**TUGAS BESAR II**

**ANALISIS TAHAP PRE PROCESSING ALGORITMA CART, NEURAL NETWORK, DAN STOCHASTIC GRADIENT DESCENT**

Mata Kuliah Pengantar Data Science



**Dosen :**

Wawan Gunawan, S.Kom, MT

**Disusun Oleh :**

Muhammad Fauzan Zamzami (41522010080)

Vemas Adi Pratama (41522010081)

Okky Tri Wahyuda (41522010114)

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS MERCU BUANA**

**2023**

# Pendahuluan

Diabetes retinopathy adalah salah satu komplikasi serius yang dapat muncul pada individu yang menderita diabetes. Kondisi ini terjadi ketika gula darah tinggi merusak pembuluh darah di mata, khususnya di retina, yang dapat mengakibatkan kehilangan penglihatan. Untuk mencegah dampak yang parah, diagnosa yang akurat dan dini menjadi kunci. Oleh karena itu, penggunaan teknik-teknik seperti algoritma CART (Classification and Regression Trees), jaringan saraf tiruan (neural network), dan SGD (Stochastic Gradient Descent) dalam klasifikasi penyakit diabetes retinopathy menjadi sangat penting.

Tujuan utama dari penelitian ini adalah untuk menguji, membandingkan, dan menganalisis ketiga algoritma, yaitu CART, neural network, dan SGD, dalam konteks klasifikasi penyakit diabetes retinopathy. Dengan demikian, penelitian ini akan memberikan wawasan yang lebih mendalam tentang manfaat dan keterbatasan masing-masing algoritma dalam mendiagnosa penyakit ini.

Penelitian ini akan dibatasi oleh penggunaan dataset dari UCI Machine Learning Repository sebagai sumber data utama. Penelitian ini sangat relevan karena diabetes retinopathy adalah masalah kesehatan global yang memengaruhi jutaan orang di seluruh dunia. Diagnosa yang tepat dan dini merupakan langkah kritis dalam mencegah hilangnya penglihatan pada penderita diabetes. Dengan membandingkan dan menganalisis berbagai algoritma dalam konteks ini, penelitian ini memiliki potensi untuk meningkatkan kualitas perawatan medis dan membantu dokter dalam membuat keputusan yang lebih baik. Selain itu, penelitian ini juga memiliki implikasi dalam pengembangan teknologi medis yang lebih baik.

Makalah ini akan disusun dalam beberapa bagian utama. Pertama, kami akan menjelaskan tahapan penelitian dan metodologi yang digunakan. Kemudian, kami akan memaparkan hasil penelitian, termasuk perbandingan kinerja ketiga algoritma.

# Daftar isi

[Pendahuluan 2](#_Toc153279778)

[Daftar isi 3](#_Toc153279779)

[I. Pengolahan Data 4](#_Toc153279780)

[a) Pengumpulan Data dan Tahap Pre Processing 4](#_Toc153279781)

[b) Langkah-langkah Normalisasi Data 6](#_Toc153279782)

[II. Analisis Perilaku Data terhadap Kelas dengan Visualisasi 14](#_Toc153279783)

[III. Perbandingan Algoritma 19](#_Toc153279784)

[CART 19](#_Toc153279785)

[Neural Network 19](#_Toc153279786)

[SGD 19](#_Toc153279787)

[IV. Penerapan 21](#_Toc153279788)

[V. Referensi 21](#_Toc153279789)

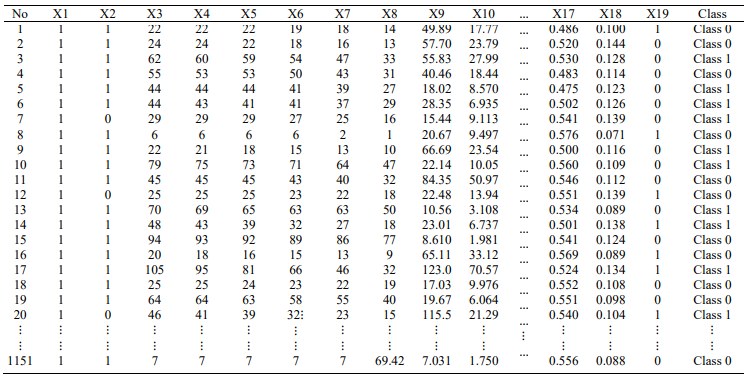
# Pengolahan Data

## Pengumpulan Data dan Tahap Pre Processing

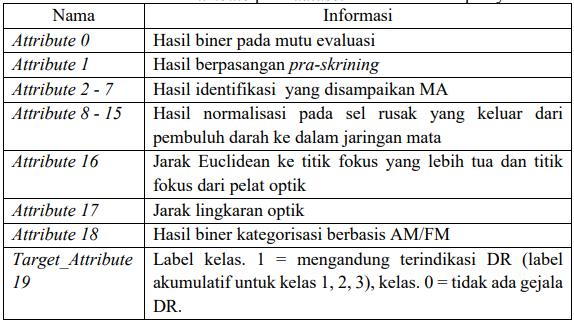
Pada tahap ini, data mentah diolah untuk menghasilkan data yang bersih, termasuk pengisian atau penghapusan nilai kosong, penyamarataan format pada setiap atribut, dan penggantian angka yang mungkin typo.

Dataset yang digunakan berasal dari UCI Machine Learning Repository dan merupakan diabetic retinopathy debrecen.

Dataset ini terdiri dari 1151 data dengan 19 atribut dan 1 kelas target. Berikut adalah rinciannya:

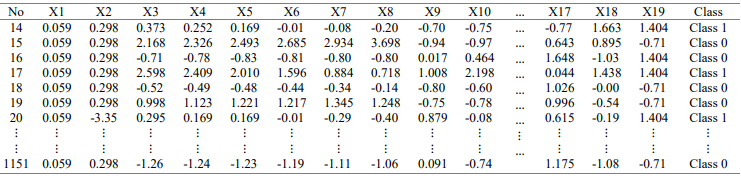


Gambar 1.2 Dataset Mentah



Gambar 1.3 Atribut dan Label Dataset

Setelahnya, data tersebut menjalani proses preprocessing, di mana perubahan yang dilakukan adalah mengubah nilai "contains signs" sebelumnya menjadi 0, dan "signs" menjadi 1.

z

Gambar 1.4 Data Normalisasi

Normalisasi dilakukan dengan bantuan aplikasi RapidMiner Studio. Author meng-klaim dataset dinormalisasi dengan normalisasi Min-Max dengan nilai minimum 0 dan nilai maksimum 1. Hal ini akan menghasilkan data pada rentang 0 dan 1. Berikut formula normalisasi Min-Max :



Gambar 1.5 Formula Normalisasi Min-Max

Namun, berdasarkan data normalisasi dalam jurnal, ditemukan adanya keanehan data, yang mana terdapat nilai data diluar rentang 0 dan 1. Setelah kami melakukan pengecekan kembali, kami menemukan bahwa data yang telah dinormalisasi pada jurnal, tidak dinormalisasi dengan normalisasi Min-Max, melainkan dengan metode normalisasi standar (Standardization). Normalisasi standar dilakukan dengan mengubah nilai nilai dataset sehingga memiliki rata-rata nol dan standar deviasi sebanyak satu. Berikut formula untuk normalisasi Standardization :



