali korelasi yang kuat, lemah, atau bahkan tidak ada korelasi antar variabel.

Tambahkan ikon **Scatter Plot** dari menu *Visualize* ke worksheet Orange.

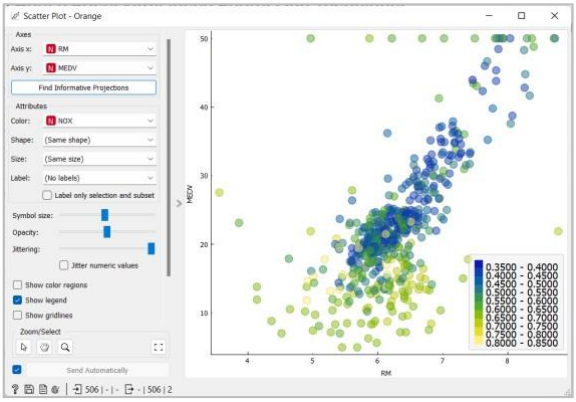


Gambar 34. Pemilihan Scatter Chart untuk Visualisasi Data

Buat koneksi antara ikon File dan Scatter Plot, lalu buka tampilannya dengan double click.



Gambar 35. Koneksi antara File dan Scatter Chart



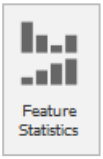
Gambar 36. Scatter Plot untuk Data Boston House Prices

Sebagai contoh, ketika RM (jumlah ruang) dipasang sebagai X axis dan MEDV sebagai Y axis, terlihat bahwa rumah dengan jumlah ruang lebih banyak cenderung memiliki nilai MEDV yang lebih tinggi. Sementara itu, jika kita gunakan NOX (tingkat polusi udara) sebagai variabel, maka pola hubungan negatif muncul — semakin tinggi NOX, nilai MEDV cenderung lebih rendah. Hal ini

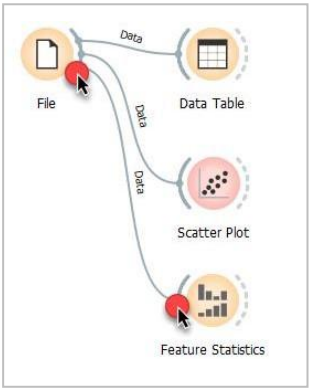
memperlihatkan bahwa lokasi dan kondisi lingkungan memengaruhi nilai properti secara signifikan.

2. Statistik Setiap Atribut

Selain visualisasi, Orange juga memungkinkan pengguna menghitung statistik deskriptif dari setiap atribut. Ini sangat berguna untuk mengetahui ukuran sentral (mean, median) dan sebaran (min, max, range).

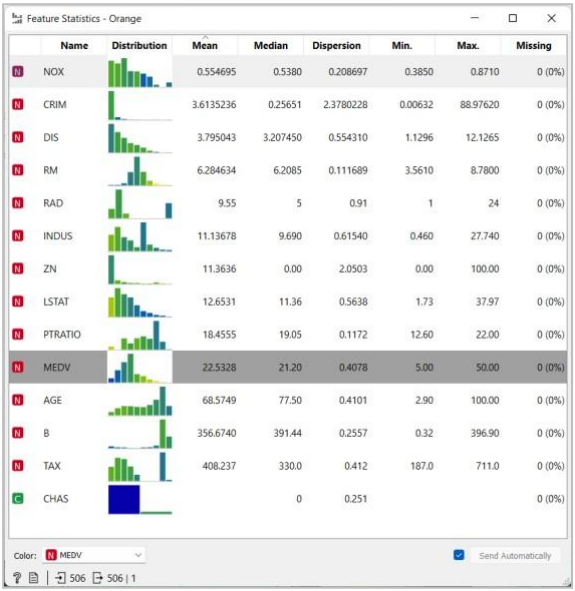


Tambahkan ikon **Feature Statistics** ke worksheet dan hubungkan dengan ikon File.



Gambar 37. Koneksi File dengan Feature Statistics

Klik dua kali ikon tersebut untuk membuka tampilan grafik batang dan tabel statistik atribut.



Gambar 38. Hasil Statistik Setiap Atribut

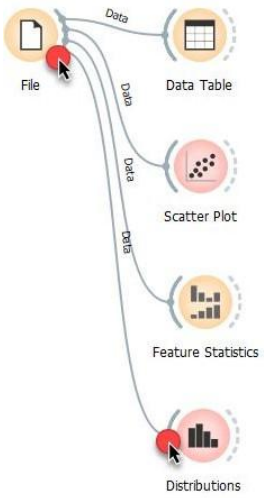
Informasi ini sangat bermanfaat untuk memahami kondisi data yang digunakan. Misalnya, jika ditemukan atribut dengan variasi sangat kecil, maka fitur tersebut mungkin kurang berkontribusi dalam prediksi dan dapat dipertimbangkan untuk dikeluarkan.

3. Distribusi Data

Untuk memahami persebaran nilai dari setiap atribut, kita dapat menggunakan fitur **Distributions**. Ini

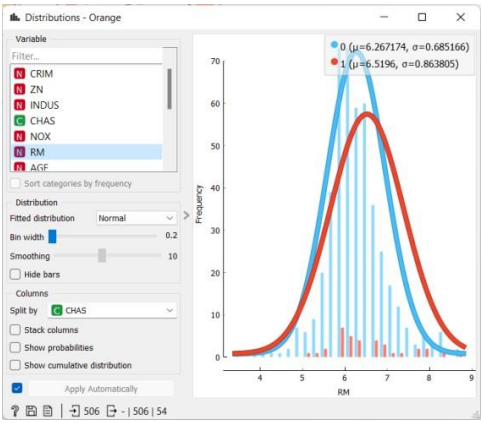
memungkinkan kita melihat seberapa seimbang data yang kita miliki, baik secara numerik maupun kategorikal.

Tambahkan ikon Distribution ke worksheet dan hubungkan dengan ikon File.



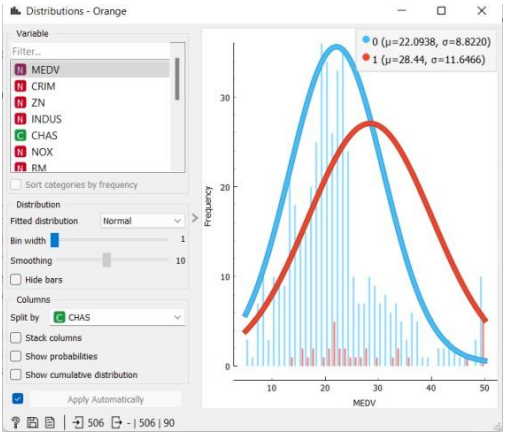
Gambar 39. Koneksi File dengan Distributions

Klik dua kali ikon Distribution untuk melihat distribusi RM berdasarkan variabel CHAS, yang menunjukkan apakah rumah terletak di dekat sungai Charles.



Gambar 40. Distribusi Data RM berdasarkan CHAS

Distribusi ini memberikan gambaran apakah harga rumah berbeda antara properti yang berlokasi dekat sungai dan yang tidak. Ini membantu kita menilai apakah variabel CHAS memiliki pengaruh signifikan dalam model prediktif.

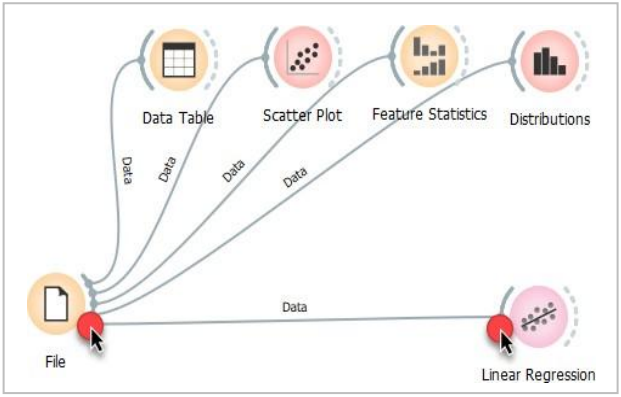


Gambar 41. Kemiringan Data MEDV Berdasarkan CHAS

C. Pembuatan Model

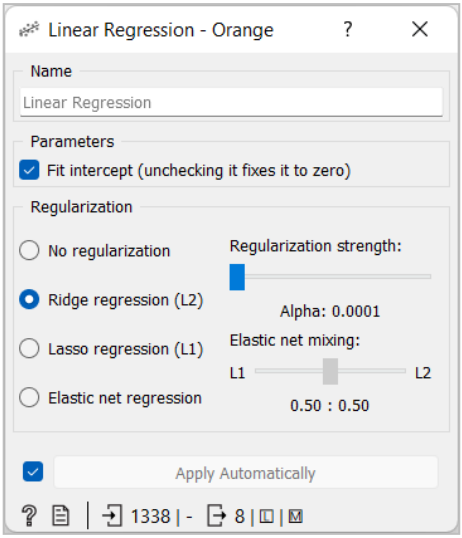
Dengan pemahaman data yang matang, kini saatnya membangun model prediksi. Dalam kasus ini, kita menggunakan algoritma **Linear Regression** karena sifatnya yang sederhana namun efektif dalam memodelkan hubungan linier.

Tambahkan ikon Linear Regression dari menu *Model* ke worksheet.



Gambar 42. Koneksi File Dataset dengan Linear Regression Model

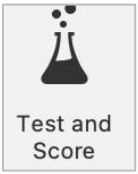
Hubungkan dengan ikon File dan klik dua kali untuk membuka pengaturan. Kita dapat menyesuaikan parameter seperti alpha, meskipun untuk tahap awal parameter default sudah mencukupi.



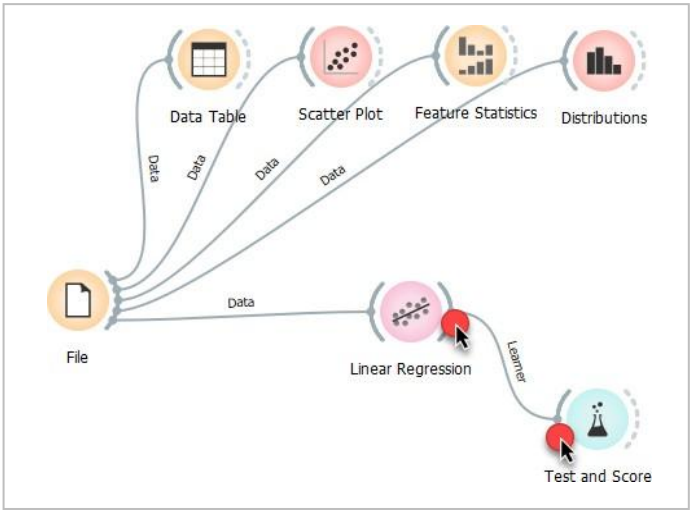
Gambar 43. Setting Parameter untuk Linear Regression

D. Evaluasi

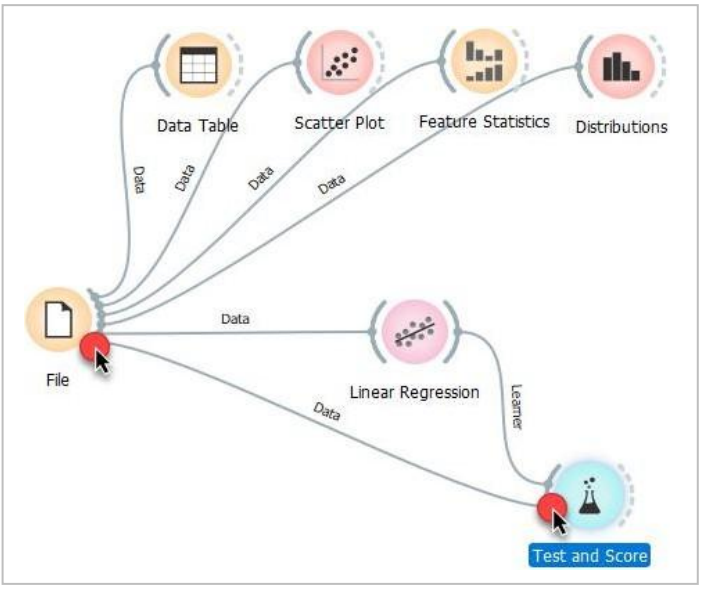
Setiap model yang dibangun harus diuji performanya melalui evaluasi. Tanpa evaluasi, kita tidak dapat mengetahui sejauh mana prediksi model dapat dipercaya.



