Modul 6

Regression

1.1. Tujuan Praktikum

Setelah menyelesaikan praktikum ini, mahasiswa mampu:

- Mengetahui tentang konsep Regression
- Mengimplementasikan Regression menggunakan RapidMiner

1.2. Peralatan yang dibutuhkan

Beberapa peralatan yang dibutuhkan dalam menyelesaikan praktikum ini adalah:

- Aplikasi RapidMiner versi 7 ke atas
- Dataset terkait

1.3. Dasar Teori

1.3.1. Regression

Regresi adalah salah satu teknik dalam statistika yang digunakan untuk memodelkan hubungan antara satu atau lebih variabel independen (disebut juga sebagai variabel prediktor atau variabel eksplanatori) dan variabel dependen (variabel respons). Tujuannya adalah untuk memahami dan menganalisis bagaimana perubahan dalam variabel independen mempengaruhi variabel dependen.

Berikut adalah beberapa konsep penting dalam regresi:

1. **Variabel Dependensi**: Variabel dependen adalah variabel yang ingin kita prediksi atau jelaskan. Ini adalah variabel yang kita harapkan akan dipengaruhi oleh variabel independen. Misalnya, jika kita sedang mempelajari hubungan antara waktu belajar dengan nilai ujian, maka nilai ujian adalah variabel dependen.

2. **Variabel Independensi**: Variabel independen adalah variabel yang digunakan untuk memprediksi atau menjelaskan variabel dependen. Ini adalah variabel yang kita harapkan

mempengaruhi variabel dependen. Dalam contoh sebelumnya, waktu belajar adalah variabel independen.

3. **Hubungan Linear**: Regresi linear adalah bentuk paling umum dari regresi di mana hubungan antara variabel independen dan dependen diasumsikan sebagai linear. Ini berarti bahwa perubahan dalam variabel independen berhubungan secara linier dengan perubahan dalam variabel dependen. Namun, ada juga model regresi non-linear yang memungkinkan hubungan yang lebih kompleks.

4. **Model Regresi**: Model regresi adalah representasi matematis dari hubungan antara variabel independen dan dependen. Model ini dinyatakan dalam bentuk persamaan matematis yang memungkinkan kita untuk memprediksi nilai variabel dependen berdasarkan nilai variabel independen.

5. **Koefisien Regresi**: Koefisien regresi adalah angka-angka yang menggambarkan kekuatan dan arah hubungan antara variabel independen dan dependen dalam model regresi. Dalam regresi linear sederhana, terdapat dua koefisien: koefisien kemiringan (slope) yang menunjukkan seberapa banyak variabel dependen berubah untuk setiap satuan perubahan dalam variabel independen, dan koefisien perpotongan (intercept) yang menunjukkan nilai rata-rata dari variabel dependen ketika variabel independen sama dengan nol.

6. **Evaluasi Model**: Evaluasi model regresi penting untuk mengetahui seberapa baik model tersebut cocok dengan data. Salah satu metode evaluasi yang umum digunakan adalah menggunakan metrik seperti R-squared (koefisien determinasi) untuk menentukan seberapa banyak variabilitas dalam variabel dependen yang dijelaskan oleh model.

Regresi digunakan dalam berbagai bidang seperti ilmu sosial, ekonomi, ilmu alam, dan banyak lagi untuk menganalisis dan memprediksi hubungan antara variabel-variabel yang ada. Teknik regresi juga memiliki berbagai variasi dan pengembangan yang telah dikembangkan untuk menangani situasi yang lebih kompleks.

Ada beberapa jenis regresi yang umum digunakan, dan berikut ini adalah beberapa di antaranya:

1. **Regresi Linear Sederhana**: Ini adalah bentuk paling dasar dari regresi di mana hubungan antara variabel independen dan dependen diasumsikan sebagai linear. Dalam regresi linear

sederhana, hanya ada satu variabel independen yang digunakan untuk memprediksi variabel dependen.

2. **Regresi Linear Berganda (Multiple Linear)**: Regresi linear berganda melibatkan lebih dari satu variabel independen untuk memprediksi variabel dependen. Model ini memungkinkan untuk mempertimbangkan pengaruh variabel-variabel independen yang lebih kompleks terhadap variabel dependen.

3. **Regresi Logistik**: Regresi logistik digunakan ketika variabel dependen adalah biner (dua kategori). Ini digunakan untuk memodelkan probabilitas kejadian suatu peristiwa berdasarkan satu atau lebih variabel independen. Misalnya, dapat digunakan dalam prediksi probabilitas keberhasilan atau kegagalan suatu peristiwa.

4. **Regresi Polinomial**: Regresi polinomial memodelkan hubungan antara variabel independen dan dependen dengan menggunakan fungsi polinomial daripada fungsi linear. Ini memungkinkan model untuk menangkap hubungan non-linear antara variabel-variabel tersebut.

5. **Regresi Ridge dan Lasso**: Regresi Ridge dan Lasso adalah metode regresi yang digunakan untuk menangani masalah multikolinieritas dan seleksi fitur dalam regresi linear berganda. Keduanya menggunakan regularisasi untuk mencegah overfitting dan meningkatkan kinerja model.

6. **Regresi Nonparametrik**: Regresi nonparametrik adalah pendekatan regresi yang tidak membuat asumsi tertentu tentang bentuk fungsional hubungan antara variabel independen dan dependen. Metode seperti regresi kernel dan regresi loess termasuk dalam kategori ini.

7. **Regresi Bayesian**: Regresi Bayesian menggunakan pendekatan Bayesian untuk memodelkan hubungan antara variabel independen dan dependen. Ini memungkinkan untuk memperhitungkan ketidakpastian dalam model dan memberikan distribusi probabilitas atas parameter-parameter model.