*Xstandardized* : Nilai yang telah dinormalisasi standar

*X* : Nilai asli

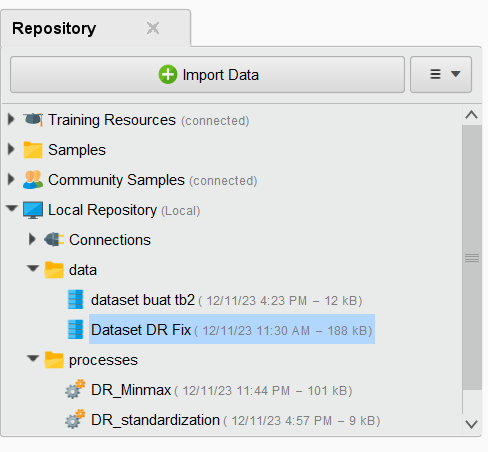
*X* : Rata-rata nilai data

*σ* : Standar deviasi dataset

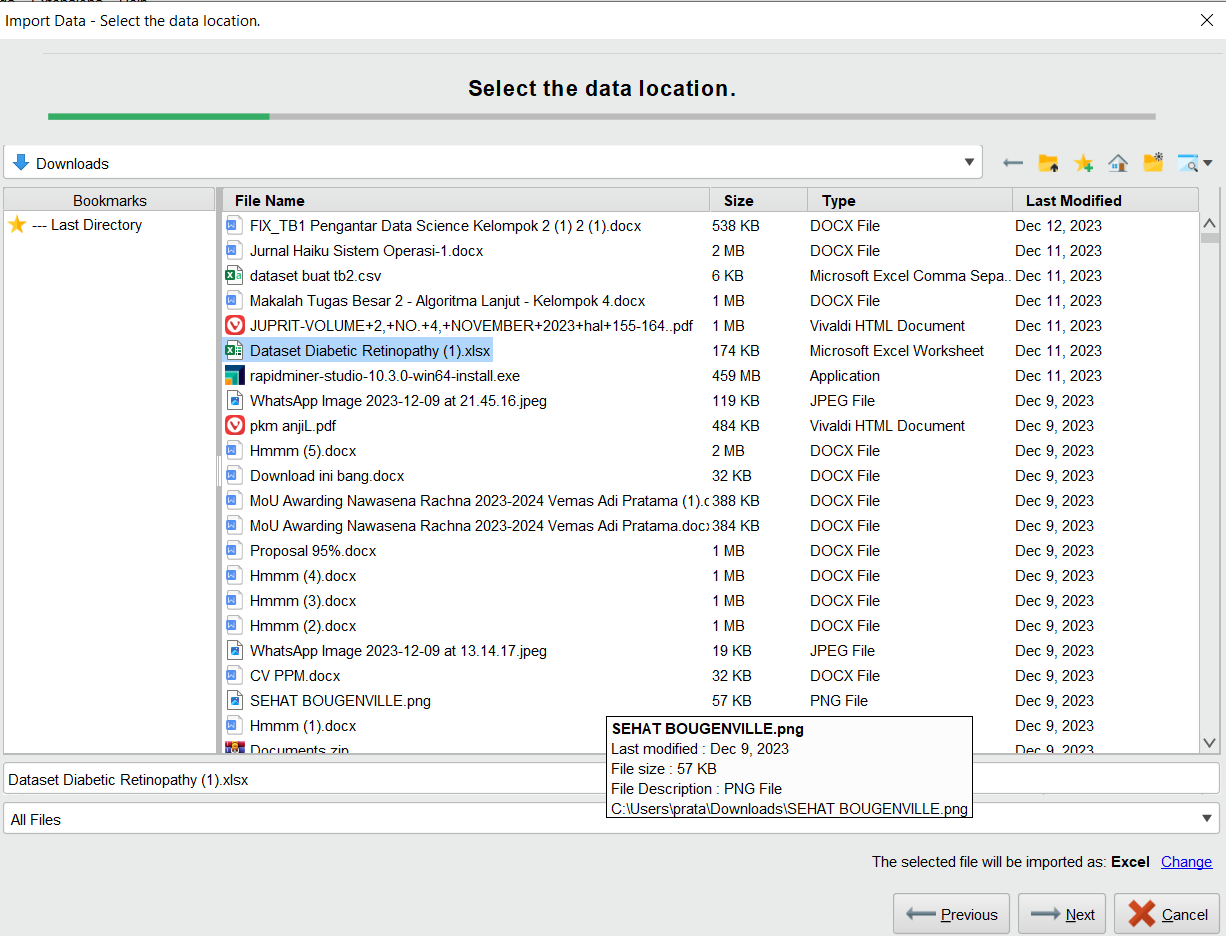
## Langkah-langkah Normalisasi Data

Berikut Langkah-Langkah Normalisasi Data :

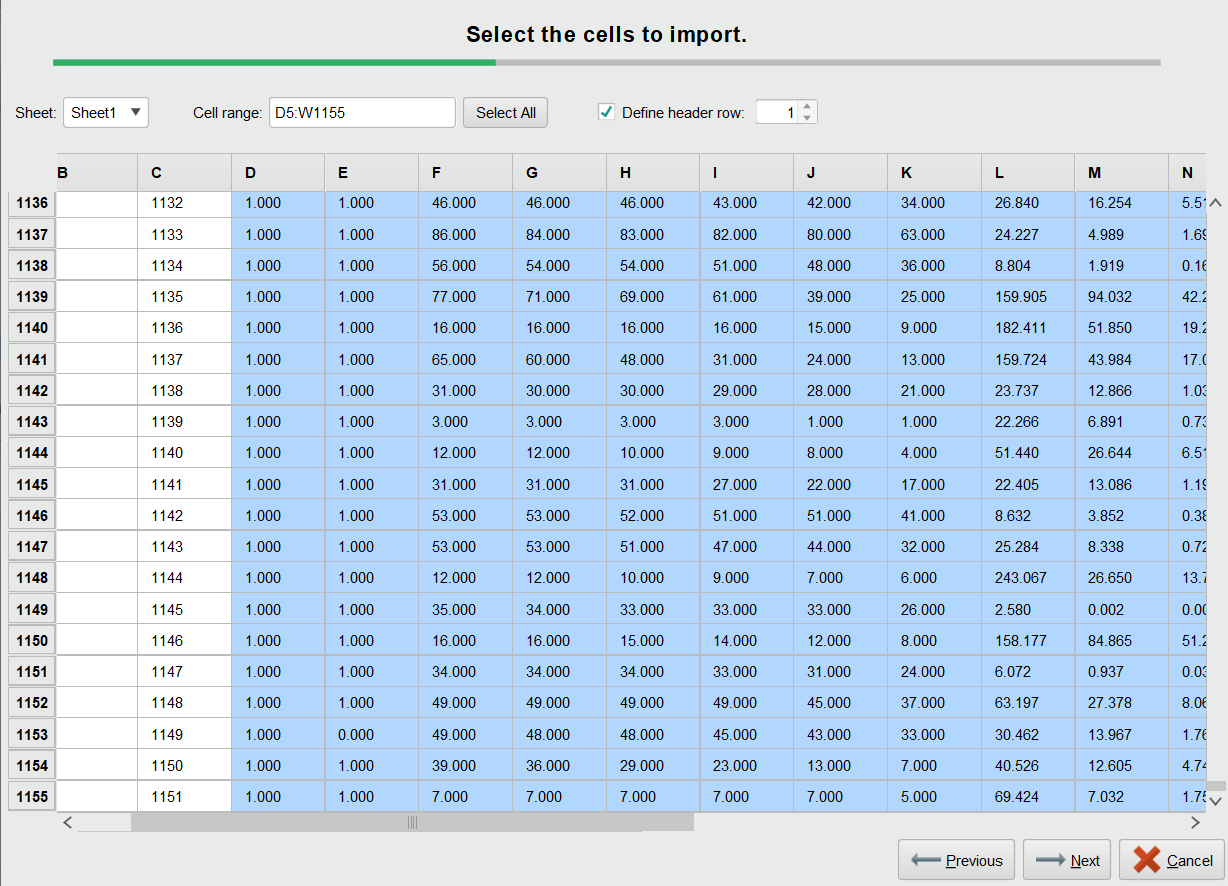
1. Siapkan Dataset, bisa dalam bentuk excel atau csv untuk memudahkan pengimportan. Setelah disiapkan masuk pada aplikasi RapidMiner Studio lalu import data tersebut pada aplikasi.