Tambahkan ikon **Test and Score** dari menu *Evaluate* ke worksheet dan sambungkan dengan ikon Model serta ikon File.

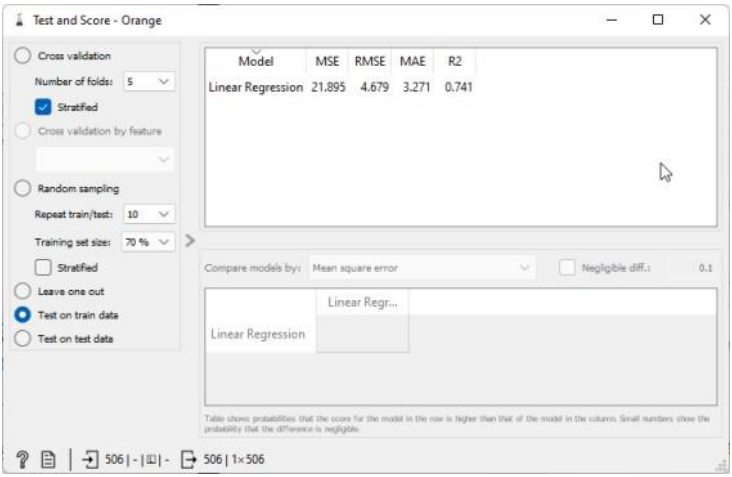


Gambar 44. Hubungan antara Model Prediksi dengan Evaluasi Test and Score



Gambar 45. Hubungan Dataset dengan Test and Score

Klik dua kali ikon Test and Score untuk melihat metrik evaluasi. Metrik yang digunakan termasuk MSE, MAE, RMSE, dan R^2, yang masing-masing menunjukkan tingkat kesalahan dan kemampuan model dalam menjelaskan variansi data.

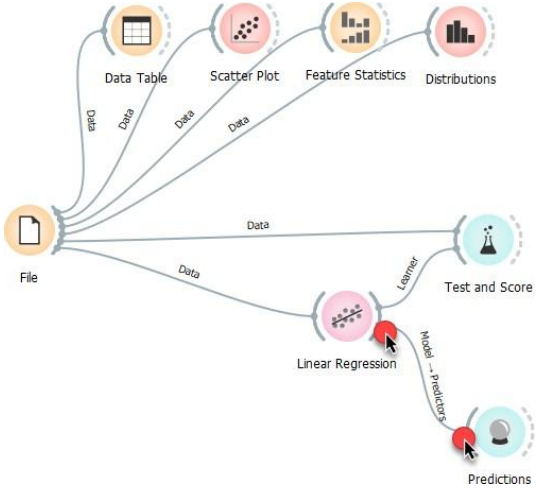


Gambar 46. Hasil Evaluasi Model untuk Linear Regression

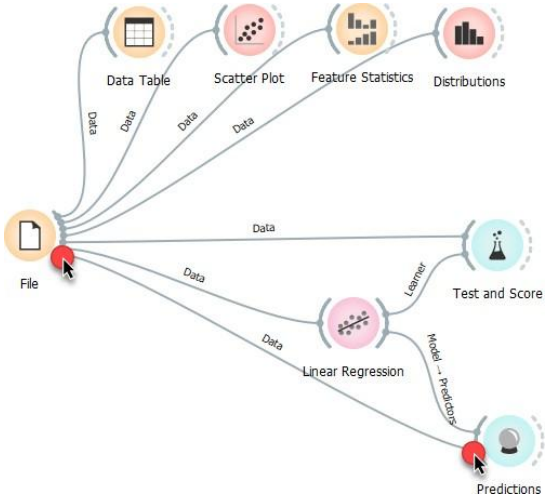
E. Prediksi

Langkah akhir dalam proses analisis ini adalah menerapkan model yang telah dievaluasi untuk melakukan prediksi harga rumah. Ini memungkinkan pengguna memproyeksikan nilai properti baru berdasarkan karakteristiknya.

Tambahkan ikon **Predictions** ke worksheet dan sambungkan dengan Model dan File.

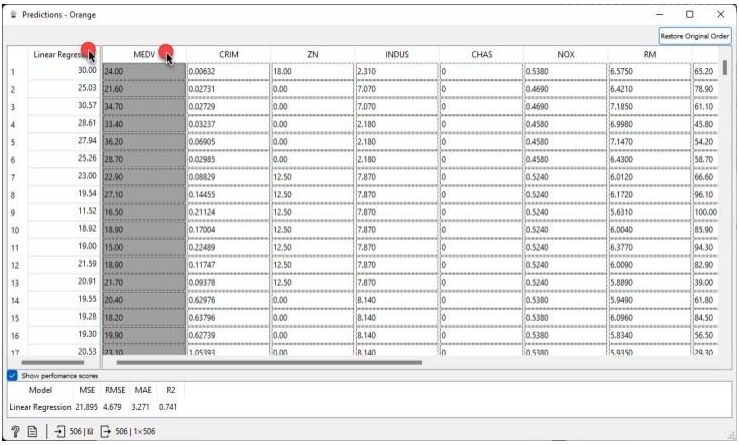


Gambar 47. Link Model dan Prediction



Gambar 48. Koneksi antara File dengan Prediction

Klik dua kali ikon Predictions untuk menampilkan hasil prediksi.



Gambar 49. Hasil Prediksi MEDV dengan Linear Regression

Hasil prediksi ini sangat berguna bagi agen properti, pengembang, dan pembeli rumah untuk membuat keputusan berbasis data dan bukan sekadar intuisi.

Bab 4
Latihan Soal Regresi

Latihan dalam proses pembelajaran bukan hanya sarana untuk menguji pemahaman terhadap materi, tetapi juga menjadi ruang aktualisasi bagi mahasiswa untuk membangun kompetensi dalam situasi yang lebih otonom. Pada bab ini, latihan regresi disusun agar mahasiswa dapat melatih pemahaman mereka terhadap konsep analisis data dan pengembangan model prediktif berbasis regresi. Dengan menggunakan dua sumber data yang berbeda—*US Births 2018* dan *Auto MPG Dataset*—mahasiswa didorong untuk mengalami sendiri bagaimana siklus data science bekerja, mulai dari akuisisi data, pemahaman data, pemodelan, hingga refleksi terhadap hasil prediksi. Pembelajaran tidak lagi bersifat top-down, melainkan bersifat reflektif dan partisipatif. Mahasiswa bukan sekadar menyimak, tetapi melakukan, menemukan, dan menilai.

A. Latihan 1: Prediksi Berat Badan Bayi – US Births 2018

Permasalahan gizi dan kesehatan bayi merupakan isu yang sangat penting dan berdampak luas. Salah satu indikator kesehatan bayi yang paling sering digunakan adalah berat badan saat lahir. Oleh karena itu, prediksi berat badan bayi berdasarkan faktor-faktor kehamilan dan demografi ibu bisa menjadi masukan penting dalam perencanaan kebijakan dan intervensi kesehatan publik. Pada latihan ini, mahasiswa akan bekerja menggunakan dataset *US Births (2018)* dari Kaggle untuk memprediksi berat badan bayi berdasarkan atribut-atribut yang tersedia. Proyek ini tidak hanya bersifat teknis, tetapi juga memupuk kepekaan terhadap isu-isu sosial dan kesehatan yang aktual.

Langkah 1: Akuisisi Dataset

Langkah pertama dimulai dengan mengakses platform *Kaggle* dan mencari dataset berjudul “US Births (2018)” yang dipublikasikan oleh “U.S. *Government Works*”. Dataset ini bersifat publik dan bebas digunakan untuk kegiatan akademik. Mahasiswa diharapkan dapat

mengunduh file data, mengekstraknya, dan menyiapkannya dalam format .csv. Proses ini adalah latihan awal yang sederhana namun penting, karena mengajarkan tanggung jawab dalam memilih data yang kredibel serta memahami etika penggunaan data terbuka.

Langkah 2: Menyiapkan Dataset di Orange

Setelah data siap, bukalah Orange dan buat proyek baru. Masukkan dataset menggunakan ikon **File** lalu hubungkan ke **Data Table** untuk melihat isi datanya. Pastikan seluruh kolom terbaca dengan baik. Proses ini akan melatih keterampilan visualisasi struktur data awal sebelum melakukan transformasi atau analisis lebih lanjut.

Langkah 3: Eksplorasi Visualisasi – *Scatter Plot*

Gunakan ikon **Scatter Plot** untuk memetakan hubungan antara berat badan bayi dan atribut lain, seperti usia ibu atau lama kehamilan. Mahasiswa didorong untuk menggunakan fitur *color*, *size*, dan *shape* agar visualisasi lebih bermakna. Dalam visualisasi ini, mahasiswa akan

mengasah intuisi mereka untuk membaca pola, outlier, dan korelasi visual—sebuah keterampilan yang esensial dalam dunia analitik data.

Langkah 4: Distribusi Data

Gunakan widget **Distribution** untuk mengeksplorasi sebaran data. Identifikasi atribut mana yang mendekati distribusi normal dan mana yang tidak. Diskusikan bagaimana distribusi ini bisa mempengaruhi hasil model regresi yang akan dibangun. Di sinilah mahasiswa belajar bahwa regresi tidak bisa lepas dari pemahaman statistik dasar.

Langkah 5: Pembuatan dan Evaluasi Model

Selanjutnya, buatlah model **Linear Regression** dan evaluasi performanya menggunakan **Test and Score**. Pelajari nilai MSE, RMSE, MAE, dan R^2. Apakah model cukup akurat? Adakah fitur yang menyebabkan error tinggi? Proses ini mendorong mahasiswa mengembangkan sikap kritis terhadap hasil model dan tidak semata-mata menerima angka tanpa refleksi.

Langkah 6: Melakukan Prediksi

Tambahkan widget **Predictions** dan lihat bagaimana model memperkirakan berat badan bayi dari fitur input. Bandingkan nilai prediksi dengan nilai aktual. Apakah terdapat deviasi besar? Apa yang menyebabkannya? Pertanyaan-pertanyaan ini perlu dijawab bukan hanya dari angka, tetapi dari narasi data yang terbaca.

Langkah 7: Refleksi dan Dokumentasi

Sebagai penutup, mahasiswa harus menulis hasil refleksi mereka dalam bentuk catatan atau laporan. Fokuskan pada tiga aspek: proses (apa yang dilakukan), hasil (apa yang ditemukan), dan pemaknaan (apa artinya dalam konteks dunia nyata). Dengan cara ini, latihan menjadi bagian dari perjalanan belajar, bukan sekadar tugas.