8. **Regresi Log-Linear**: Regresi log-linear digunakan ketika hubungan antara variabel independen dan dependen diungkapkan dalam bentuk logaritmik. Ini sering digunakan dalam analisis data tabulasi dan penelitian survei.

Setiap jenis regresi memiliki kegunaan dan asumsi yang berbeda, dan pemilihan jenis regresi yang tepat tergantung pada sifat data dan tujuan analisis.

Terdapat tiga model regression yang akan dibahas pada materi kali ini yaitu Regresi Linear, Regresi Logistik, dan Regresi Polinomial.

1. Regresi Linier (Linear Regression)

   Regresi linear adalah jenis regresi yang paling umum digunakan dan paling mudah dipahami. Ini melibatkan memodelkan hubungan linier antara satu atau lebih variabel independen (prediktor) dan variabel dependen (respons). Dalam regresi linear sederhana, hanya ada satu variabel independen yang digunakan untuk memprediksi variabel dependen, sedangkan dalam regresi linear berganda, terdapat lebih dari satu variabel independen.

   Secara matemtis, model regresi linear sederhana dapat dinyatakan sebagai berikut:

   $Y = \beta_0 + \beta_1 X + \varepsilon$

   Dimana:

   - Y adalah variabel dependen (respons).

   - X adalah variabel independen (prediktor).

   - β_0 adalah intercept (koefisien perpotongan), yaitu nilai rata-rata dari Y ketika X sama dengan nol.

   - β_1 adalah koefisien kemiringan, yang menggambarkan seberapa banyak Y berubah ketika X berubah satu satuan.

   - ε adalah kesalahan acak, yang merupakan perbedaan antara nilai yang diamati dan nilai yang diprediksi oleh model.

   Tujuan utama dari regresi linear adalah untuk memperkirakan nilai dari koefisien β_0 dan β_1 sehingga model yang dihasilkan dapat memprediksi nilai Y dengan akurat berdasarkan nilai X.

   Proses estimasi koefisien dalam regresi linear sering menggunakan metode kuadrat terkecil (Ordinary Least Squares/OLS). Metode ini mengurangi jumlah kesalahan kuadrat antara nilai yang diamati dan nilai yang diprediksi oleh model untuk menemukan estimasi yang optimal untuk koefisien.

Evaluasi kualitas model regresi linear umumnya dilakukan dengan menggunakan metrik seperti R-squared (koefisien determinasi), yang mengukur seberapa banyak variabilitas dalam variabel dependen yang dapat dijelaskan oleh model.

Regresi linear dapat diterapkan dalam berbagai bidang, termasuk ekonomi, ilmu sosial, ilmu alam, kesehatan, dan banyak lagi. Ini adalah alat analisis yang kuat untuk memahami hubungan antara variabel-variabel yang ada dalam data dan memprediksi nilai variabel dependen berdasarkan nilai variabel independen.

Pada percobaan kali ini, kita akan gunakan regresi linier pada Rapid Miner.

Dataset yang digunakan adalah dataset harga rumah (HousePrices).

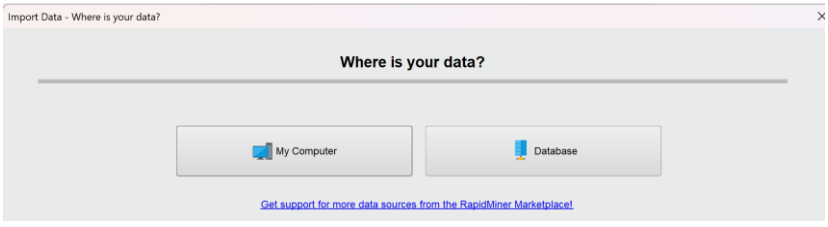
Perhatikan dataset **HousePrices.CSV** berikut:

house_sqft,num_of_bedrooms,num_of_bathrooms,year_built,tax_assessed_value,last_sold_price
1770,3,2,1990,195000,196358
1770,3,2,1990,195000,197715
1770,3,2,1990,195000,197816
1772,3,2,1990,200000,198011
1850,3,2.5,1990,200000,200530
1850,3,2.5,1990,200000,201805
1850,3,2.5,1990,205000,206175
1850,3,2.5,1990,205000,207027
1850,3,2.5,1990,205000,207121
1850,3,2.5,1990,205000,210519
1850,3,2.5,1990,212000,211099
1850,3,2.5,1990,212000,211274
1900,4,2.5,1990,212000,211414

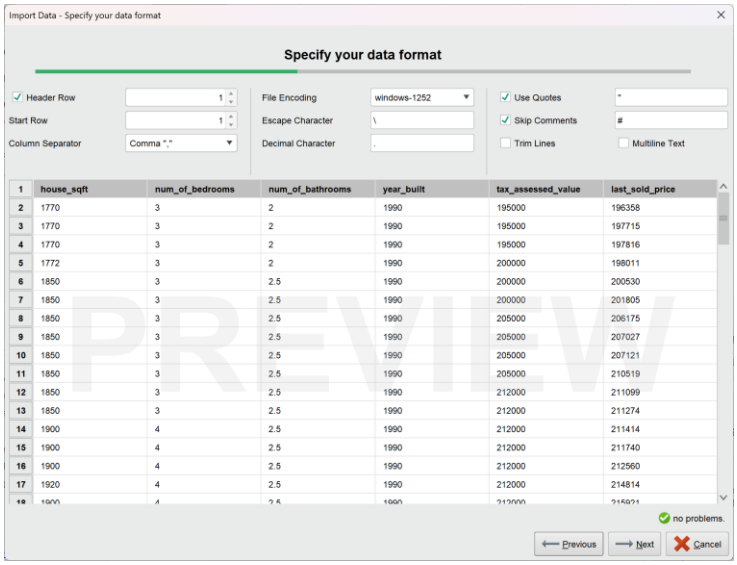
Data terdiri dari 6 atribut yaitu luas rumah (house_sqft), jumlah kamar tidur (num_of_bedrooms), tahun dibangun (year_built), biaya pajak (tax_assessed_value), harga jual (last_sold_price). Dataset terdiri dari 120 sampel data.