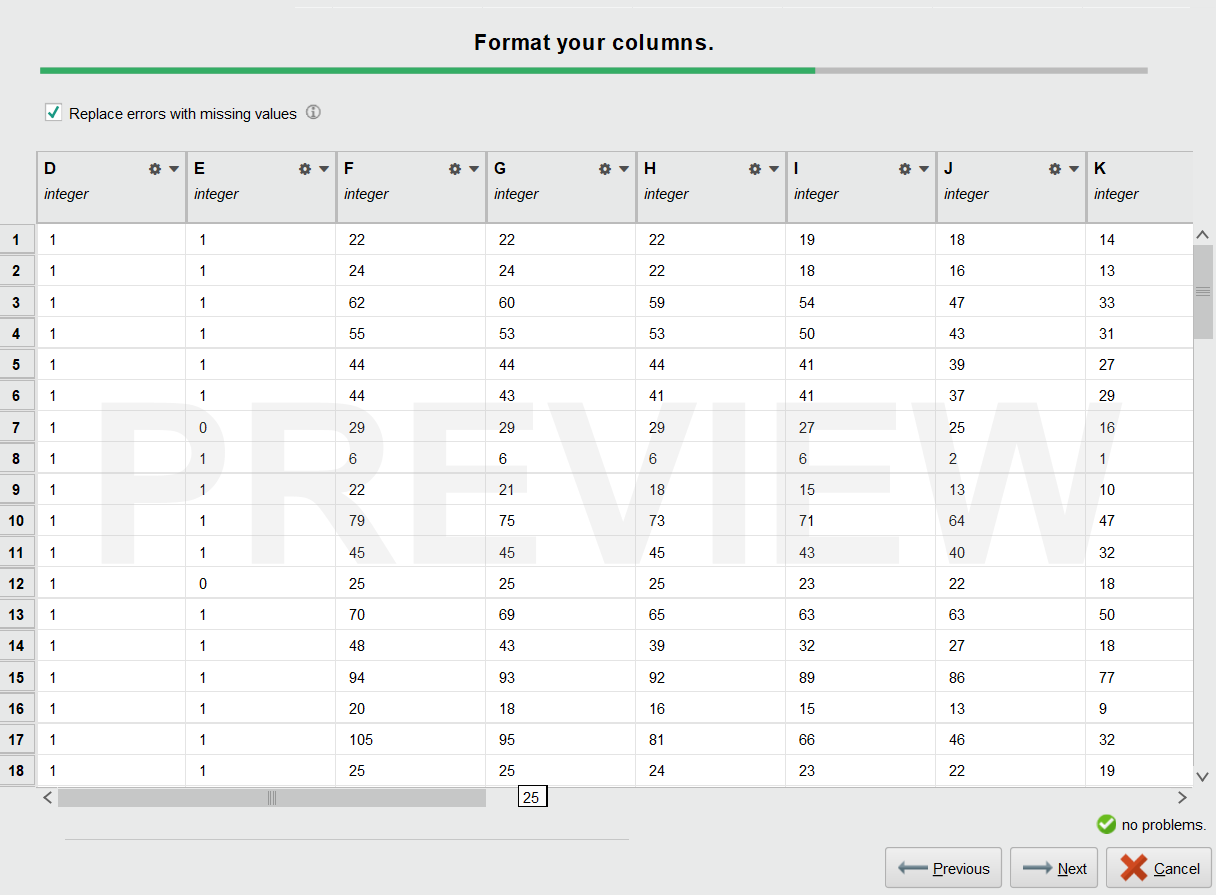
1. Pilih file yang akan diimpor



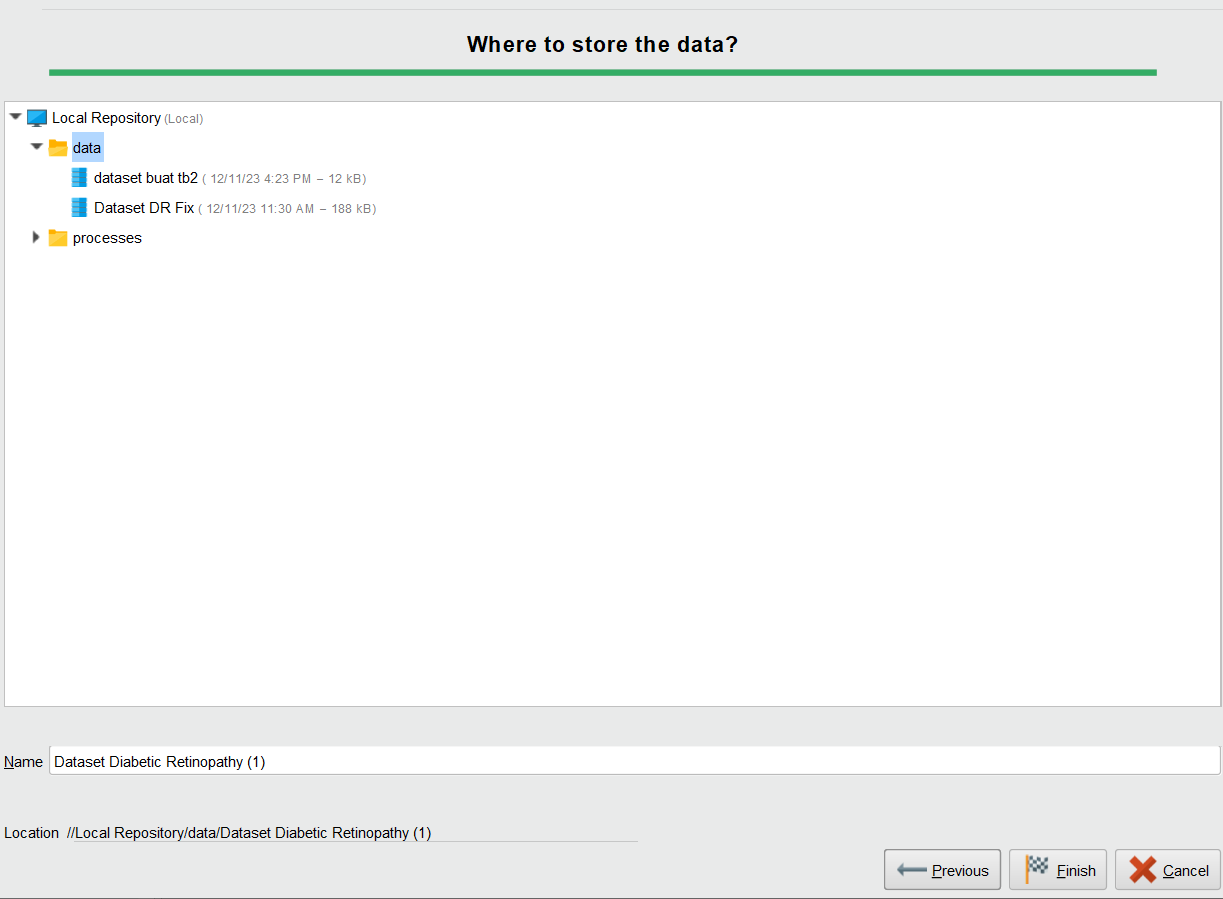
1. Pilih data yang ingin kita olah



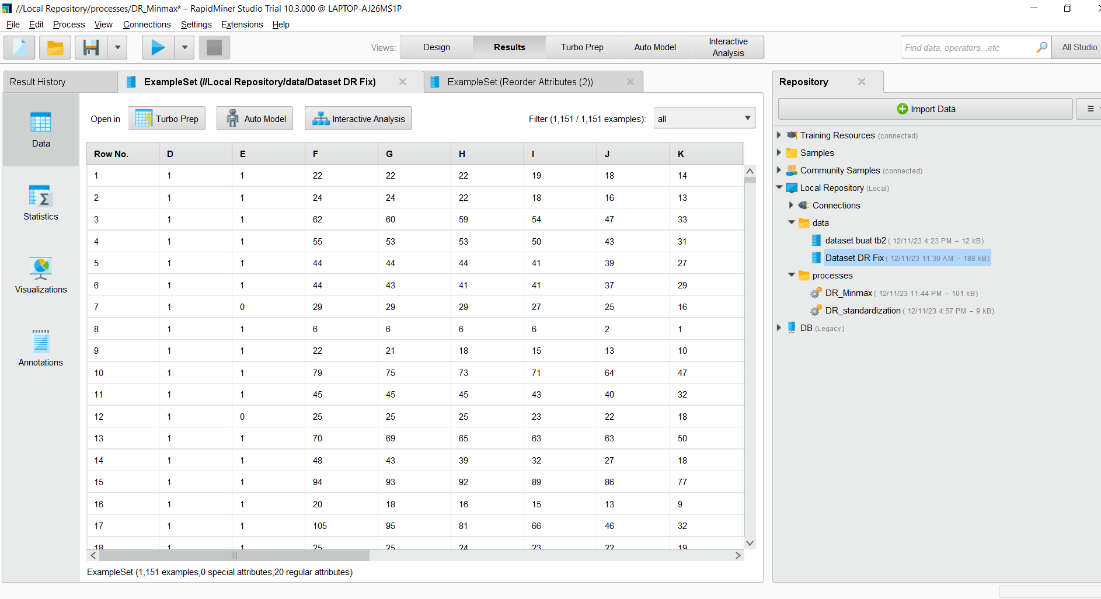
1. Cek kembali jumlah dan kesesuaian data. Periksa juga tipe data atribut dan kelas, atribut dari dataset ini seluruhnya bertipe data integer dan real, sedangkan kelas bertipe data category, jangan lupa untuk mengubah role kelas dengan label, yang menandakan bahwa itu merupakan kelas target dataset. Selanjutnya, beri ceklist pada Replace errors with missing values.