B. Latihan 2: Regresi Linier Berganda – Dataset Auto MPG

Latihan kedua menantang mahasiswa untuk menerapkan regresi linier berganda pada data otomotif. Dataset *Auto MPG* dari UC Irvine Machine Learning Repository berisi data historis kendaraan yang menggambarkan hubungan antara karakteristik mesin dan efisiensi bahan bakar

(*miles per gallon*). Proyek ini membuka ruang belajar interdisipliner antara teknologi, lingkungan, dan kebijakan transportasi berkelanjutan.

Struktur Dataset:

No	Atribut	Tipe Data
1	mpg	Continuous
2	cylinders	Discrete
3	displacement	Continuous
4	horsepower	Continuous
5	weight	Continuous
6	acceleration	Continuous
7	model year	Discrete
8	origin	Discrete
9	car name	Text (Unique)

Langkah 1: Load Dataset dan Identifikasi Masalah

Buka Orange, pilih **Datasets**, dan cari dataset *Auto MPG*. Amati bahwa terdapat missing values, khususnya pada kolom *horsepower*. Ini adalah tantangan nyata yang sering ditemui di dunia data science, dan mahasiswa harus belajar menyikapinya secara metodologis.

Langkah 2: Menangani Missing Values

Gunakan widget **Impute** dan pilih metode rata-rata (*Average*) untuk data numerik atau modus (*Most Frequent*) untuk data kategori. Imputasi tidak hanya soal mengganti nilai kosong, tapi juga mempertimbangkan logika dan dampak statistiknya terhadap model.

Langkah 3: Verifikasi Imputasi

Hubungkan **Data Table** setelah Impute untuk memastikan proses berjalan baik. Perhatikan apakah ada pola nilai ganjil atau perubahan distribusi akibat imputasi.

Langkah 4: Eksplorasi Statistik dan Korelasi

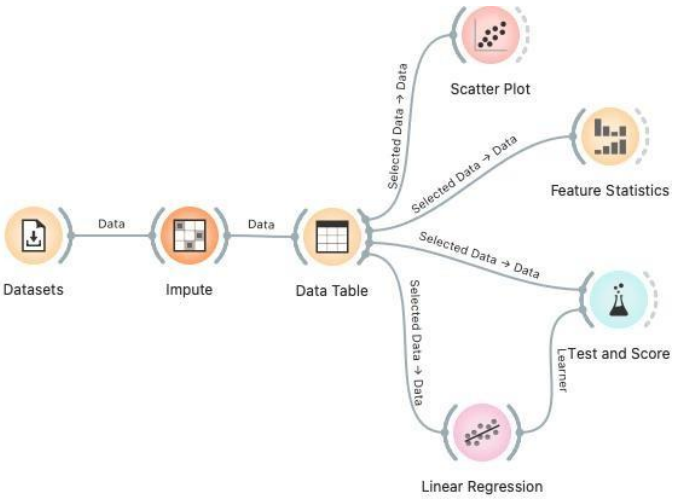
Gunakan **Feature Statistics** untuk melihat nilai pusat dan sebaran atribut. Kemudian, gunakan **Scatter Plot** untuk mengamati hubungan antara mpg dan fitur lainnya, terutama *weight* yang secara logika sangat berpengaruh. Mahasiswa diharapkan mampu mengenali pola korelasi dan menjelaskan implikasi praktisnya.

Langkah 5: Membangun dan Mengevaluasi Model

Bangun model **Linear Regression** dengan semua fitur, dan evaluasi menggunakan **Test and Score**. Nilai evaluasi seperti R^2 akan menunjukkan seberapa baik fitur-fitur tersebut menjelaskan variasi pada *mpg*.

Langkah 6: Interpretasi Hasil

Interpretasi bukan hanya soal nilai R^2 yang tinggi atau rendah. Mahasiswa harus mengevaluasi: fitur mana yang paling relevan? Adakah multikolinearitas? Apakah data outlier merusak model?



Gambar 50. Skema Orange Data Mining untuk Latihan Soal Regresi 2

Refleksi Penutup

Bab ini mengajak mahasiswa untuk berpindah dari pembelajaran pasif menjadi aktor aktif dalam eksplorasi data. Melalui dua latihan yang berbeda konteks kesehatan masyarakat dan efisiensi kendaraan mahasiswa diajak melihat bahwa data dan regresi bukan sekadar urusan teknis, tetapi juga menyentuh kebijakan, etika, dan keberlanjutan.

Kemampuan regresi yang sesungguhnya bukan hanya tentang bisa mengklik ikon model dan membaca metrik evaluasi. Ia melibatkan kemampuan memahami konteks, menimbang relevansi variabel, menangani ketidaksempurnaan data, serta menyampaikan hasil prediksi dalam bahasa yang bisa dimengerti dan bermanfaat.

Latihan dalam bab ini harus dibaca sebagai ruang pembelajaran reflektif. Tidak ada satu jawaban benar yang mutlak, karena setiap data memiliki cerita, dan setiap cerita butuh kepekaan, bukan hanya logika.

Bab 5
Klasifikasi Kanker Payudara

Kanker payudara merupakan salah satu isu kesehatan global yang berdampak luas, terutama bagi perempuan. Deteksi dini menjadi faktor kunci dalam meningkatkan harapan hidup dan efektivitas pengobatan. Di tengah kemajuan teknologi data, pendekatan klasifikasi berbasis machine learning menawarkan solusi potensial dalam membantu proses diagnosis. Dalam bab ini, kita akan mempelajari bagaimana memanfaatkan Orange Data Mining untuk mengembangkan sistem klasifikasi guna memprediksi apakah suatu sel bersifat ganas (malignant) atau jinak (benign), melalui tahapan yang terstruktur dari akuisisi data, eksplorasi visual, pemodelan, hingga prediksi dan evaluasi. Lebih dari sekadar penerapan teknis, pembelajaran ini akan mendorong mahasiswa untuk membangun sensitivitas terhadap penggunaan data medis secara etis dan bertanggung jawab.

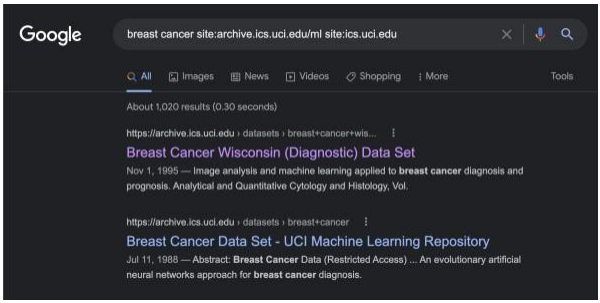
A. Akuisisi Data

Tahapan awal dalam proses klasifikasi adalah mengumpulkan data yang akurat dan representatif. Dalam konteks kanker payudara, kualitas data sangat menentukan keandalan model prediksi. Oleh karena itu, penting untuk memanfaatkan dataset terbuka dari sumber terpercaya.

1. Pengambilan Data dari UCI Repository

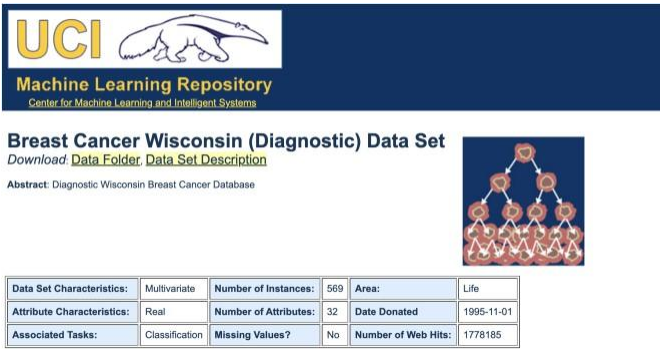
UCI Machine Learning Repository adalah sumber data terpercaya yang banyak digunakan dalam penelitian dan pembelajaran machine learning. Untuk mengaksesnya, buka laman https://archive.ics.uci.edu/ml/index.php.

Masukkan kata kunci “breast cancer” pada kolom pencarian, lalu klik tombol *Search*. Anda akan diarahkan ke halaman pencarian Google yang menampilkan dataset yang relevan.



Gambar 51. Laman Hasil Pencarian

Klik tautan teratas bertuliskan **Breast Cancer Wisconsin Dataset**. Halaman tersebut memuat informasi rinci tentang struktur dataset, jumlah atribut, dan deskripsi setiap fitur.



Gambar 52. Laman Dataset Kanker Payudara UCI Repository

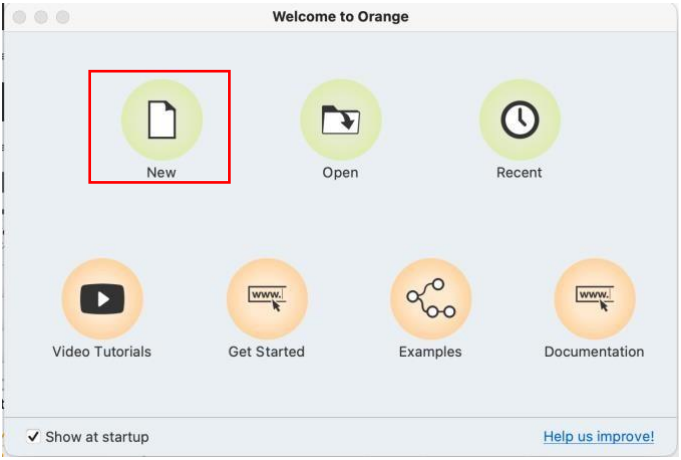
Selanjutnya, klik menu **Data Folder**, dan unduh file berformat .data. Dataset ini merupakan hasil observasi

terhadap sel kanker dan sangat ideal untuk pelatihan model klasifikasi.

Gambar 53. Laman Data Folder untuk Dataset Breast Cancer

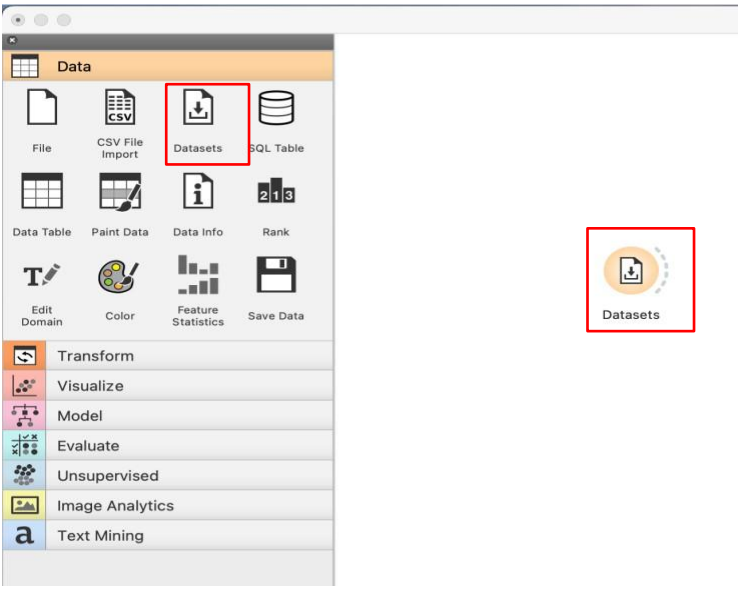
2. Set-up Data di Orange

Orange Data Mining juga menyediakan dataset bawaan. Untuk mengaksesnya, buka aplikasi Orange dan klik tombol **New** untuk memulai proyek baru.