Percobaan 1:

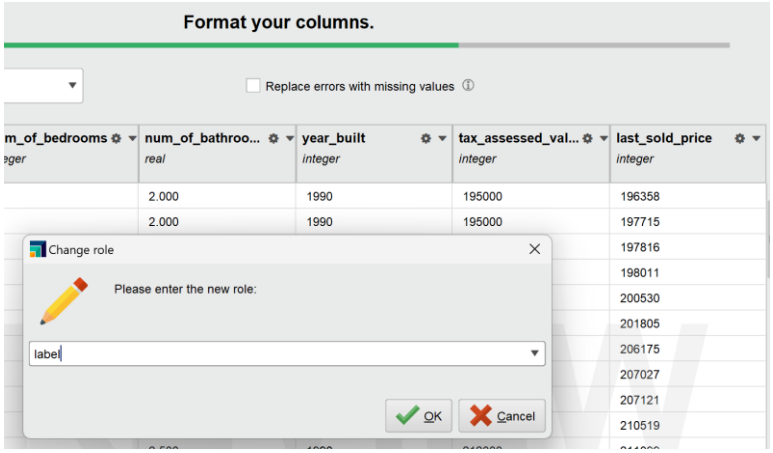
a. Import Dataset dari local disk



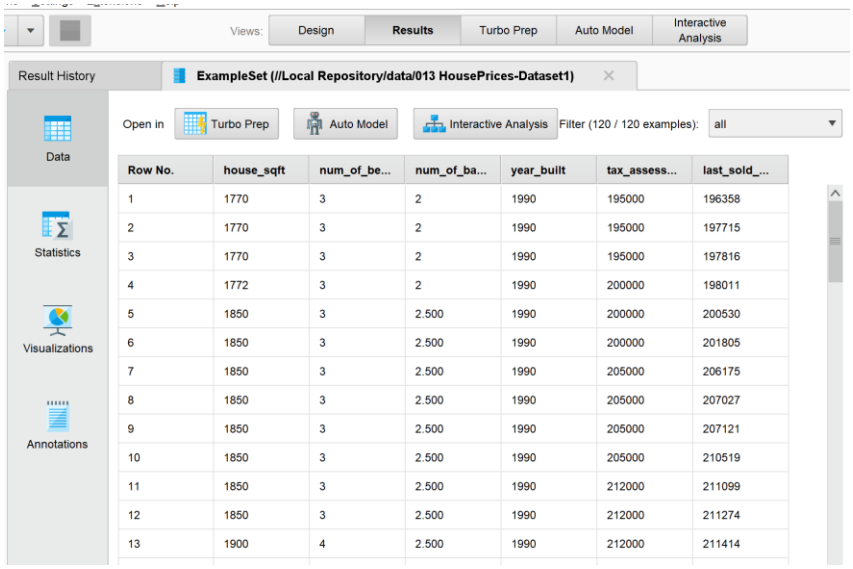
b. Pada jendela berikutnya seharusnya tertampil dataset dalam tabel yang lebih rapi



c. Langkah berikutnya adalah perlu untuk menentukan prediksi apa yang dilakukan dipilih sebagai label. Karena akan memprediksi harga rumah, maka last_sold_price perlu dipilih sebagai label.

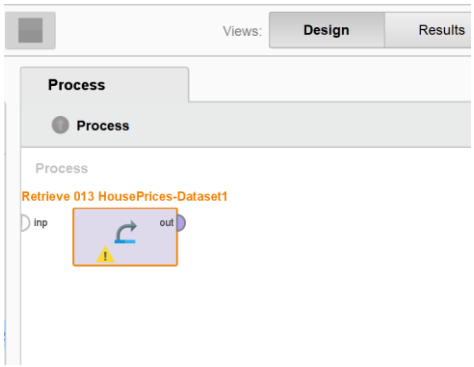


d. Jika berhasil, maka data akan ditampilkan pada menu Result seperti terlihat pada gambar berikut:

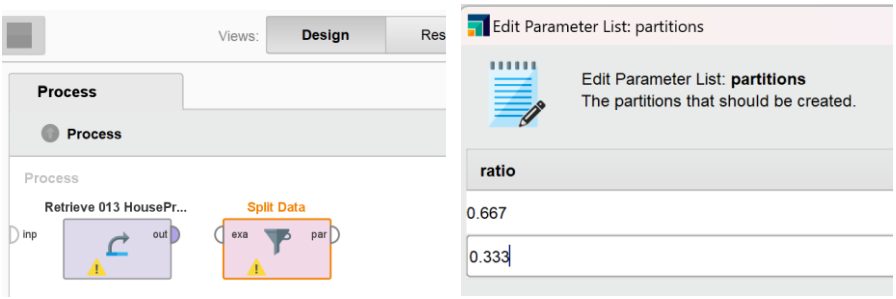


Row No.	house_sqft	num_of_be...	num_of_ba...	year_built	tax_assess...	last_sold_...
1	1770	3	2	1990	195000	196358
2	1770	3	2	1990	195000	197715
3	1770	3	2	1990	195000	197816
4	1772	3	2	1990	200000	198011
5	1850	3	2.500	1990	200000	200530
6	1850	3	2.500	1990	200000	201805
7	1850	3	2.500	1990	205000	206175
8	1850	3	2.500	1990	205000	207027
9	1850	3	2.500	1990	205000	207121
10	1850	3	2.500	1990	205000	210519
11	1850	3	2.500	1990	212000	211099
12	1850	3	2.500	1990	212000	211274
13	1900	4	2.500	1990	212000	211414

e. Langkah berikutnya yang dilakukan adalah dengan drag dataset tersebut dari jendela Repository dan drop pada jendela Design, sehingga muncul blok Retrieve… seperti terlihat pada gambar berikut:

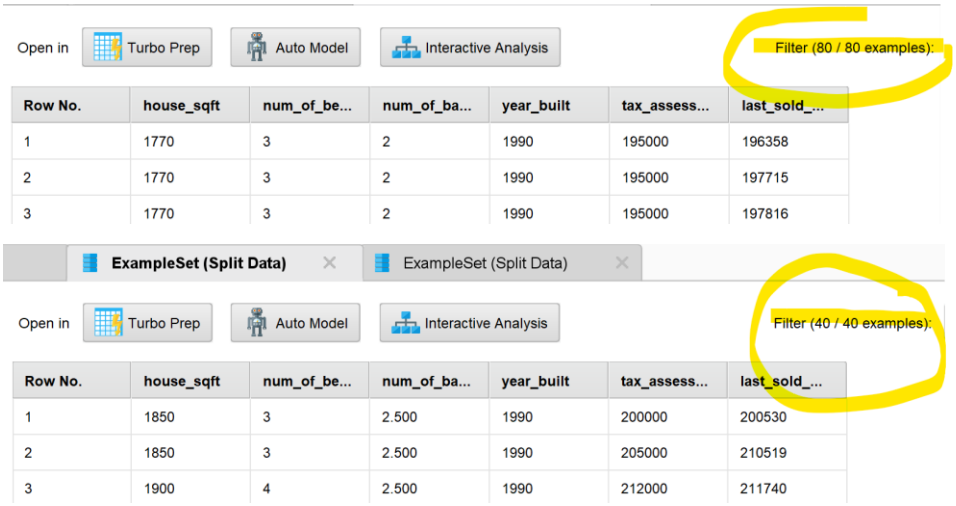


f. Perhatikan bahwa Regresi merupakan Machine Learning yang membuat model prediksi dengan menggunakan data pembelajaran (training data). Dataset yang tersedia kemudian perlu dibagi menjadi dua, data pembelajaran dan data pengujian (testing data). Data pembelajaran digunakan untuk membentuk model dengan akurasi sebaik-baiknya, sedangkan data pengujian digunakan sebagai data penguji model yang dibentuk dari data pembelajaran. Sampel data untuk pengujian tidak boleh pernah digunakan sebagai sampel data pembelajaran, sehingga langkah berikutnya yang perlu dilakukan adalah melakukan pemisahan data dari dataset (Split Data). Data akan dipisahkan menjadi data pembelajaran dan data pengujian. Secara umum biasanya dibagi dari dataset untuk pembelajaran sebanyak 2/3 sampel data dan sisanya untuk data pengujian. Gunakan operator Split Data untuk memisahkan dataset. Set Ratio dari Split Data -> Enumeration menjadi 0.667 untuk training dan 0.333 untuk testing.

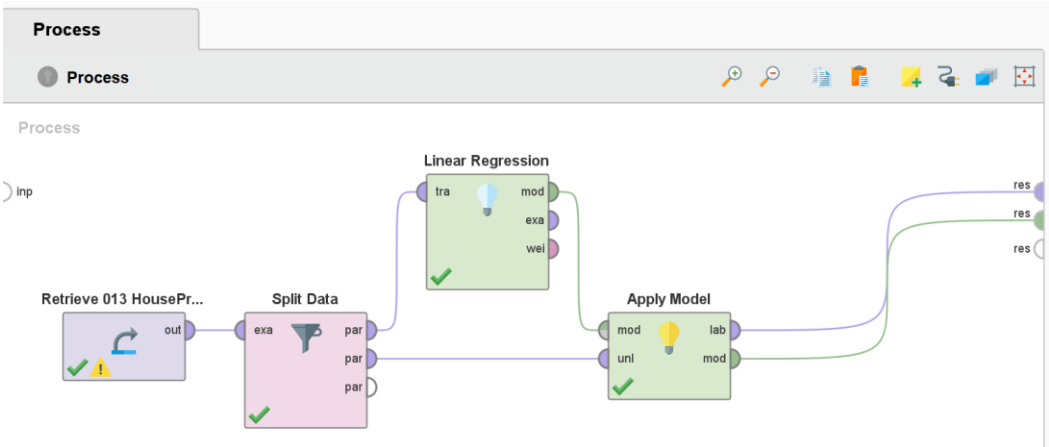


g. Hubungkan tiap relasinya seperti gambar berikut, jalankan dan lihat hasilnya:

h. Pada Data Training Terdapat 80 sampel data, dan Data Testing terdapat 40 sampel data



i. Karena akan menggunakan Linear Regression, maka pilih operator Linear Regression dan sambungkan pada data training Split Data. Model yang dihasilkan akan diuji coba menggunakan data testing dari test data, sehingga operator Apply model digunakan untuk operasi ini. Perhatikan gambar berikut untuk relasinya dan perhatikan juga hasilnya