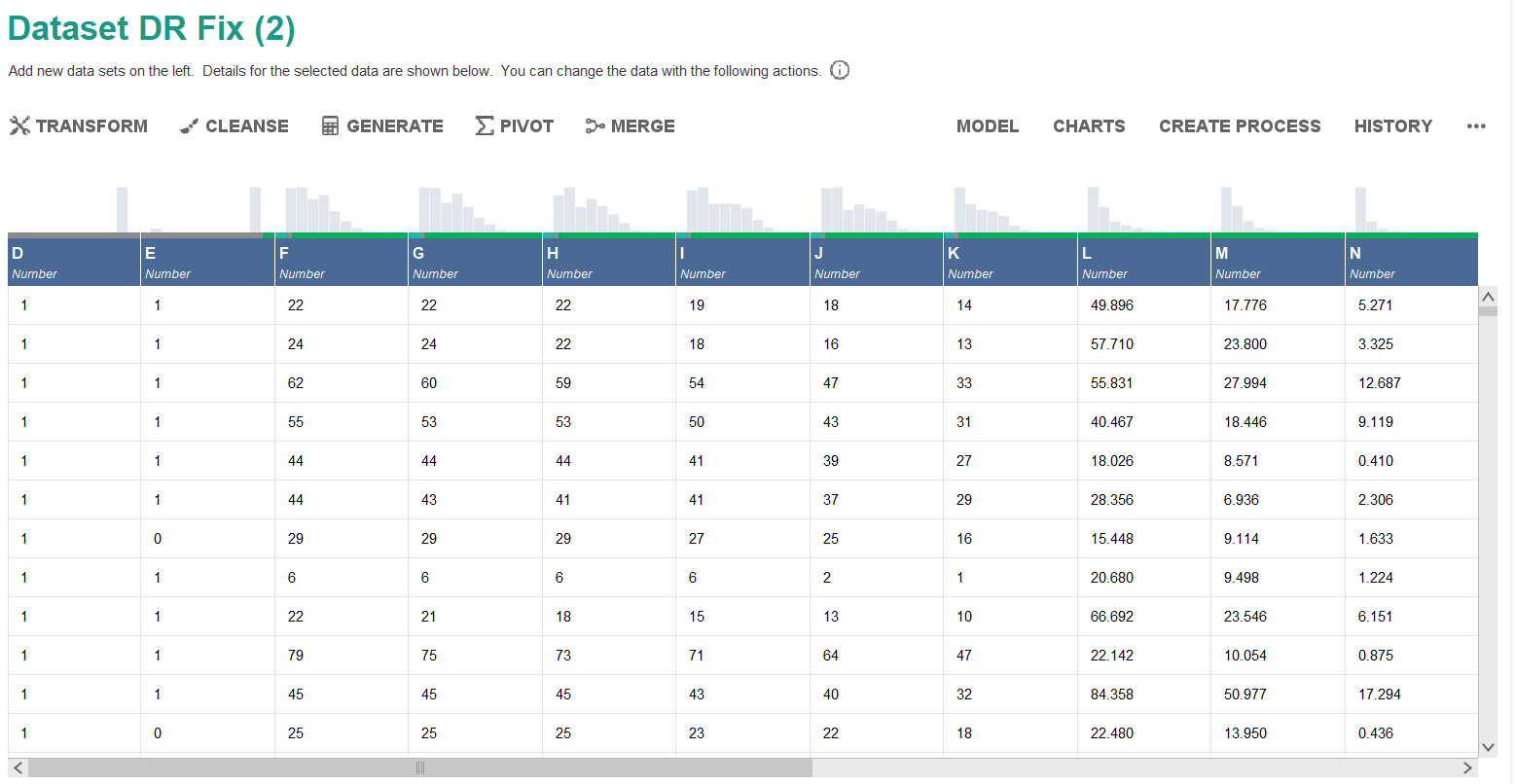
1. Pilih tempat untuk menyimpan dataset tersebut.



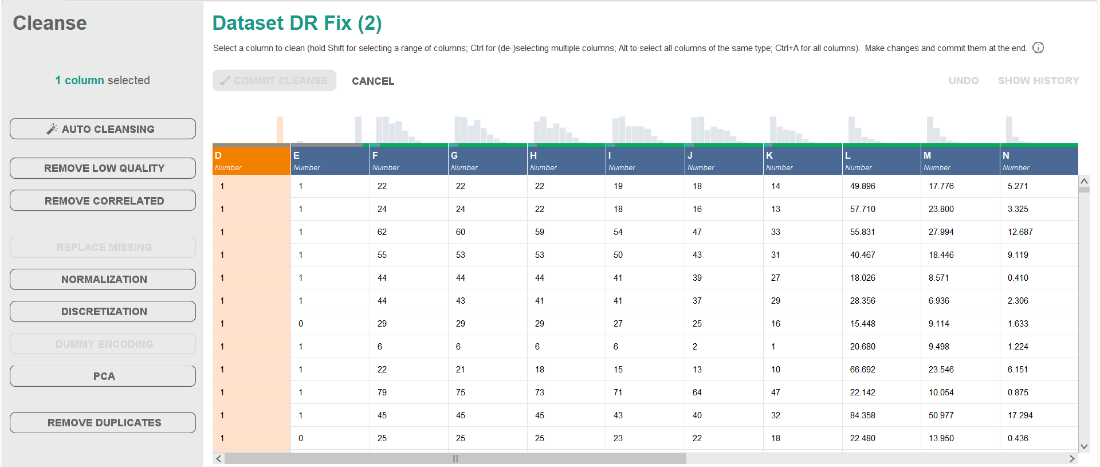
1. Setelah data diimpor, langkah selanjutnya adalah membuat proses-proses dengan Turbo Prep



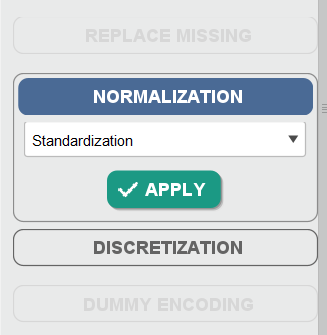
1. Dalam Turbo Prep, terdapat beberapa pilihan, seperti transform, kita dapat mengganti, nama, mungubah tipe, dan menghapus atribut. Pada pilihan cleanse, kita dapat menormalisasi dataset, bisa dengan metode auto cleansing, normalisasi manual seperti Min-Max, Standardization, dll.



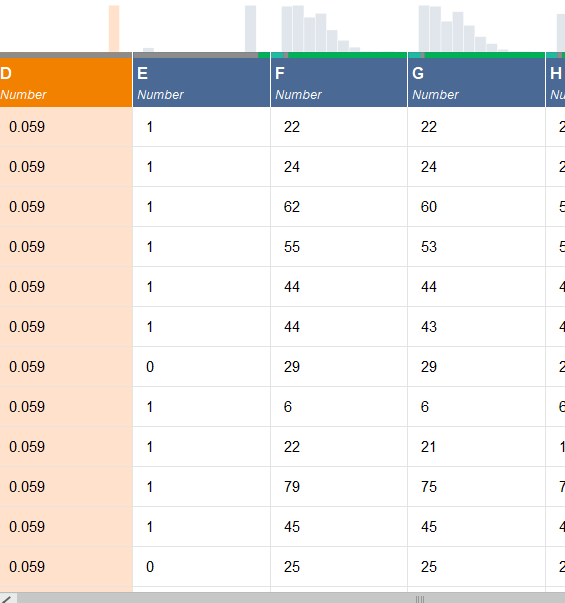
1. Jika kita ingin menormalisasi manual, kita bisa melakukannya satu demi satu setiap atribut.



1. Berikut merupakan pilihan untuk normalisasi.



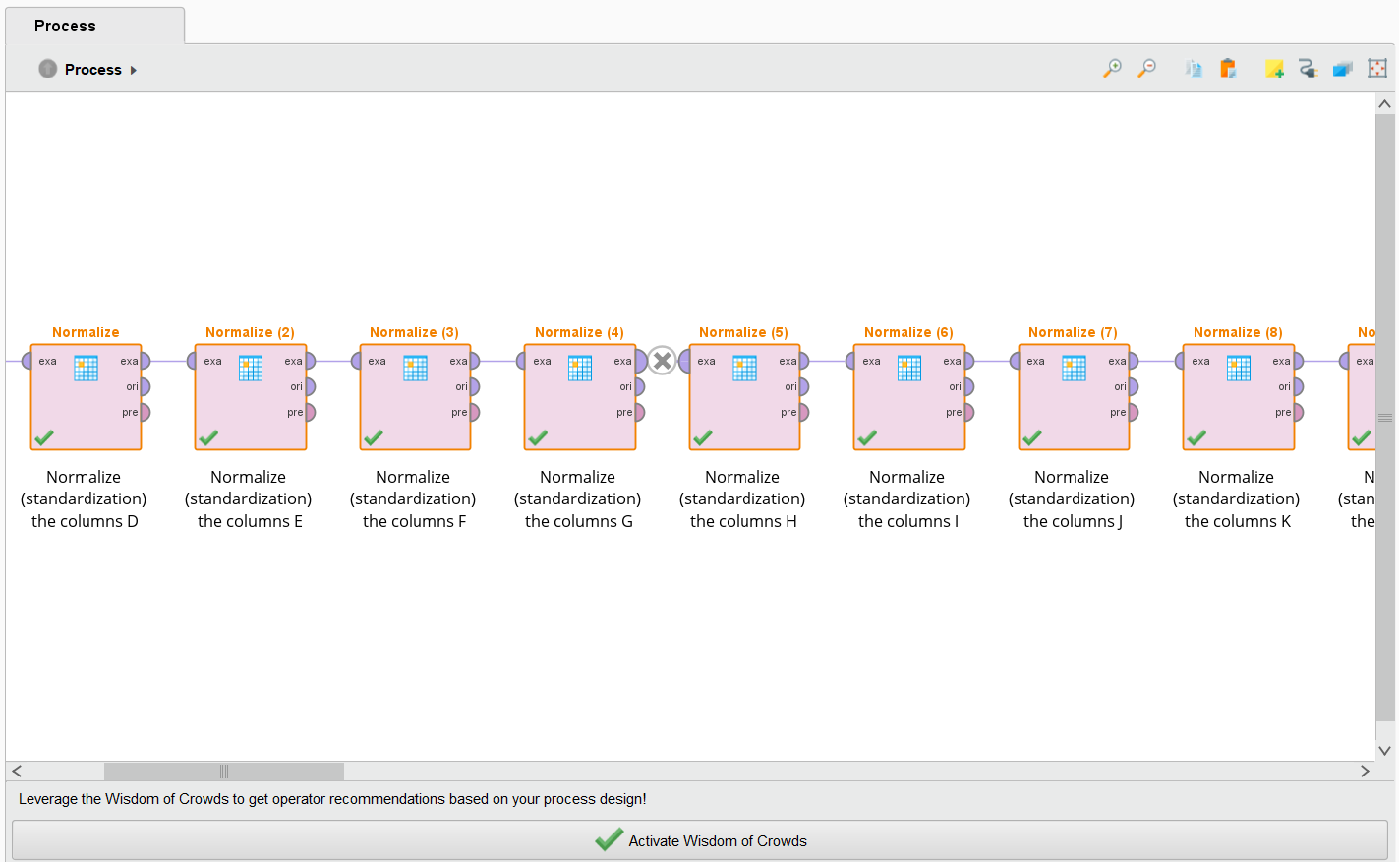
1. Setelah normalisasi, nilai data akan berubah, seperti pada gambar dibawah ini, menggunakan normalisasi standar pada atribut D.