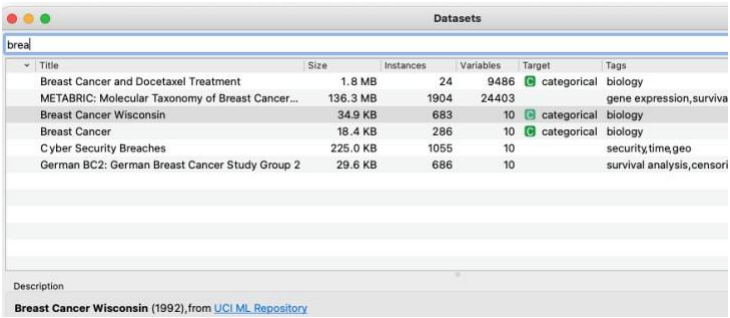


Gambar 54. Tampilan Awal Orange

Seret ikon **Datasets** ke worksheet dan klik dua kali ikon tersebut untuk menampilkan daftar dataset yang tersedia.

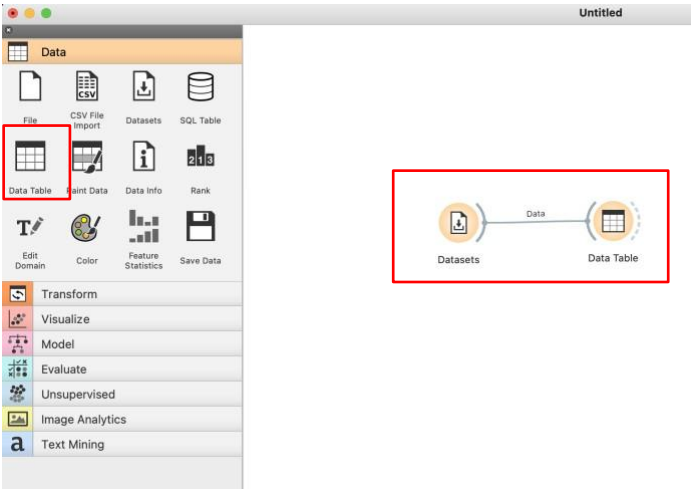


Gambar 55. Drag Ikon Dataset ke Worksheet



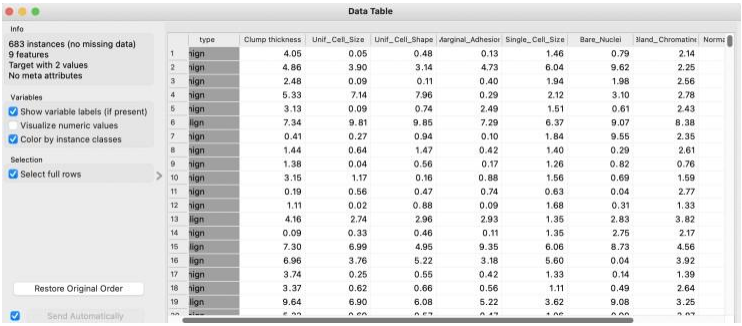
Gambar 56. Daftar Dataset yang Disediakan oleh Orange

Pilih **Breast Cancer (Wisconsin)** lalu hubungkan ikon tersebut dengan ikon **Data Table** untuk melihat isi datanya.



Gambar 57. Koneksi Dataset dan Data Table

Klik dua kali pada ikon Data Table untuk menampilkan data: sebanyak 683 instance, 9 fitur, dan dua kelas target (*malignant* dan *benign*). Dataset ini tidak memiliki missing value, sehingga tidak perlu dilakukan pembersihan data.



Gambar 58. Isi Data Breast Cancer pada Orange

B. Pemahaman Data

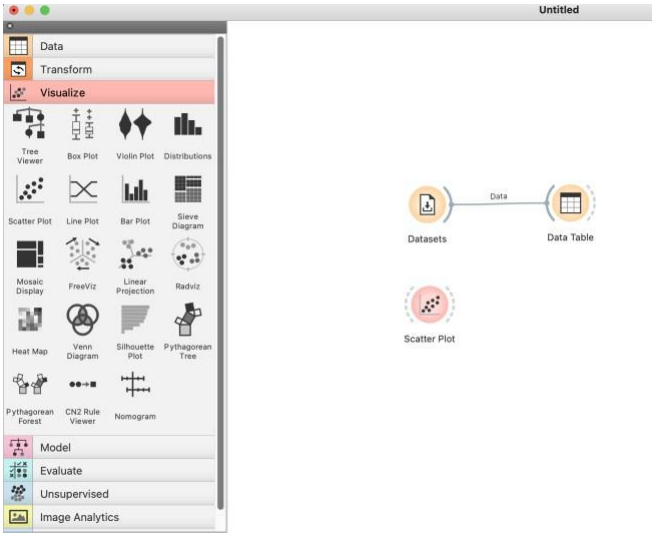
Sebelum melangkah ke pembangunan model, pemahaman terhadap karakteristik data sangatlah penting. Proses ini mencakup identifikasi pola, korelasi antar fitur, serta distribusi nilai atribut yang bisa memengaruhi hasil klasifikasi.

1. Visualisasi Data dengan Scatter Chart

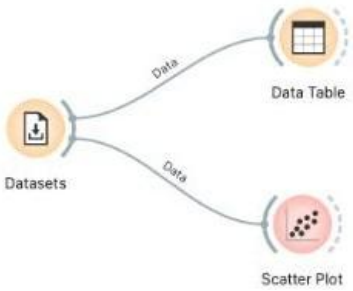
Scatter chart membantu dalam memahami relasi visual antara dua atribut. Misalnya, *Bare Nuclei* dan *Cell Shape* seringkali menunjukkan kecenderungan terhadap keganasan.

Scatter Plot

Tarik ikon **Scatter Plot** ke worksheet, lalu sambungkan ke dataset.

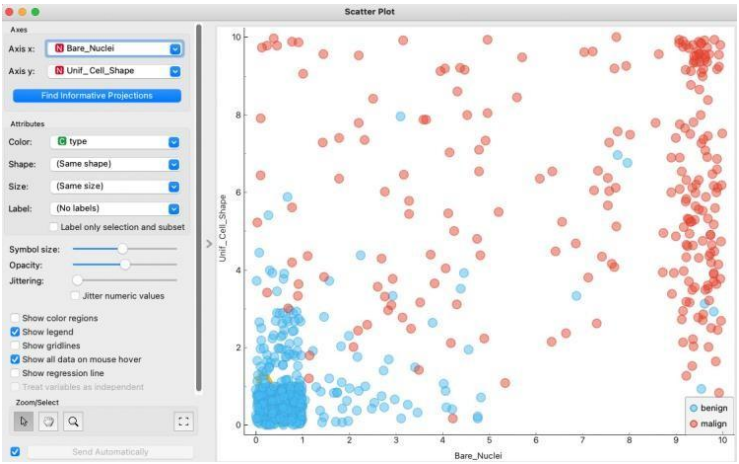


Gambar 59. Scatter Plot untuk Data Kanker



Gambar 60. Koneksi Dataset dengan Scatter Chart

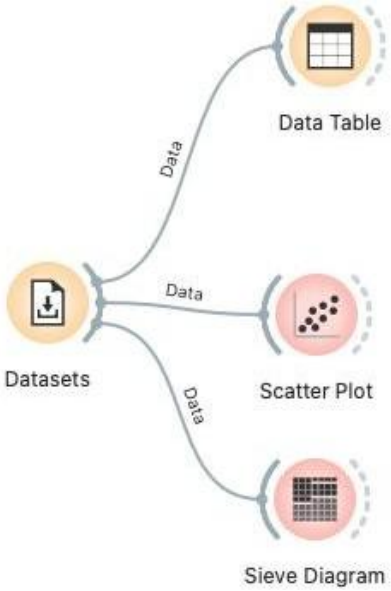
Klik dua kali ikon scatter plot, lalu pilih dua fitur untuk sumbu X dan Y. Misalnya, *Bare Nuclei* di X-axis dan *Cell Shape* di Y-axis. Hasil menunjukkan bahwa nilai-nilai tinggi dari kedua fitur tersebut lebih sering dikaitkan dengan label *malignant*.



Gambar 61. Visualisasi Scatter untuk Prediksi Kanker

2. Visualisasi dengan Sieve Diagram

Sieve diagram berguna untuk melihat keseimbangan data. Ketidakseimbangan kelas dapat menyebabkan bias model. Tarik ikon **Sieve Diagram** ke worksheet dan hubungkan ke dataset.

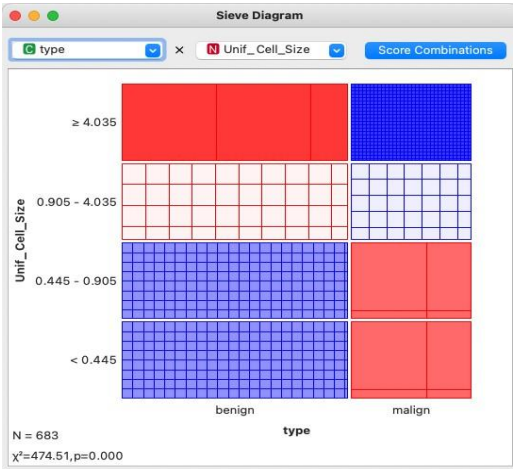


Gambar 62. Sieve Diagram pada Dataset Kanker Payudara

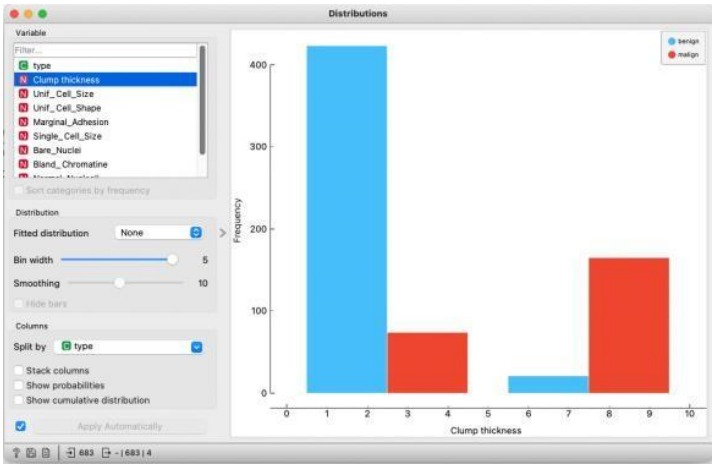
Klik dua kali ikon tersebut. Diagram menunjukkan bahwa jumlah data *benign* lebih banyak dibanding *malignant*. Meski tidak ekstrem, ketimpangan ini perlu diperhatikan saat evaluasi.

3. Distribusi Atribut

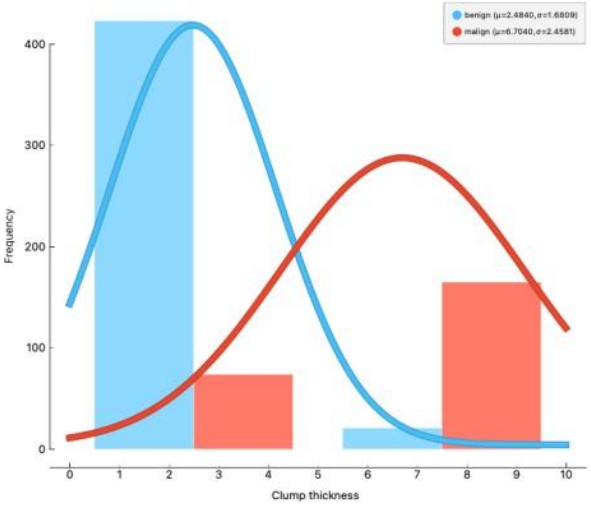
Distribusi data penting untuk memahami apakah data tersebar normal atau condong (skewed). Gunakan ikon **Distribution**, sambungkan ke dataset dan klik dua kali untuk menampilkan visualisasi.