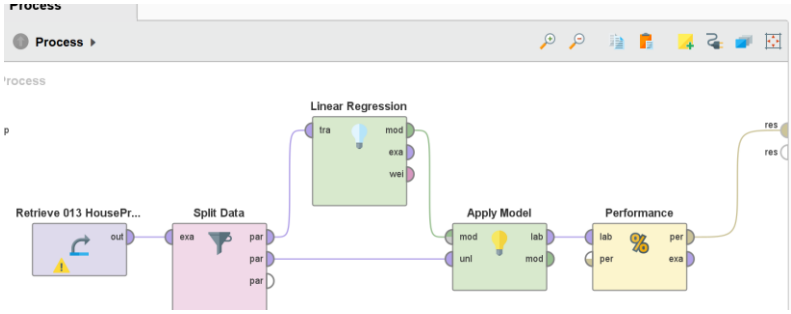


Row No.	last_sold_...	prediction(...	house_sqft	num
1	200530	201229.851	1850	3
2	210519	206200.805	1850	3
3	211740	213260.422	1900	4
4	215921	213260.422	1900	4
5	216623	216283.108	1920	4

LinearRegression (Linear Regression) | ExampleSet (Apply Model)

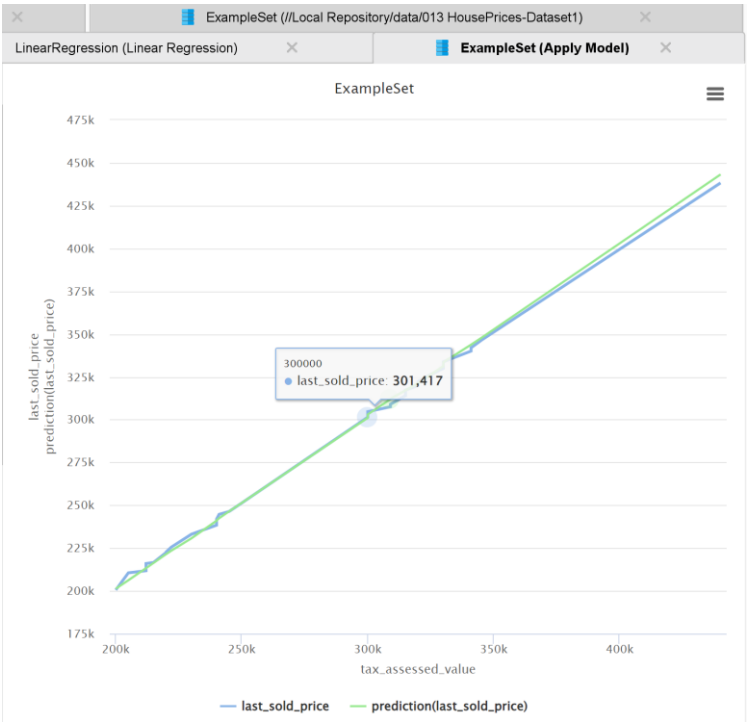
Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
house_sqft	2.006	0.790	0.018	0.554	2.538	0.013	**
num_of_bathro...	-1680.723	703.366	-0.011	0.736	-2.390	0.019	**
year_built	90.455	78.657	0.011	0.176	1.150	0.254	
tax_assessed_...	0.994	0.012	0.984	0.054	86.206	0	****
(Intercept)	-177122.074	154625.372	?	?	-1.145	0.256	

j. Blok Performance dapat ditambahkan untuk melihat sejauh mana error yang terjadi pada model yang dibentuk. Perhatikan gambar berikut dan lihat hasil RMSE nya. Tampilkan grafik antara data real dan data hasil prediksi dengan nilai sumbu X adalah Tax assessed dan perhatikan hasilnya. Tuliskan kesimpulan pada laporan anda.



root_mean_squared_error

root_mean_squared_error: 2085.444 +/- 0.000



2. Regresi Logistik (Logistic Regression)

Regresi logistik adalah teknik statistik yang digunakan untuk memodelkan hubungan antara satu atau lebih variabel independen (prediktor) dengan variabel dependen yang bersifat biner (dua kategori). Ini adalah alat yang sangat berguna dalam analisis klasifikasi dan prediksi di mana variabel dependen adalah variabel biner, seperti sukses/gagal, ya/tidak, atau 1/0.

Secara umum, regresi logistik menghasilkan probabilitas bahwa suatu kejadian terjadi berdasarkan nilai-nilai variabel independen. Dalam konteks klasifikasi biner, model regresi logistik mencoba memprediksi probabilitas bahwa suatu observasi akan termasuk dalam kategori yang ditentukan (misalnya, "positif" atau "negatif").

Model regresi logistik menghasilkan output dalam bentuk odds (rasio probabilitas sukses terhadap probabilitas gagal) yang diperlakukan sebagai fungsi linier dari variabel independen. Namun, karena nilai odds dapat bervariasi dari nol hingga tak terbatas, regresi logistik menggunakan fungsi logistik (sigmoid) untuk mengubah output menjadi probabilitas yang berada dalam rentang 0 hingga 1.

Secara matematis, regresi logistik dapat dijelaskan sebagai berikut:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}}$$

Dimana:

- $P(Y = 1|X)$ adalah probabilitas bahwa variabel dependen Y adalah 1 (sukses) berdasarkan nilai-nilai variabel independen X

-$\beta_0, \beta_1, \dots, \beta_n$ adalah koefisien regresi yang harus diestimasi.

- $X_1, X_2, \dots X_n$ adalah nilai-nilai variabel independen.