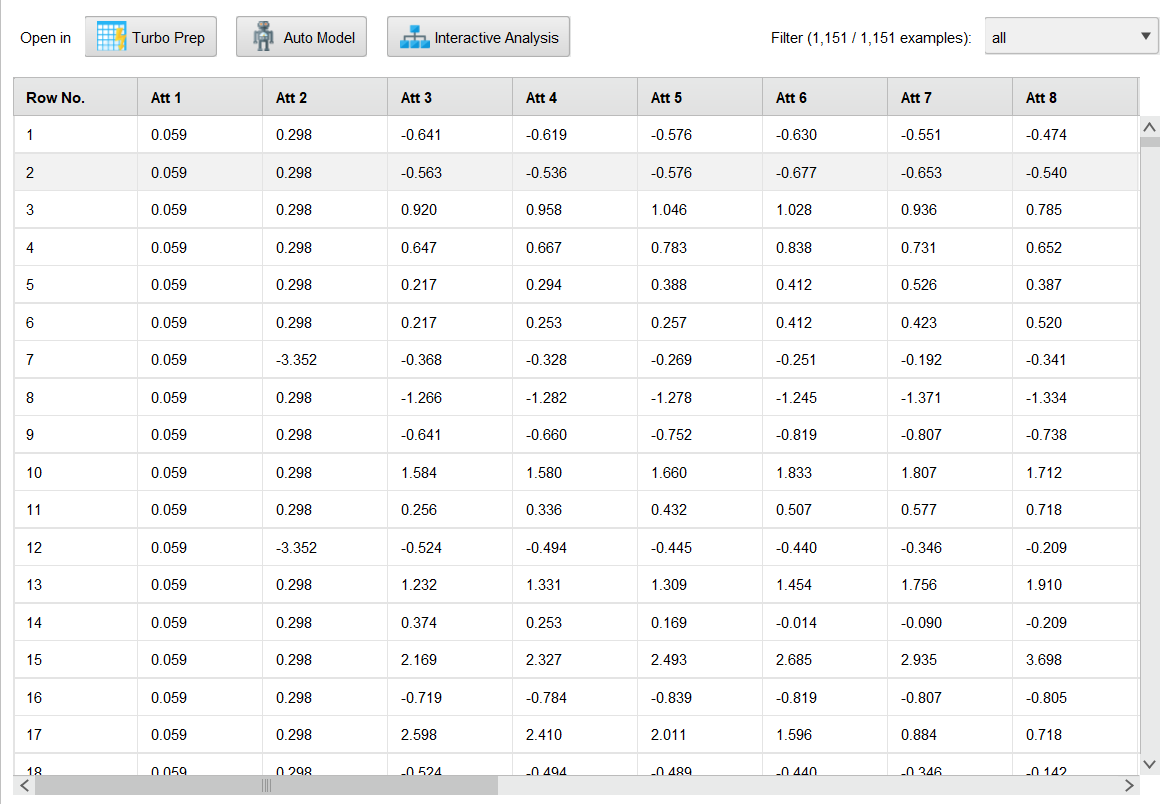
1. Setelah semua atribut dinormalisasi, jangan lupa simpan proses tersebut dengan create process.



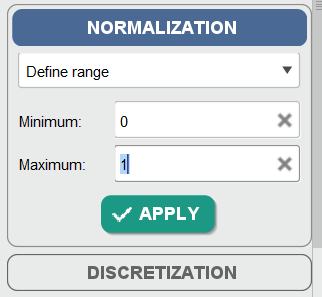
1. Berikut tampilan design pada proses, jika kita ingin menerapkan pada Result, kita dapat memilih semua proses yang ada, kemudian Run.



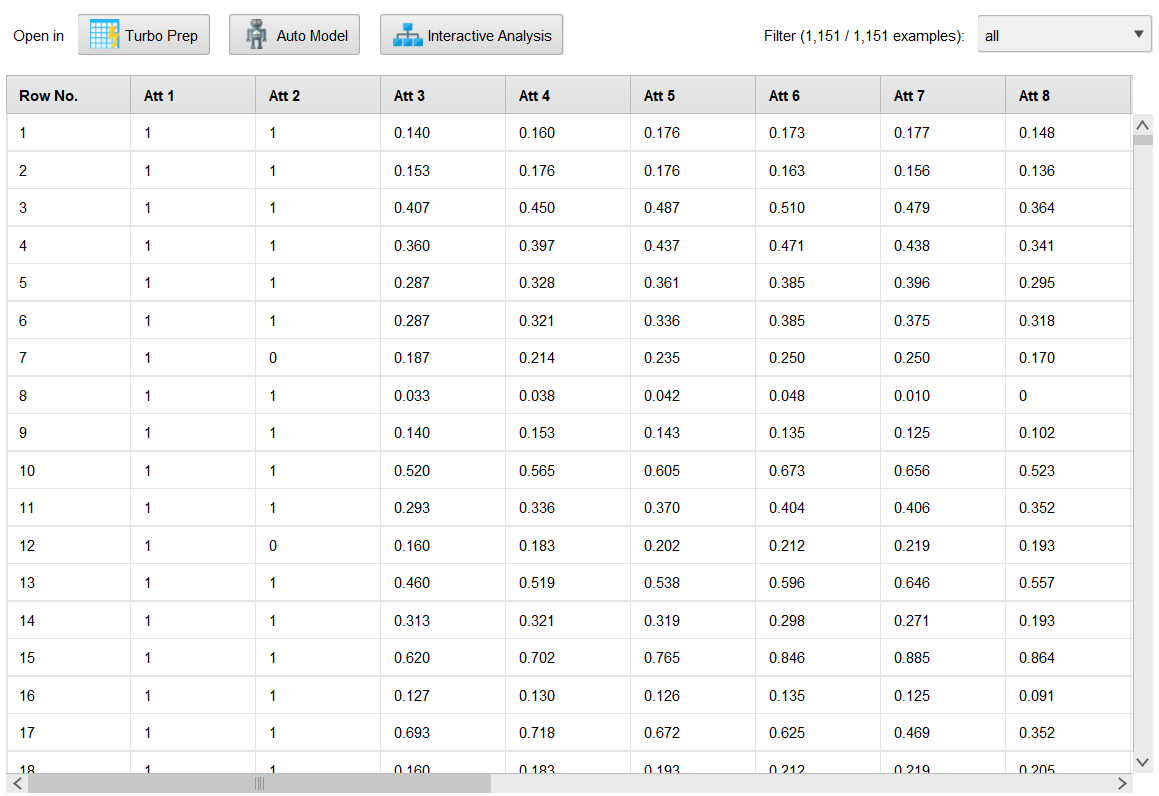
1. Setelah dijalankan, data pada Result akan berubah kedalam bentuk yang telah dinormalisasi.