Gambar 63. Distribusi Dataset Kanker



Gambar 64. Bar Chart Distribusi



Gambar 65. Kemiringan Distribusi per Atribut

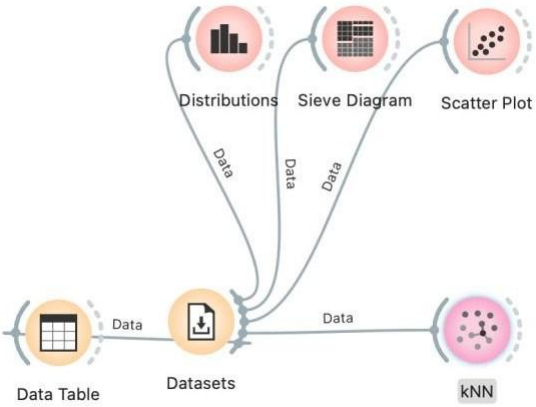
Dengan memilih atribut tertentu seperti *Clump Thickness* atau *Cell Size*, kita bisa menilai apakah fitur tersebut memiliki distribusi simetris atau tidak, dan mempertimbangkan apakah perlu normalisasi.

C. Pembuatan Model

Tahapan berikutnya adalah membangun model klasifikasi. Algoritma yang digunakan di sini adalah **K-Nearest Neighbor (KNN)**, karena kesederhanaannya dan efektivitasnya dalam kasus dataset kecil hingga menengah.

Tarik ikon **KNN** ke worksheet dari menu **Model**, lalu sambungkan ke dataset.

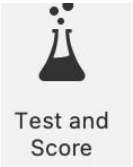
Gambar 66. Pemilihan Model KNN



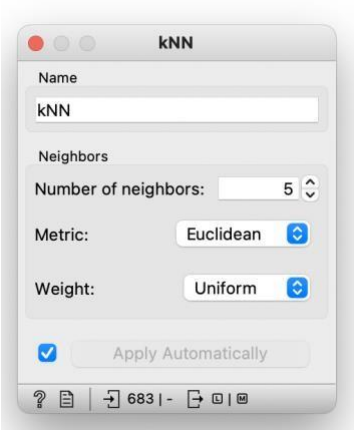
Gambar 67. Koneksi Dataset dan KNN

Klik dua kali ikon KNN. Anda bisa mengatur parameter seperti jumlah tetangga (*n_neighbors*), metode pembobotan (*uniform* atau *distance*), dan metrik jarak.

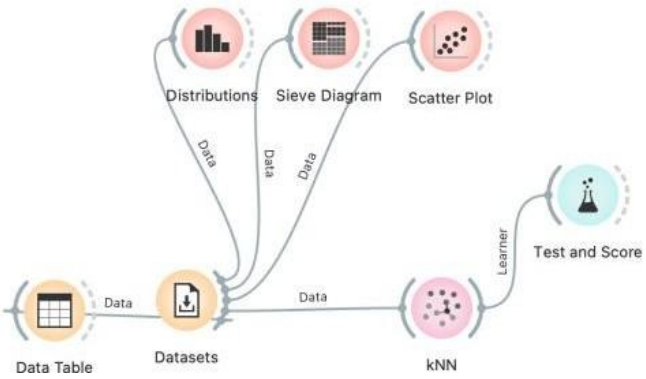
D. Evaluasi



Evaluasi digunakan untuk mengetahui seberapa baik model mengenali pola data. Gunakan ikon **Test and Score**, lalu hubungkan ke KNN model dan dataset.

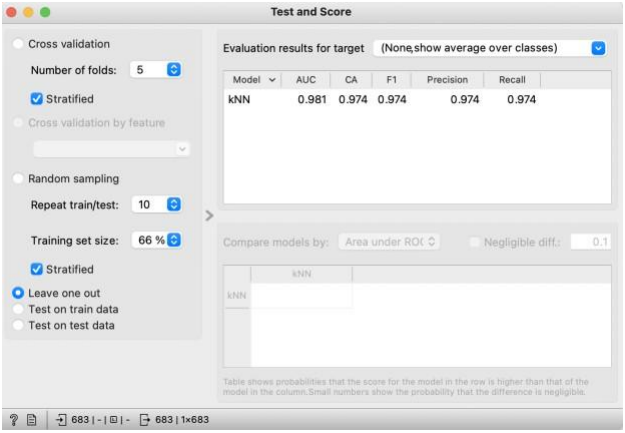


Gambar 68. Test and Score KNN



Gambar 69. Hubungan Dataset dan Evaluasi

Klik dua kali ikon evaluasi, dan hasil evaluasi seperti Accuracy, Precision, Recall, dan F1 Score akan ditampilkan.

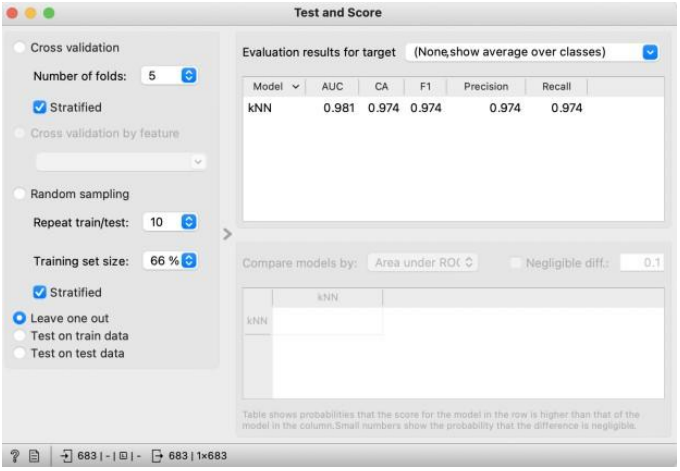


Gambar 70. Hasil Evaluasi Model KNN

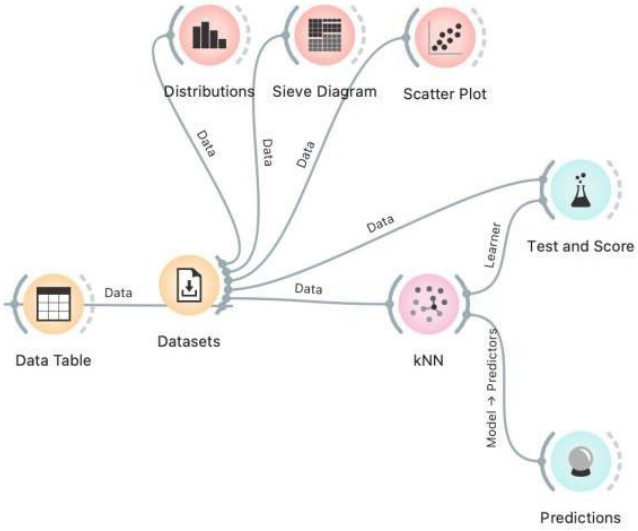
Nilai akurasi di atas 95% menunjukkan bahwa model cukup akurat. Namun, jangan hanya terpaku pada akurasi; perhatikan juga F1 Score, terutama jika terdapat ketimpangan kelas.

E. Prediksi

Langkah terakhir adalah melakukan prediksi terhadap data baru. Tarik ikon **Predictions**, lalu hubungkan ke KNN model dan dataset.

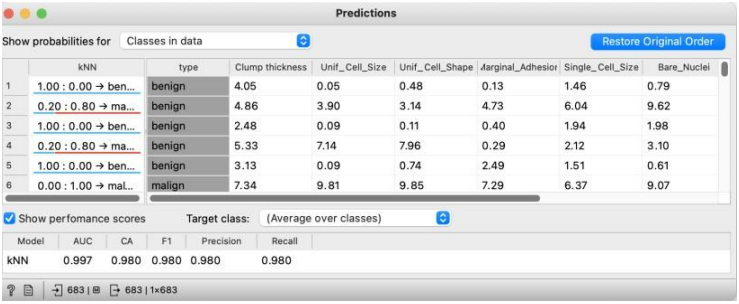


Gambar 71. Prediksi Model KNN



Gambar 72. Koneksi Dataset dan Predictions

Klik dua kali ikon Predictions untuk melihat hasil klasifikasi. Anda dapat membandingkan hasil prediksi dengan label sebenarnya pada setiap instance.



Gambar 73. Hasil Prediksi Visual

Hasil ini memungkinkan kita menilai seberapa tepat model memetakan instance baru, serta mengidentifikasi kemungkinan kesalahan klasifikasi yang harus diantisipasi dalam dunia nyata.

Kesimpulan

Melalui bab ini, mahasiswa tidak hanya dilatih secara teknis untuk melakukan klasifikasi menggunakan Orange, tetapi juga diajak memahami implikasi sosial dan etis dari penggunaan data medis. Dengan menggabungkan pemahaman konseptual, keterampilan teknis, dan refleksi kritis, pembelajaran klasifikasi menjadi alat

transformasional dalam pendidikan data science yang humanistik dan inklusif.

Bab 6
Klasifikasi Diabetes

Penyakit diabetes merupakan salah satu tantangan kesehatan global yang terus meningkat setiap tahunnya. *World Health Organization* (WHO) mencatat bahwa diabetes adalah penyebab utama kematian dini akibat komplikasi seperti gagal ginjal, kebutaan, dan penyakit jantung. Oleh karena itu, pendekatan deteksi dini berbasis data menjadi sangat penting untuk mendukung sistem kesehatan berbasis pencegahan. Salah satu cara untuk mewujudkannya adalah dengan menggunakan teknik klasifikasi dalam ilmu data untuk memprediksi risiko diabetes berdasarkan hasil pemeriksaan kesehatan rutin.

Pada bab ini, kita akan membahas secara mendalam bagaimana membangun model klasifikasi menggunakan dataset medis untuk mendeteksi potensi diabetes. Langkah-langkahnya mencakup akuisisi data dari sumber daring, eksplorasi data secara visual dan statistik, pelatihan model klasifikasi menggunakan algoritma *K*-

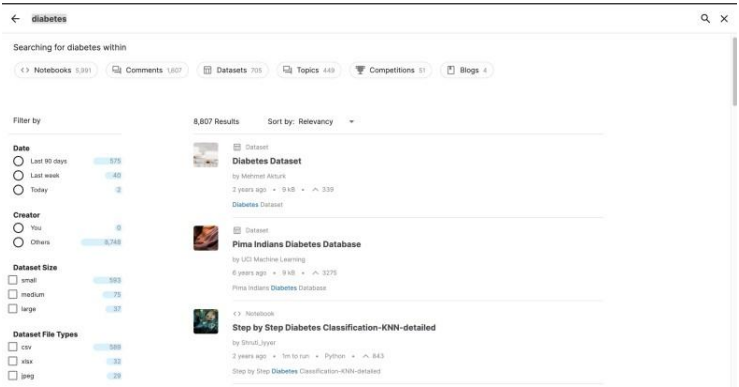
Nearest Neighbor (KNN), evaluasi model dengan metrik performa standar, hingga proses prediksi. Semua proses ini dilakukan secara intuitif melalui aplikasi Orange Data Mining, sebuah alat visualisasi berbasis antarmuka grafis yang sangat ideal bagi pemula maupun praktisi tanpa latar belakang pemrograman.

A. Akuisisi Data

Akuisisi data merupakan pondasi awal dalam proses analisis data. Tanpa data yang tepat dan representatif, seluruh proses analitik dapat menghasilkan kesimpulan yang keliru atau bahkan menyesatkan. Dalam studi ini, kita menggunakan dataset terkenal yaitu *Pima Indian Diabetes Dataset*, yang merupakan data survei kesehatan dari perempuan keturunan Indian-Pima yang mengandung delapan parameter medis, seperti jumlah kehamilan, kadar glukosa, tekanan darah, dan indeks massa tubuh.