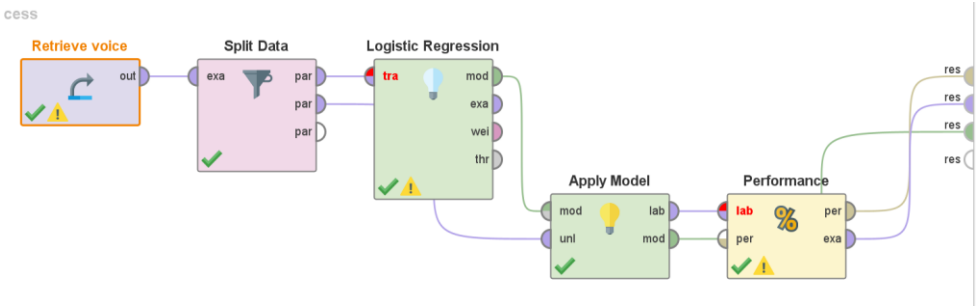
- e adalah basis logaritma natural (euler's number).

Koefisien regresi $\beta_0, \beta_1, \dots, \beta_n$ diestimasi menggunakan teknik optimasi seperti metode maksimum likelihood.

Regresi logistik sangat berguna dalam berbagai aplikasi seperti analisis risiko kredit, pemodelan risiko kesehatan, prediksi keberhasilan pemasaran, dan lain-lain. Model ini adalah alat yang penting dalam analisis data klasifikasi di mana variabel dependen adalah biner.

Percobaan 2:

a. Lakukan percobaan yang sama pada dataset Voice. Data ini berisi sampel suara dari pria dan wanita. Logistic Regression akan digunakan untuk memprediksi suara pria atau wanitakah pada data testing yang ditentukan. Tambahkan performance (binomial) untuk melihat performance dari model yang dibuat

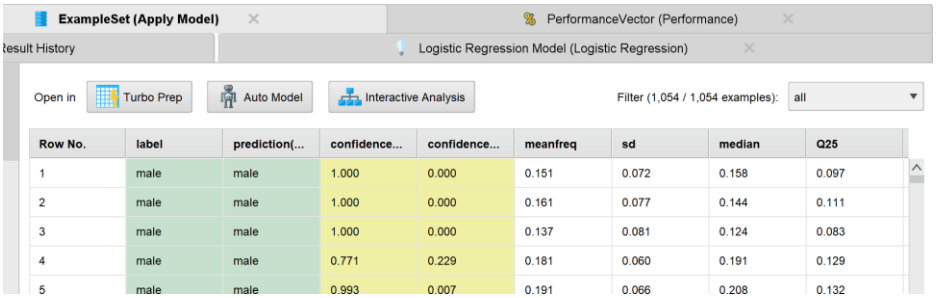


b. Perhatikan hasilnya, dan buat laporan terkait hasilnya

sult History | Logistic Regression Model (Logistic Regression)

Warning: Removed collinear columns [IQR, centroid, dfrange]

Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
meanfreq	-118.067	-3.548	62.985	-1.875	0.061
sd	-153.199	-2.558	50.482	-3.035	0.002
median	27.034	0.986	16.837	1.606	0.108
Q25	61.481	3.040	15.819	3.886	0.000
Q75	22.218	0.511	28.061	0.792	0.429
IQR	0	0	?	?	?
skew	-0.040	-0.173	0.203	-0.196	0.845
kurt	0.006	0.869	0.005	1.155	0.248



Row No.	label	prediction(...	confidence...	confidence...	meanfreq	sd	median	Q25
1	male	male	1.000	0.000	0.151	0.072	0.158	0.097
2	male	male	1.000	0.000	0.161	0.077	0.144	0.111
3	male	male	1.000	0.000	0.137	0.081	0.124	0.083
4	male	male	0.771	0.229	0.181	0.060	0.191	0.129
5	male	male	0.993	0.007	0.191	0.066	0.208	0.132

accuracy: 97.25%

	true male	true female	class precision
pred. male	512	14	97.34%
pred. female	15	513	97.16%
class recall	97.15%	97.34%	

3. Regresi Polinomial (Polynomial Regression)

Regresi polinomial adalah jenis regresi yang memodelkan hubungan antara variabel independen dan dependen menggunakan fungsi polinomial, bukan fungsi linear seperti dalam regresi linear. Ini memungkinkan kita untuk menangkap hubungan yang lebih kompleks antara variabel-variabel tersebut.

Dalam regresi polinomial, variabel independen tidak hanya dipertimbangkan dalam bentuk linear, tetapi juga dalam bentuk pangkat-pangkat tertentu (misalnya, kuadrat, kubik, dll.). Oleh karena itu, model regresi polinomial memiliki bentuk umum sebagai berikut:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots + \beta_n X^n + \varepsilon$$

Dimana:

- Y adalah variabel dependen.
- X adalah variabel independen.
- $\beta_0,\ \beta_1,\ \beta_2,\ \beta_3, \ldots, \beta_n$ adalah koefisien regresi yang harus diestimasi.
- ε adalah kesalahan acak.

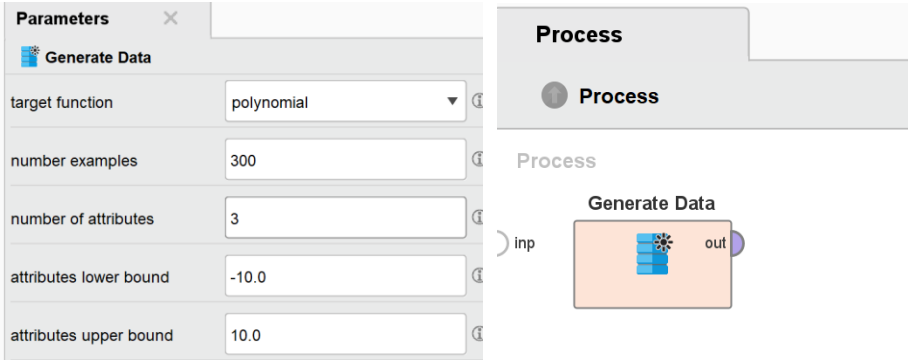
Jumlah pangkat tertinggi dari X dalam model polinomial ditentukan oleh derajat polinomial yang dipilih. Sebagai contoh, regresi polinomial derajat dua akan menghasilkan model dengan satu variabel independen X dan satu pangkat kedua dari $X(X^2)$, sedangkan regresi polinomial derajat tiga akan menambahkan pangkat ketiga (X^3), dan seterusnya.