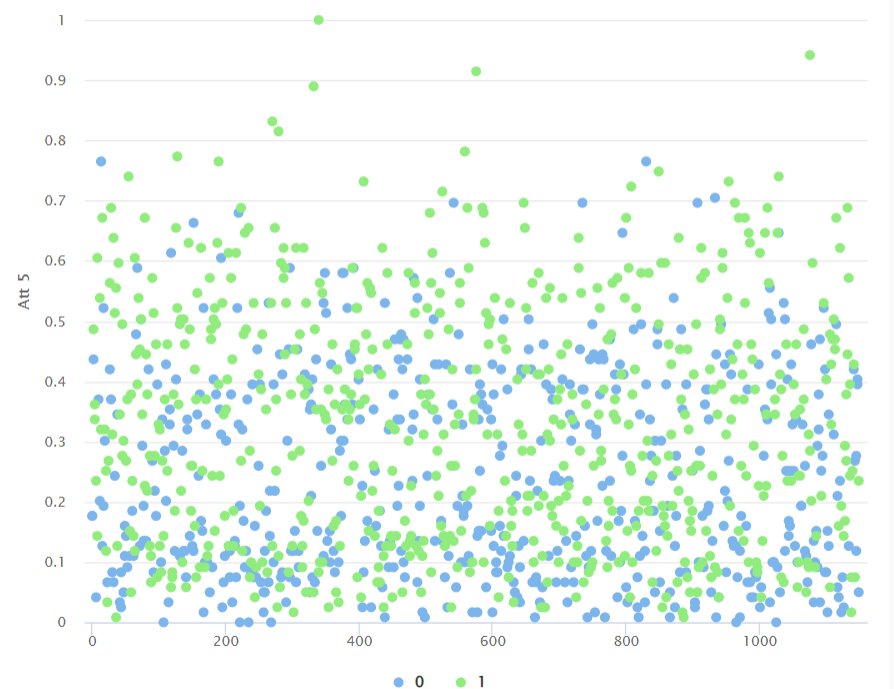
1. Berikut adalah jika kita menormalisasi data dengan normalisasi Min-Max, dengan range 0 dan 1.



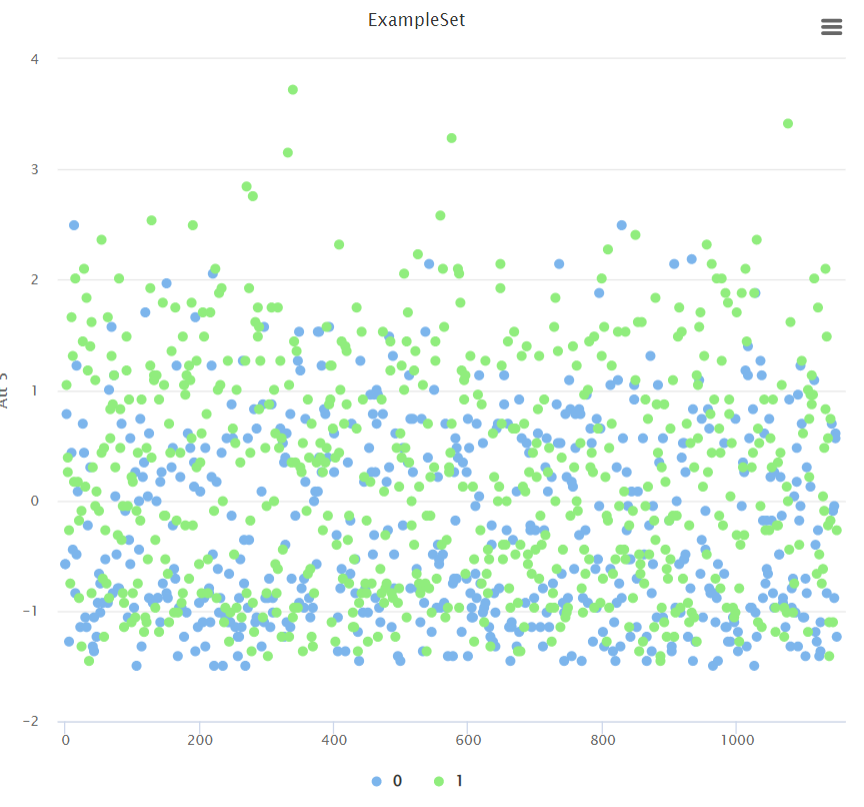
1. Berikut data jika kita menormalisasi dataset dengan normalisasi Min-Max



1. Tahap selanjutnya adalah mengekspor file yang sudah dinormalisasi, direkomendasikan menggunakan format .csv agar bisa digenerate pada aplikasi weka saat pengujian dengan algoritma klasifikasi.