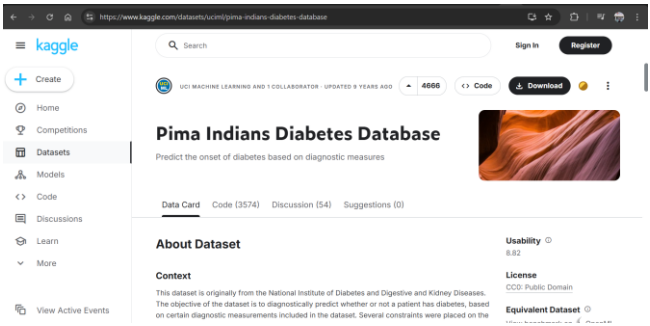
Langkah pertama yang harus dilakukan adalah mengakses situs Kaggle di https://www.kaggle.com/ . Kaggle adalah repositori terbuka yang menyimpan ribuan dataset dari berbagai disiplin ilmu dan sering digunakan

oleh komunitas data science. Pengguna dapat mengetikkan kata kunci "*Pima Indian Diabetes*" di kolom pencarian untuk menemukan dataset tersebut.



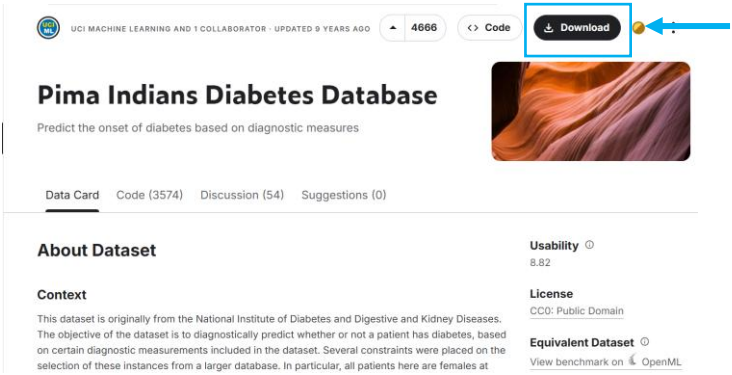
Gambar 74. Laman hasil pencarian pada Kaggle

Dari hasil pencarian tersebut, pengguna memilih salah satu tautan yang mengarah ke dataset yang telah terverifikasi dan digunakan oleh banyak pengguna. Hal ini penting agar data yang digunakan memiliki integritas yang tinggi. Setiap dataset di Kaggle biasanya dilengkapi dokumentasi, deskripsi variabel, dan petunjuk penggunaannya.



Gambar 75. Laman dataset Pima Indian Diabetes Dataset pada Kaggle Repository

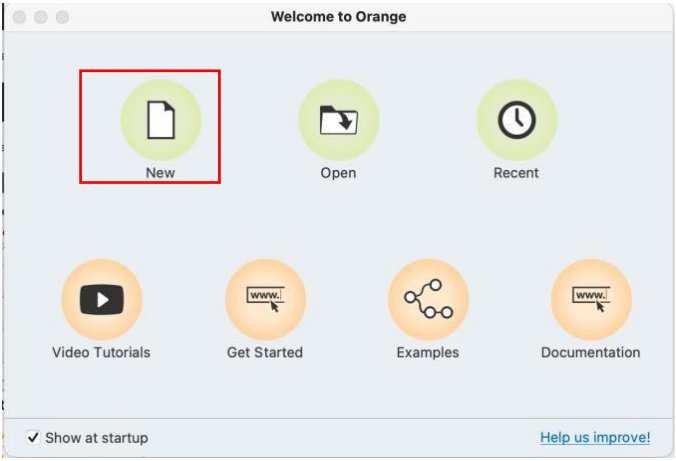
Setelah mengakses halaman tersebut, pengguna perlu login ke akun Kaggle mereka agar tombol unduh dataset menjadi aktif. Dataset ini biasanya tersedia dalam format .csv yang mudah digunakan dalam berbagai aplikasi analisis data, termasuk Orange.



Gambar 76. Unduh dataset Pima Indian Diabetes

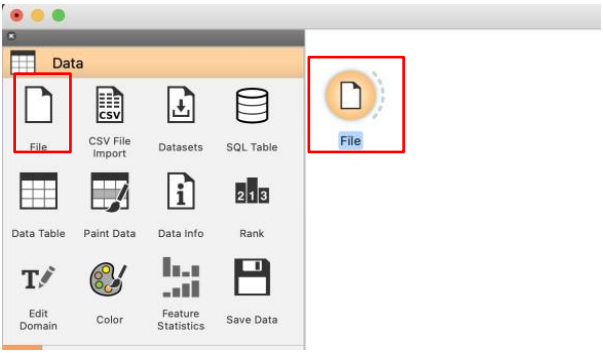
1. Set-up Dataset Eksternal di Orange

Setelah dataset berhasil diunduh, pengguna dapat langsung mengimpor data tersebut ke dalam Orange. Tahapan ini disebut set-up data, di mana kita mempersiapkan dataset agar bisa digunakan dalam proses machine learning. Di Orange, proses ini tidak membutuhkan kode sama sekali, cukup dengan menambahkan ikon File dan memilih dataset .csv dari komputer.



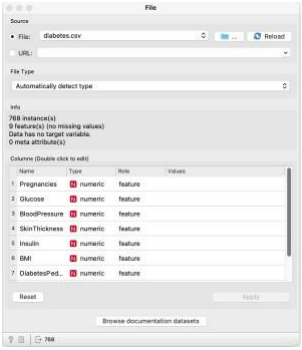
Gambar 77. Tampilan depan dari Aplikasi Orange untuk proyek klasifikasi diabetes

Widget File kemudian di-drag ke area kerja Orange, dan pengguna melakukan klik dua kali untuk memilih file yang diinginkan.



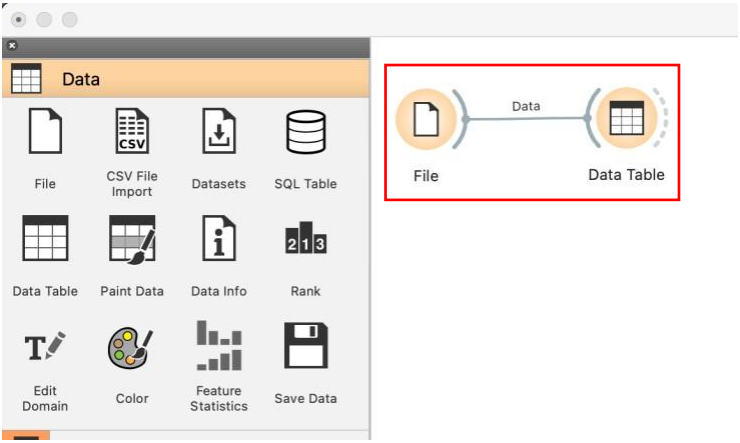
Gambar 78. Drag ikon file pada worksheet

Langkah ini dilanjutkan dengan memilih file dari komputer yang berisi data diabetes. Pastikan file telah disimpan di lokasi yang mudah ditemukan.



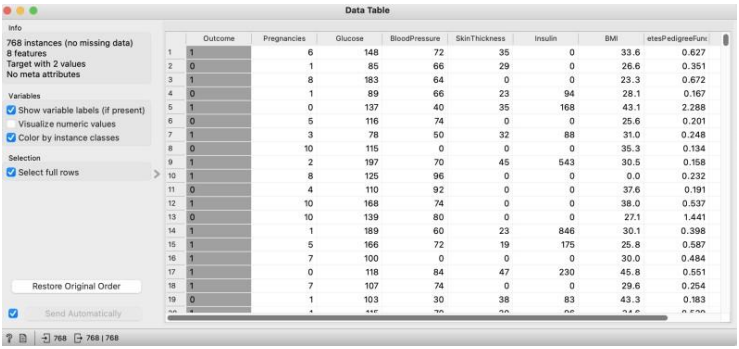
Gambar 79. Pemilihan file untuk dataset diabetes

Setelah file terpilih, kita perlu menambahkan widget Data Table untuk melihat isi data secara keseluruhan. Ini adalah proses verifikasi awal untuk memastikan bahwa semua data berhasil diimpor dan tidak ada atribut yang kosong atau tidak terbaca.



Gambar 80. Pembuatan link pada dataset dan Data Table

Dengan membuka Data Table, kita bisa melihat jumlah instance, jumlah fitur, dan target variabel yang digunakan. Dalam hal ini, data memiliki 768 baris dan 8 fitur dengan 1 kolom target (Outcome) yang menunjukkan apakah seseorang terkena diabetes (1) atau tidak (0).



	Outcome	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	etesPedigreeFunc
1	1	6	148	72	35	0	33.6	0.627
2	0	1	85	66	29	0	26.6	0.351
3	1	8	183	64	0	0	23.3	0.672
4	0	1	89	66	23	94	28.1	0.167
5	1	0	137	40	35	168	43.1	2.288
6	0	5	116	74	0	0	25.6	0.201
7	1	3	78	50	32	88	31.0	0.248
8	0	10	115	0	0	0	35.3	0.134
9	1	2	197	70	45	543	30.5	0.158
10	1	8	125	96	0	0	0.0	0.232
11	0	4	110	92	0	0	37.6	0.191
12	1	10	168	74	0	0	38.0	0.537
13	0	10	139	80	0	0	27.1	1.441
14	1	1	189	60	23	846	30.1	0.398
15	1	5	166	72	19	175	25.8	0.587
16	1	7	100	0	0	0	30.0	0.484
17	1	0	118	84	47	230	45.8	0.551
18	1	7	107	74	0	0	29.6	0.254
19	0	1	103	30	38	83	43.3	0.183

Gambar 81. List data penyakit diabetes dari data yang disediakan oleh Kaggle

B. Pemahaman Data

Pemahaman terhadap struktur dan isi data sangat penting sebelum melakukan pelatihan model. Pemahaman ini bertujuan untuk mengetahui pola hubungan antar fitur, mendeteksi nilai-nilai ekstrem (outlier), serta menilai keseimbangan kelas target.

1. Visualisasi dengan Scatter Chart

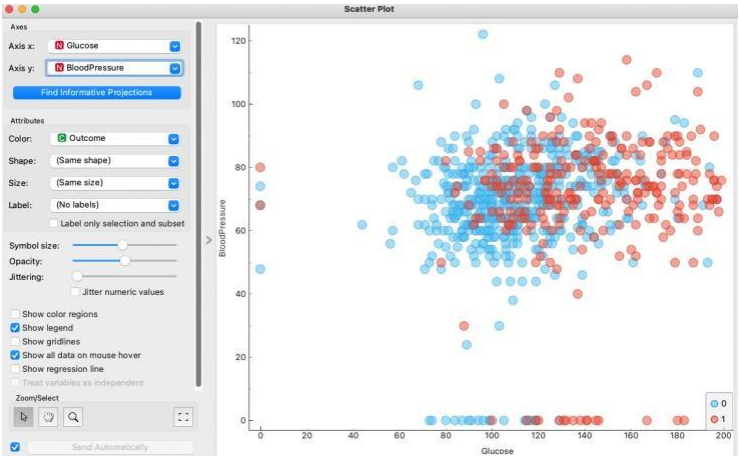
Scatter plot merupakan salah satu alat visual yang sangat berguna untuk mengeksplorasi hubungan dua variabel numerik. Dalam konteks data diabetes, kita bisa menganalisis apakah terdapat hubungan antara kadar glukosa dan tekanan darah.

Gambar 82. Pemilihan scatter chart untuk visualisasi data Pima diabetes

Kita kemudian menghubungkan ikon File ke scatter chart untuk menampilkan data.