Proses estimasi koefisien dalam regresi polinomial melibatkan teknik yang mirip dengan regresi linear, di mana kita mencari koefisien yang meminimalkan jumlah kuadrat kesalahan antara nilai yang diamati dan nilai yang diprediksi oleh model.

Regresi polinomial sangat berguna ketika hubungan antara variabel independen dan dependen tidak dapat dijelaskan dengan baik oleh model linear. Ini memungkinkan kita untuk menangkap pola-pola yang lebih kompleks dalam data. Namun, perlu diingat bahwa regresi polinomial dengan derajat yang tinggi cenderung lebih rentan terhadap overfitting, sehingga pemilihan derajat yang tepat penting untuk menghindari masalah tersebut.

Percobaan 3:

a. Pada dua percobaan sebelumnya anda menggunakan dataset yang telah ada, pada percobaan kali ini akan dibuat data baru berbentuk polinomial. Blok proses Generate Data dapat digunakan untuk membuat dataset khususnya polinomial (target function). Jumlah example dan atribut bisa kita set sesuai keinginan, pada percobaan ini dibuat 300 example dan 3 atribut. Perhatikan hasil generate datanya.

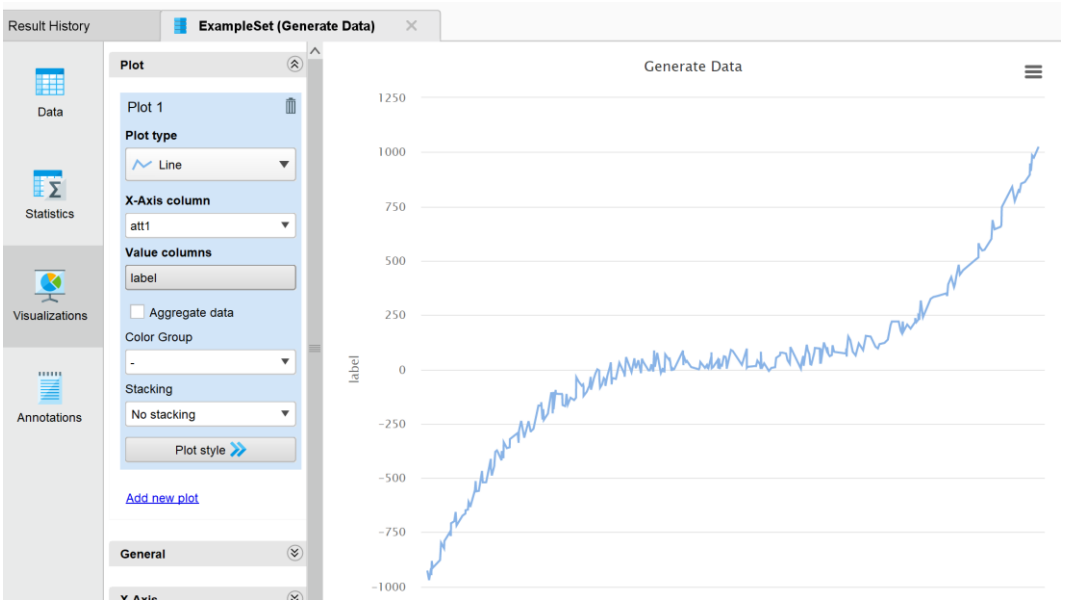


ExampleSet (Generate Data)

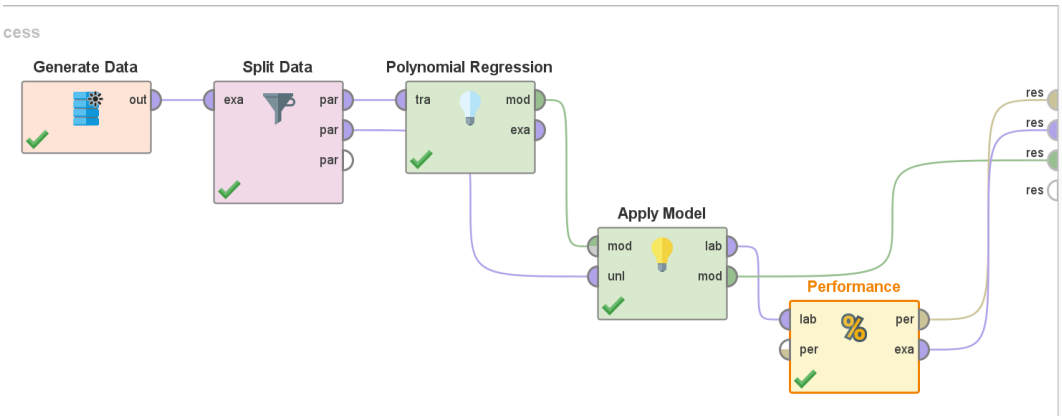
Open in Turbo Prep Auto Model Interactive Analysis