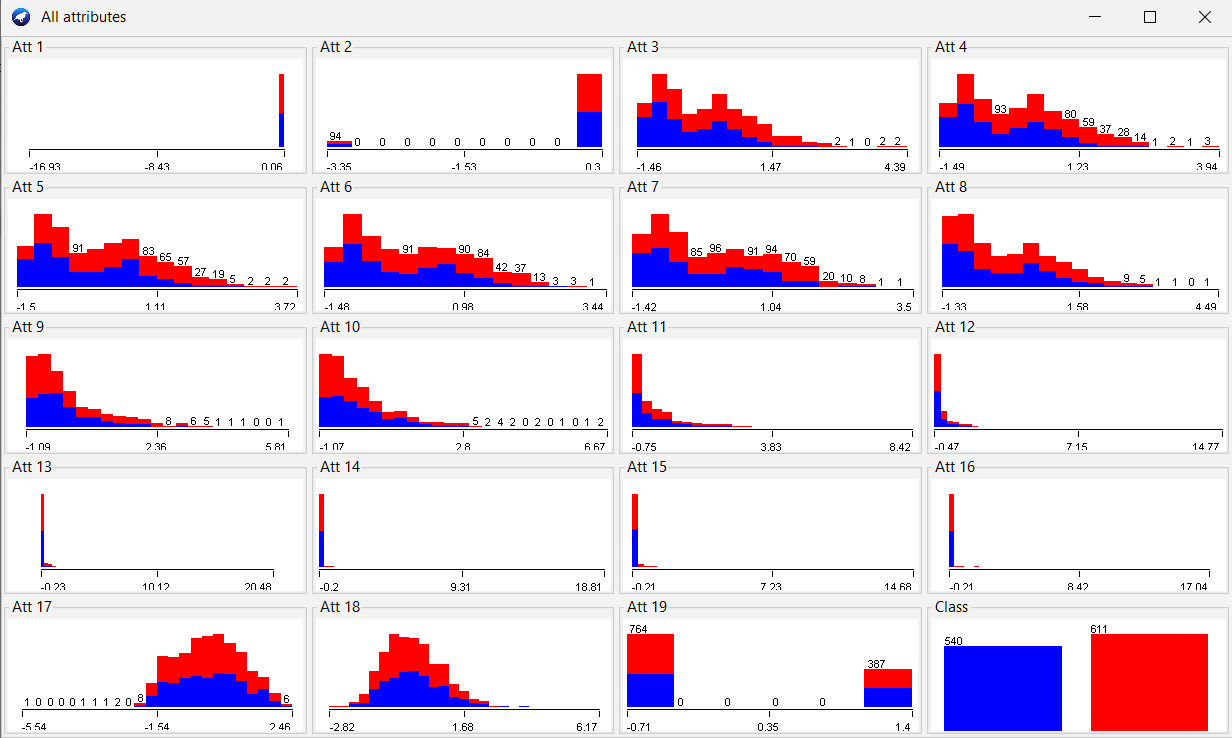


Gambar 1.5 Persebaran Data dengan Normalisasi Min-Max

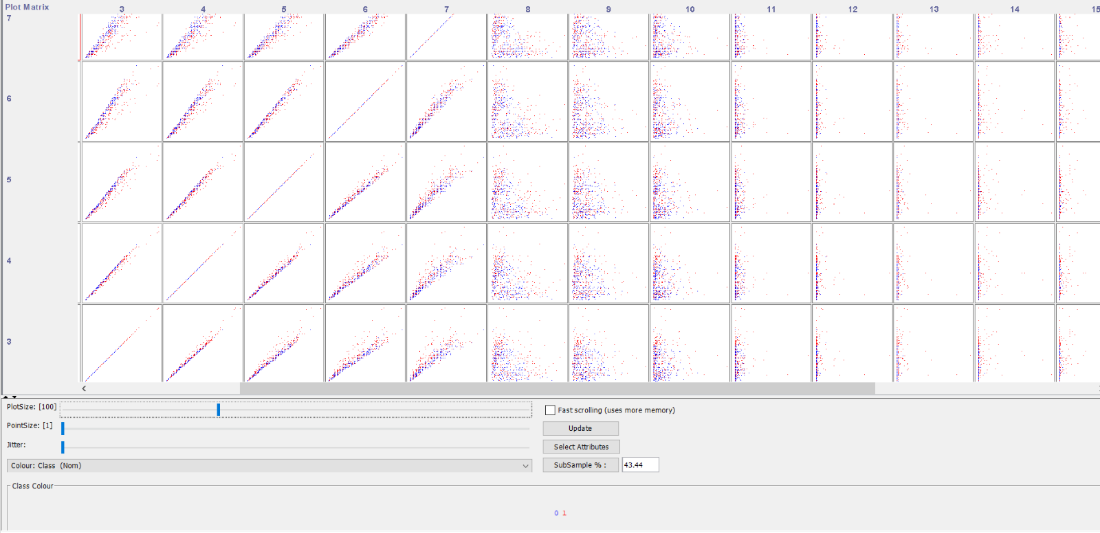


Gambar 1.6 Persebaran Data dengan Normalisasi Standar

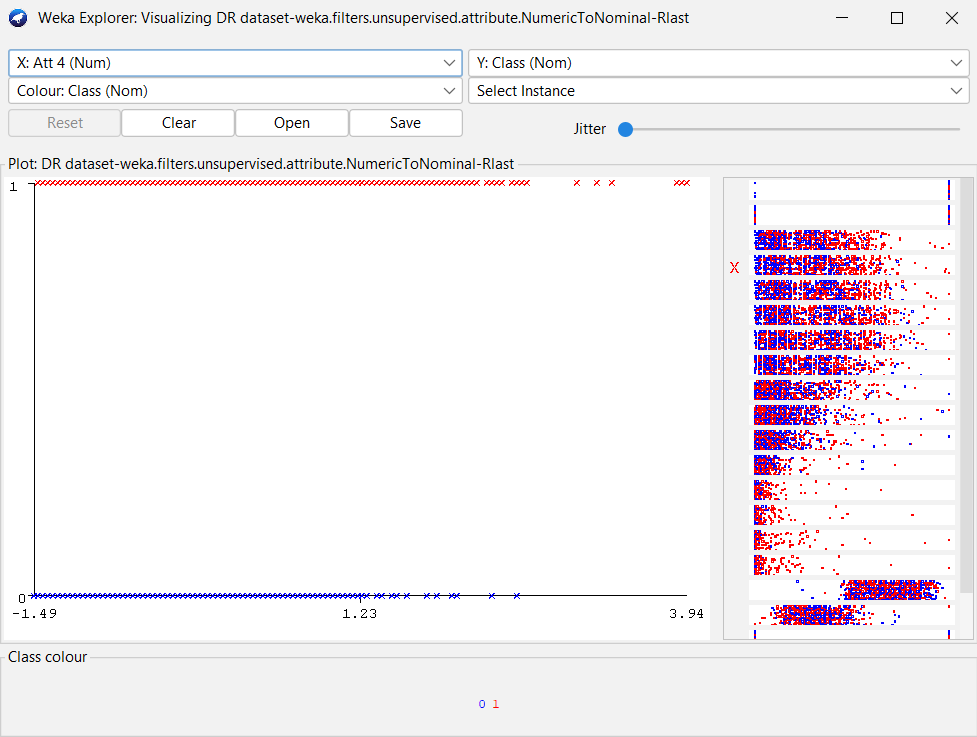
# Analisis Perilaku Data terhadap Kelas dengan Visualisasi



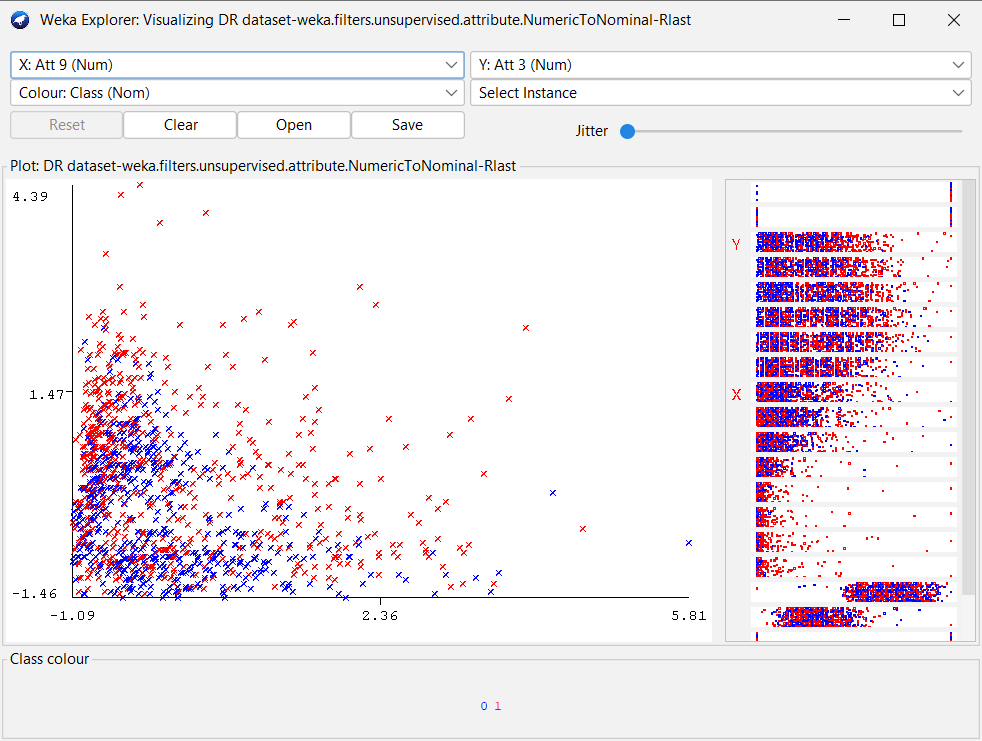
Grafik menunjukkan akumulasi jumlah pasien dengan dan tanpa diagnosa diabetes retinopathy. Warna biru menunjukkan pasien tanpa diagnosa, sedangkan warna merah menunjukkan pasien dengan diagnosa. Dari grafik, terlihat bahwa perilaku atribut dapat memengaruhi akumulasi kelas. Beberapa grafik menunjukkan sebaran yang merata, sementara yang lain memiliki mayoritas data terkumpul dalam satu rentang.



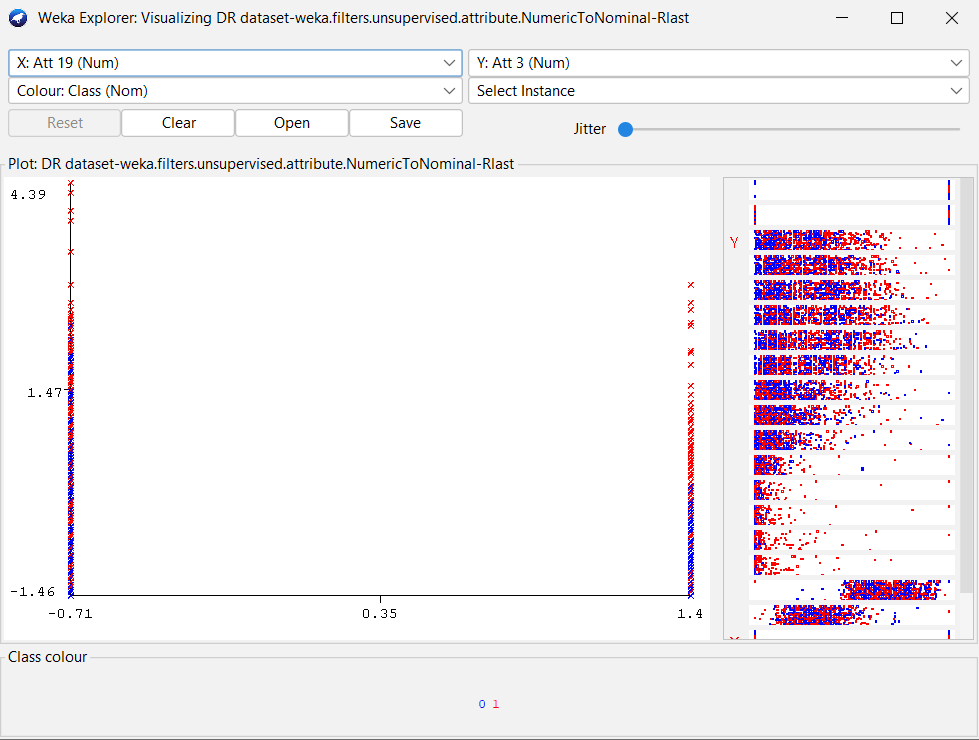
Gambar di atas menampilkan serangkaian visualisasi persebaran yang menggambarkan interaksi antara atribut dan juga dengan kelas. Terdapat 19x19 grafik visualisasi data, masing-masing menunjukkan karakteristik yang unik. Dalam grafik tersebut, teramati beberapa pola penyebaran data, baik yang tercampur maupun yang terpisah.



Grafik di atas menggambarkan kombinasi perilaku atribut 4 (deteksi MA) dan kelas. Terlihat bahwa seluruh titik merah terletak di y=1, sementara seluruh titik biru terletak di y=0, menunjukkan perbedaan kelas data dalam setiap entitas dataset. Atribut 3 ditampilkan pada sumbu X, dengan nilai minimum x=-1,49 dan nilai maksimum x=3,94. Mayoritas data terdistribusi dalam rentang x=-1,49 hingga x=1,23. Perbedaan akumulasi antar kelas cukup terlihat pada rentang x=1,23 hingga x=3,94, dimana terlihat bahwa kelas 0 memiliki jumlah data yang lebih sedikit dibandingkan kelas 1.