Gambar 83. Koneksi antara dataset dan scatter chart untuk visualisasi data diabetes

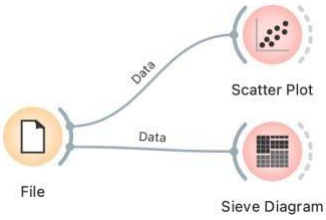
Setelah memilih sumbu X dan Y, kita akan melihat pola distribusi data. Jika data menunjukkan tren yang jelas, maka fitur tersebut memiliki potensi sebagai prediktor dalam model.



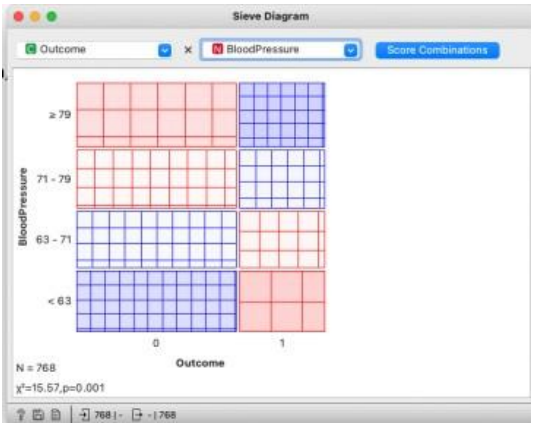
Gambar 84. Scatter plot untuk data Pima diabetes

2. Visualisasi dengan Sieve Diagram

Sieve Diagram membantu melihat apakah kelas target (diabetes dan non-diabetes) memiliki jumlah data yang seimbang.



Ketidakseimbangan kelas dapat memengaruhi kinerja model karena model cenderung bias pada kelas yang dominan.

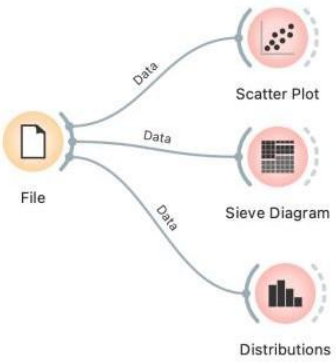


Gambar 85. Visualisasi sieve diagram pada data Pima diabetes

Dari diagram ini kita dapat melihat bahwa data non-diabetes lebih banyak daripada diabetes. Informasi ini penting untuk menentukan strategi evaluasi model ke depan.

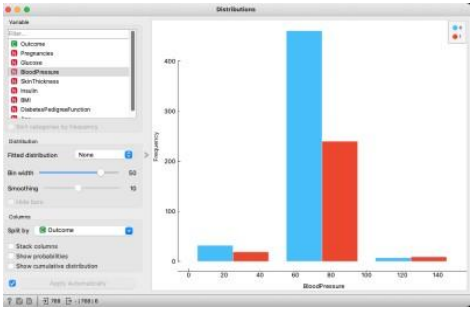
3. Visualisasi Distribusi Atribut

Diagram distribusi menunjukkan bentuk penyebaran nilai pada setiap fitur. Kita bisa mendeteksi apakah data menyebar normal, skewed, atau memiliki outlier.

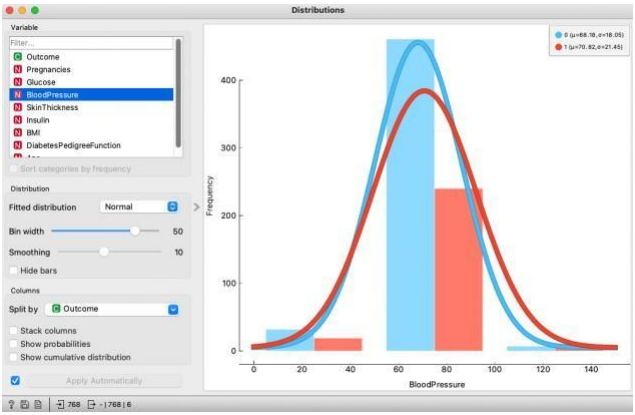


Gambar 86. Koneksi data Pima Indian diabetes dengan distributions

Distribusi ini membantu memahami karakter data numerik seperti Glucose, BMI, atau Age.

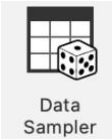


Gambar 87. Distribusi data Pima Indian diabetes

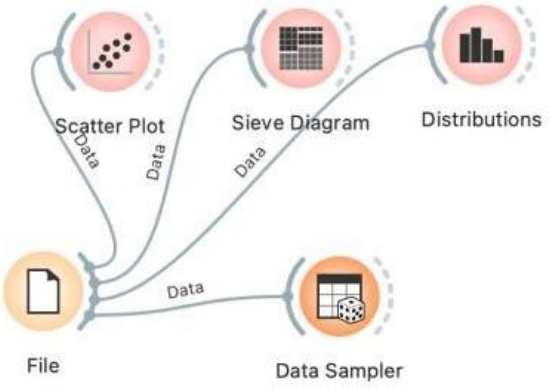


Gambar 88. Kemiringan data penyakit diabetes

C. Pembuatan Model

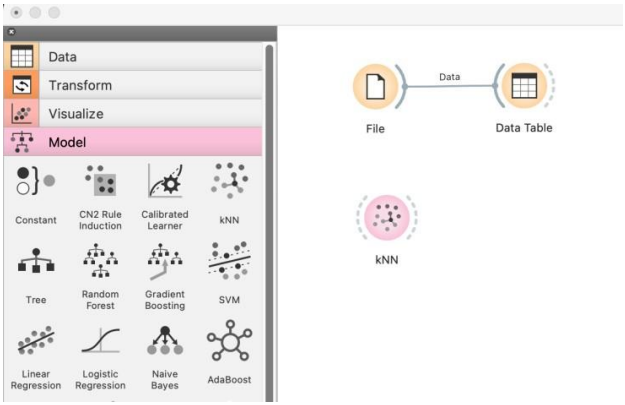


Langkah berikutnya adalah membuat model klasifikasi. Kita menggunakan algoritma K-Nearest Neighbor (KNN) yang bekerja berdasarkan kedekatan jarak antar instance. Kita juga menggunakan Data Sampler untuk membagi data menjadi data latih dan data uji.

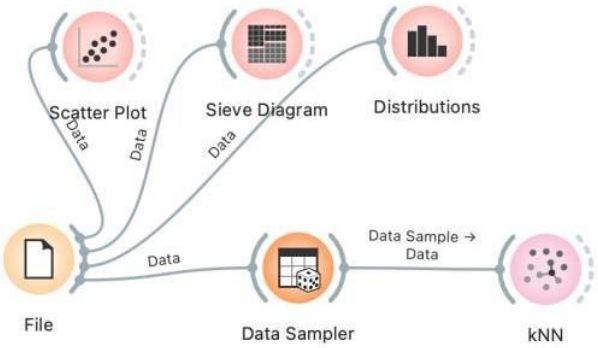


Gambar 89. Koneksi antara Data Sampler dan File

Selanjutnya, ikon KNN ditambahkan dan dikaitkan dengan output dari data sampler.

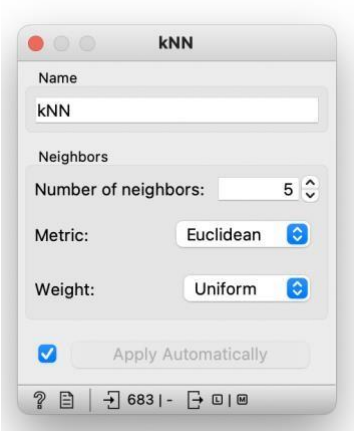


Gambar 90. Pemilihan model KNN untuk klasifikasi diabetes



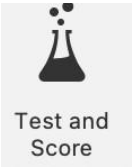
Gambar 91. Koneksi data Pima Indian diabetes dengan KNN model

Kita bisa mengatur nilai k, metrik jarak, dan bobot pembobotan sesuai kebutuhan.

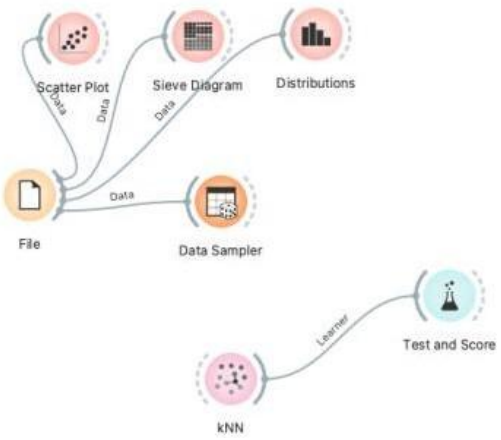


Gambar 92. Setting parameter untuk klasifikasi data diabetes

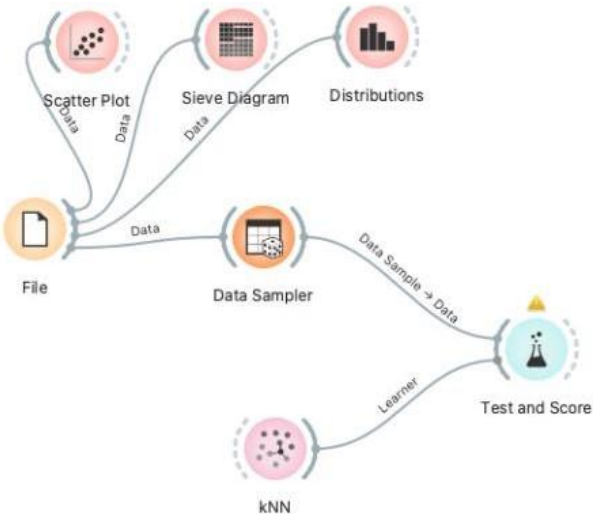
D. Evaluasi



Evaluasi model dilakukan untuk mengukur kinerja model secara objektif. Kita menggunakan Test and Score untuk menghitung metrik seperti akurasi, precision, recall, dan F1-score.

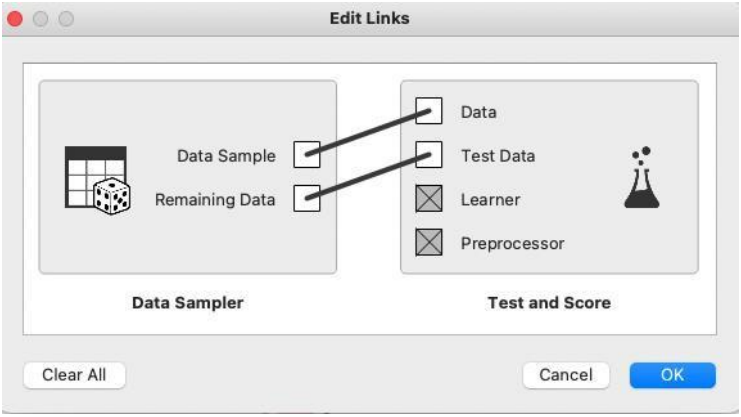


Gambar 93. Hubungan antara prediksi diabetes model KNN dengan Test and Score



Gambar 94. Hubungan dataset diabetes dengan Test and Score

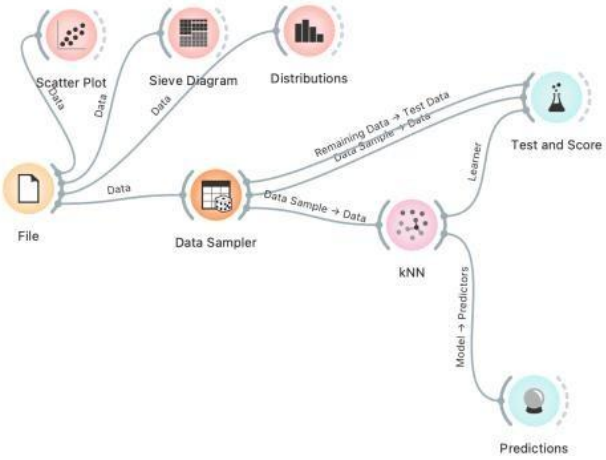
Kita juga bisa mengatur rasio data latih dan data uji.



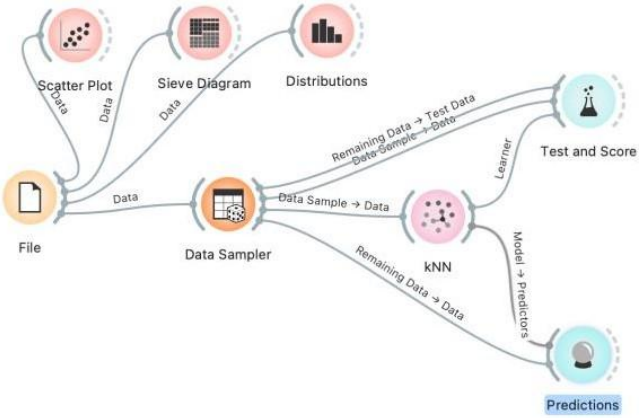
Gambar 95. Pemilihan data training dan testing

E. Prediksi

Langkah terakhir adalah prediksi terhadap data baru. Hasil prediksi membantu kita mengetahui apakah seseorang berada pada kategori berisiko atau tidak.

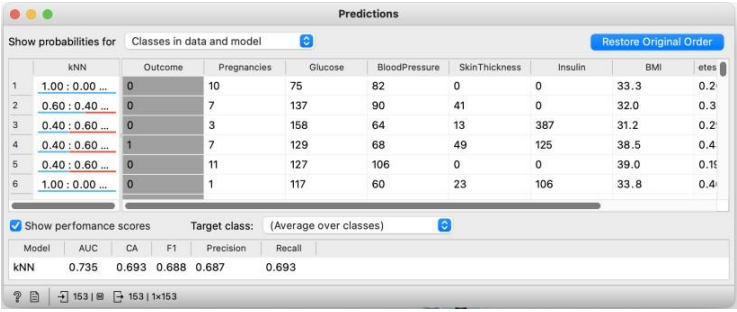


Gambar 96. Link model KNN dan Prediction untuk prediksi data diabetes



Gambar 97. Koneksi antara KNN model dengan Prediction untuk data sampler Pima diabetes

Hasil prediksi ditampilkan secara visual dan dapat dibandingkan dengan label sebenarnya.



Gambar 98. Hasil prediksi diabetes

Bab 7
Latihan Soal Klasifikasi

A. Soal Pilihan Ganda

1. Tujuan utama dari klasifikasi dalam konteks analisis data kematian dunia adalah …
   a. Menentukan korelasi antar fitur numerik
   b. Memprediksi jumlah kasus kematian di masa depan
   c. Mempetakan penyebab kematian ke dalam kategori tertentu
   d. Menjumlahkan data mortalitas dari setiap negara
2. Langkah awal dalam membangun model klasifikasi menggunakan Orange adalah …
   a. Visualisasi
   b. Akuisisi data
   c. Confusion Matrix
   d. ROC Analysis
3. Widget Orange yang digunakan untuk mengevaluasi performa model klasifikasi adalah…
   a. Scatter Plot
   b. Box Plot
   c. Test and Score
   d. Data Sampler
4. Berikut ini adalah metrik evaluasi model klasifikasi, kecuali …
   a. Accuracy
   b. Precision
   c. R2 Score

d. Recall

5. Distribusi data pada Orange Data Mining dapat divisualisasikan dengan menggunakan …
   a. Test and Score
   b. Confusion Matrix
   c. Sieve Diagram
   d. Feature Constructor

B. Studi Kasus: Analisis Klasifikasi Penyebab Kematian Global

1. Latar Belakang:

   Kematian di seluruh dunia terjadi karena berbagai penyebab, mulai dari penyakit kronis seperti kanker dan jantung, hingga faktor lingkungan seperti kecelakaan lalu lintas dan bencana alam. Organisasi Kesehatan Dunia (WHO) dan lembaga-lembaga statistik global telah mengumpulkan data mengenai penyebab kematian di berbagai negara dengan kategori usia, jenis kelamin, wilayah, dan faktor risiko.

   Dengan klasifikasi berbasis machine learning, kita dapat mencoba memetakan pola penyebab kematian berdasarkan indikator demografis dan sosial, sehingga membantu peneliti atau pembuat kebijakan untuk membuat strategi pencegahan yang lebih terarah.

2. Sumber Data:

   Dataset “Cause of Deaths Around the World” tersedia di Kaggle: https://www.kaggle.com/datasets/kashnitsky/deaths-from-various-causes

Langkah-Langkah Analisis dengan Orange Data Mining

1. Akuisisi Data
   - Unduh dataset dari Kaggle (format CSV).
   - Import ke Orange menggunakan widget File. Pastikan semua kolom terbaca dengan benar, khususnya kolom target (misal: "Cause").
2. Eksplorasi Awal dan Pembersihan Data
   - Gunakan Feature Statistics dan Distributions untuk mengevaluasi nilai-nilai ekstrem, missing values, dan distribusi kelas.
3. Visualisasi Data
   - Gunakan Scatter Plot untuk melihat hubungan antara misalnya “age_group”, “region”, dan “deaths_per_100k”.
   - Gunakan Sieve Diagram untuk melihat keseimbangan kelas berdasarkan jenis kelamin atau wilayah.
4. Pemilihan Target dan Fitur
   - Atur “Cause” (penyebab kematian) sebagai target. Gunakan fitur seperti

"age", "sex", "region", dan "year" sebagai prediktor.

5. Pembuatan Model Klasifikasi
    - Gunakan 2–3 algoritma klasifikasi seperti Naive Bayes, Decision Tree, dan Random Forest.
    - Hubungkan semua model ke widget Test and Score untuk evaluasi performa model.
6. Evaluasi Model
    - Perhatikan metrik Accuracy, Precision, Recall, dan F1-Score. Bandingkan model mana yang lebih akurat dalam mengklasifikasikan penyebab kematian.
    - Tambahkan Confusion Matrix untuk melihat di mana model paling banyak melakukan kesalahan klasifikasi.
7. Prediksi Data Baru
    - Gunakan Predictions untuk mencoba memetakan data individu (misalnya: usia 60 tahun, laki-laki, tinggal di Asia Selatan) dan lihat prediksi kemungkinan penyebab kematiannya menurut model.

Daftar Pustaka

1. Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.
2. Alpaydin, E. (2020). *Introduction to Machine Learning*. MIT Press.
3. Amrin, A., & Pahlevi, O. (2022). Implementation of Logistic Regression Classification Algorithm and Support Vector Machine for Credit Eligibility Prediction. *Jurnal Informatics Telecommunication Engineering*, 5(2), 433–441. https://doi.org/10.31289/jite.v5i2.6220
4. Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12), 64–73.
5. Geron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.
6. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Elsevier.

7. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer.
8. Kelleher, J. D., Mac Carthy, M., & Korvir, B. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. MIT Press.
9. Khan, W., Zaki, N., Masud, M. M., Ahmad, A., Ali, L., Ali, N., & Ahmed, L. A. (2022). Infant birth weight estimation and low birth weight classification in United Arab Emirates using machine learning algorithms. *Scientific Reports*, 12(1), 1–12.
10. Mateo, F., Tarazona, A., & Mateo, E. M. (2021). Comparative study of several machine learning algorithms for classification of unifloral honeys. *Foods*, 10(7), 1543.
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine

learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

12. Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*. O'Reilly Media.
13. Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.
14. Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.
15. Sinha, A. (n.d.). Utilization of Machine Learning Models in Real Estate House Price Prediction. *Amity Journal of Computational Sciences (AJCS)*, 4(1), 18–23. Retrieved from www.amity.edu/ajcs
16. Tan, P. N., Steinbach, M., Karpatne, A., & Kumar, V. (2018). *Introduction to Data Mining*

(2nd ed.). Pearson.

17. Wiradinata, T., Tanamal, R., Saputri, T. R., & Soekamto, Y. S. (2021, September). An Implementation of Support Vector Machine Classification for Developer Academy Acceptance Prediction Model. In *2021 2nd International Conference on Innovative and Creative Information Technology (ICITech)* (pp. 110–116). IEEE.
18. Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a Standard Process Model for Data Mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (Vol. 1, pp. 29–39).
19. Yang, L., Chen, Y., Xu, N., Zhao, R., Chau, K. W., & Hong, S. (2020). Place-varying impacts of urban rail transit on property prices in Shenzhen, China: Insights for value capture. *Sustainable Cities and Society*, 58, 102140.

https://doi.org/10.1016/j.scs.2020.102140

20. Zhou, Z. H. (2021). *Machine Learning*. Springer Nature.

SMART ANALYTICS

Panduan Visual Regresi dan Klasifikasi dengan Orange Data Mining di Era Data Digital

Buku ini merupakan panduan praktikum yang dirancang secara visual dan aplikatif untuk membantu mahasiswa memahami konsep regresi dan klasifikasi menggunakan Orange Data Mining, sebuah perangkat lunak open source yang intuitif. Di tengah kemajuan teknologi digital dan kebutuhan pengambilan keputusan berbasis data, kemampuan analisis menjadi kompetensi penting bagi akademisi dan praktisi. Dengan pendekatan langkah demi langkah dan studi kasus sosial-politik yang kontekstual, buku ini tidak hanya mengajarkan teknik analisis data, tetapi juga menanamkan pemahaman mendalam tentang penggunaannya dalam menjawab persoalan kebijakan publik. Disusun oleh dosen Fakultas Ilmu Sosial dan Ilmu Politik Universitas Muhammadiyah Palangka Raya, buku ini merupakan kolaborasi antara keilmuan sosial dan kecakapan teknologi, yang diharapkan mampu meningkatkan literasi data serta mendukung pembentukan kebijakan yang lebih responsif di era digital.

Dr. Irwani, S.Sos., M.A.P. adalah dosen dan peneliti di bidang Administrasi Publik dan Sosiologi. Ia aktif mengembangkan kajian tentang kebijakan publik, kepemimpinan digital, serta pemberdayaan masyarakat berbasis data. Kontribusinya pada buku ini mencerminkan pemahamannya yang mendalam terhadap isu-isu transformasi digital di sektor publik dan pentingnya pendekatan analitik dalam mendukung tata kelola pemerintahan yang lebih adaptif dan berbasis bukti.

Muhammad Anzarach Pratama, S.A.N., M.P.A. adalah akademisi dan praktisi di bidang administrasi publik dan manajemen sektor publik. Ia memiliki minat dalam pengembangan tata kelola pemerintahan, transformasi digital, dan kebijakan berbasis data. Kontribusinya pada buku ini memperkuat pemahaman konseptual dan praktis mengenai analisis data sebagai landasan penting dalam proses perumusan dan evaluasi kebijakan publik yang responsif dan efektif di era digital.

Ir. Doddy Teguh Yuwono, M.Kom. adalah dosen dan peneliti yang aktif di bidang Teknologi Informasi. Ia memiliki minat pada analisis data, literasi digital, dan pemanfaatan teknologi untuk pembelajaran dan pengambilan keputusan. Pada buku ini, ia turut menyusun panduan praktis penggunaan Orange Data Mining untuk membantu pembaca memahami regresi dan klasifikasi data secara visual dan aplikatif.