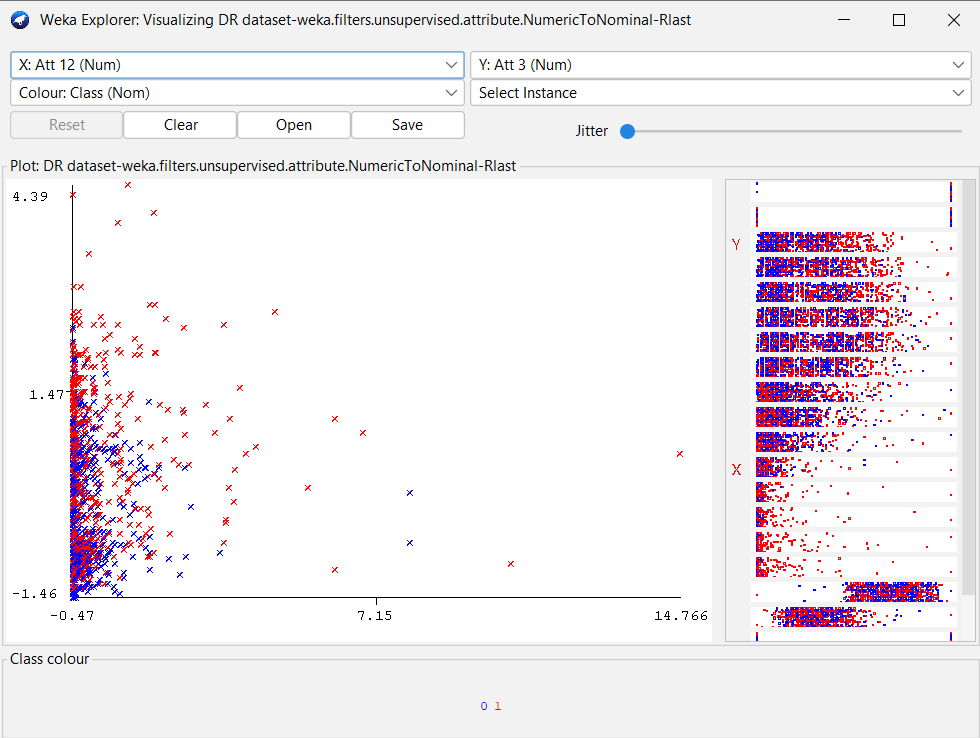


Grafik tersebut memperlihatkan korelasi antara atribut 3 (Deteksi MA) dan atribut 9 (Normalisasi pada nilai eksudat). Dari grafik, terlihat bahwa data cenderung terkumpul dan tercampur antara kelas 0 dan 1. Oleh karena itu, hanya menggunakan dua atribut tersebut saja tidak cukup untuk membuat keputusan yang jelas terkait dengan jenis kelasnya. Diperlukan atribut tambahan atau model yang lebih kompleks untuk melakukan klasifikasi yang lebih akurat.



Grafik tersebut mencerminkan perilaku data bersama-sama antara atribut 19 (Hasil Biner Kategorisasi Berbasis AM/FM) dan atribut 3 (Deteksi MA). Atribut 19 diproyeksikan pada sumbu X dengan dua basis kategori, yaitu hanya -0,71 dan 1,4. Hal ini menunjukkan bahwa data dengan nilai selain –0,71 dan 1,4 kemungkinan merupakan kesalahan penginputan, walaupun pada data tersebut jelas tidak terdapat data yang mengindikasikan kesalahan.

Terlihat bahwa grafik terbagi di sebelah kiri dan kanan. Sumbu Y memproyeksikan nilai atribut 3. Pada rentang y=-1,46 hingga y=1,47, mayoritas data berkumpul, sedangkan pada rentang y=1,47 hingga y=4,39, hanya sebagian kecil data. Pada rentang ini, mayoritas data termasuk dalam kelas 1



Grafik tersebut menggambarkan sebaran data dari korelasi antara atribut 12 (Normalisasi pada Eksudat) dan atribut 3 (Deteksi MA). Dari grafik, terlihat bahwa mayoritas data terkonsentrasi di sepanjang sumbu Y atau rentang x=-1,46 hingga 1,47. Ada sedikit nilai yang berada di atas rentang tersebut. Mayoritas data berkisar antara -0,47 hingga 7,15, namun terdapat beberapa data dengan nilai yang lebih tinggi, melebihi 14,76.

# Perbandingan Algoritma

|  |  |  |  |
| --- | --- | --- | --- |
| **Aspek** | CART | Neural Network | SGD |
| **General Explanation** | CART adalah singkatan dari Classification and Regression Trees, yang merupakan sebuah algoritma pengelompokan dan regresi yang menggunakan struktur pohon untuk memprediksi variabel target berdasarkan fitur-fitur yang ada. Algoritma CART ini diperkenalkan oleh Leo Breiman, Jerome Friedman, Richard Olshen, dan Charles Stone pada tahun 1984. Pohon keputusan yang dihasilkan oleh algoritma CART terdiri dari simpul (node) dan cabang (branch) yang mewakili keputusan dan pengelompokan data berdasarkan aturan yang didefinisikan. | Neural Network adalah termasuk salah satu cabang dari bidang soft computing.Kemampuan dasar otak manusia menjadi acuan untuk diadopsi dalam Neural Network. Kemmapuan tersebut adalah kemampuan untuk menerima stimulus, kemudian mengolah stimulus tersebut, dan memberikan tanggapan atau ouput sebagai hasil dari pengolahan stimulus. Variasi tanggapan atau Output merupakan fungsi dari dari variasi stimulasi dan kemampuan pengolahan yang dilakukan oleh otak | Stochastic Gradient Descent (SGD) adalah metode optimasi dalam machine learning untuk meminimalkan kesalahan model dengan mengatur parameter. SGD melakukan iterasi berulang, menghitung gradien fungsi error terhadap parameter, dan memperbarui parameter dengan learning rate. Fungsi error J(θ) dalam SGD melibatkan loss function (L) dan regularisasi (R). Algoritma berlanjut hingga kondisi berhenti terpenuhi, mencari parameter optimal untuk meminimalkan kesalahan dan mengoptimalkan model. SGD digunakan dalam berbagai konteks, termasuk pelatihan jaringan saraf tiruan dan regresi linier, memainkan peran kunci dalam mencari parameter yang paling sesuai dengan data. |
| **Kategori** | Klasifikasi | | |
| **Dataset Uji** | Diabetic Retinopathy Debrecen Dataset, UCI Machine Learning Repository | | |
| **Aplikasi Normalisasi** | *RapidMiner Studio* | | |
| **Metode Normalisasi** | Standardisasi | | |
| **Software Uji** | Waikato Environment for Knowledge Analysis | | |
| **Metode Pengujian** | *10 – Fold Cross Validation* | | |
| **Langkah Uji** | 1. Buka WEKA 2. Impor dataset 3. Sesuaikan atribut dan kelas 4. Cek persebaran data 5. Lakukan uji menggunakan Algoritma : 6. CART : J48 7. NN : MultilayerPerceptron 8. SGD : SGD 9. Uji folds dari 2-10 10. Ringkasan Uji | | |
| **Persen Keakuratan Model** | 63,4231% | 72.285% | 70,16% |
| **Precision** | 0,64 | 0.723 | 0,717 |
| **Recall** | 0,634 | 0.723 | 0,699 |
| **F1 - Measure** | 0,634 | 0.723 | 0,697 |
| **Performa** | Cukup Baik | Baik | Baik |

# Penerapan

Setelah menghitung persentase keakuratan menggunakan algoritma, kita dapat mengevaluasi seberapa banyak data yang diklasifikasikan dengan benar dan yang tidak. Misalnya, dengan akurasi sebesar 70% pada model klasifikasi Diabetes Retinopathy, artinya model dapat memprediksi dengan tepat kelas (ada atau tidak ada diabetes retinopathy) untuk sekitar 70% dari seluruh sampel dataset. Namun, untuk mendapatkan pemahaman yang lebih holistik, perlu diperhatikan pula metrik evaluasi lain seperti presisi, recall, dan F1-score. Setelah evaluasi, langkah selanjutnya mungkin melibatkan penyesuaian parameter model, peningkatan fitur, atau eksplorasi model lain untuk meningkatkan performa dan kemampuan generalisasi pada data yang belum pernah dilihat sebelumnya. Dengan demikian, dapat dilakukan iterasi dan penyempurnaan model untuk mencapai hasil yang optimal.

# Referensi

Antal, Balint and Hajdu, Andras. (2014). Diabetic Retinopathy Debrecen. UCI Machine Learning Repository. <https://doi.org/10.24432/C5XP4P>.

Hadistio, R. R., Mawengkang, H., & Zarlis, M. (2022). Perbandingan Algoritma Stochastic Gradient Descent dan Naïve Bayes Pada Klasifikasi Diabetic Retinopathy. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *6*(1), 271-277.

Subarkah, P. (2020). Penerapan Algoritme Klasifikasi Classification And Regression Trees (Cart) Untuk Diagnosis Penyakit Diabetes Retinopathy. *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, *19*(2), 294-301.

Arifin, T., & Herliana, A. (2018). Optimasi metode klasifikasi dengan menggunakan particle swarm optimization untuk identifikasi penyakit diabetes retinopathy. *Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika*, *4*(2), 77-81.