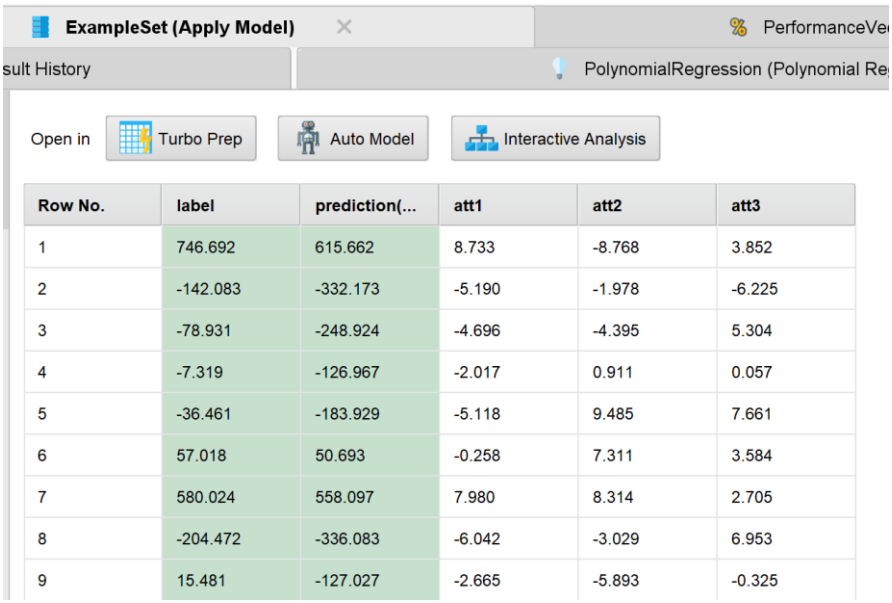
Row No.	label	att1	att2	att3
1	69.129	2.468	7.267	1.292
2	-610.619	-8.625	-5.590	-0.307
3	-167.123	-6.351	9.370	1.241
4	746.692	8.733	-8.768	3.852
5	-142.083	-5.190	-1.978	-6.225
6	-791.210	-9.413	-7.154	-8.453

b. Tampilkan grafik menggunakan line dan muncul grafik polinomial pada labelnya seperti gambar berikut:



c. Lakukan prediksi menggunakan Polynomial Regression dan perhatikan hasilnya:





Row No.	label	prediction(...	att1	att2	att3
1	746.692	615.662	8.733	-8.768	3.852
2	-142.083	-332.173	-5.190	-1.978	-6.225
3	-78.931	-248.924	-4.696	-4.395	5.304
4	-7.319	-126.967	-2.017	0.911	0.057
5	-36.461	-183.929	-5.118	9.485	7.661
6	57.018	50.693	-0.258	7.311	3.584
7	580.024	558.097	7.980	8.314	2.705
8	-204.472	-336.083	-6.042	-3.029	6.953
9	15.481	-127.027	-2.665	-5.893	-0.325

root_mean_squared_error

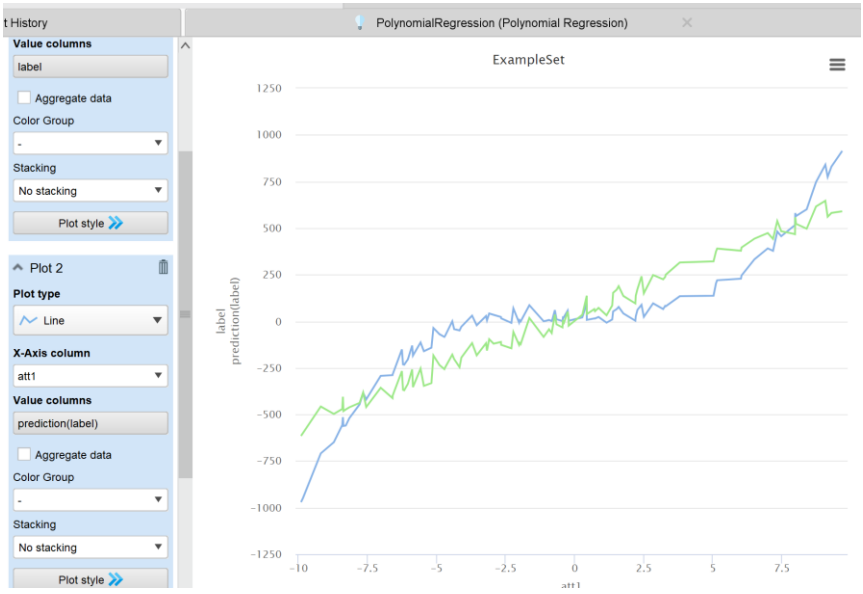
```
root_mean_squared_error: 135.437 +/- 0.000
```

ult History

PolynomialRegression

```
  59.689 * att1 ^ 1.000
+ 1.174 * att2 ^ 2.000
+ 3.097 * att3 ^ 1.000
- 7.725
```

d. Gambar berikut merupakan grafik hasil prediksi dan data asli dan bandingkan hasilnya. Catat hasil percobaan dan buat kesimpulan



I. Latihan Praktikum 1

1. Pilih dua dataset diantara dataset berikut dan lakukan Linear Regression dan tuliskan hasil percobaan anda. Awali dengan membuat deskripsi terkait dataset tersebut sesuai dengan deskripsi dari sumbernya!
   a. https://archive.ics.uci.edu/dataset/186/wine+quality
   b. https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho

c. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who
d. https://www.kaggle.com/datasets/quantbruce/real-estate-price-prediction
e. https://www.kaggle.com/datasets/mirichoi0218/insurance
f. https://www.kaggle.com/datasets/dgawlik/nyse
g. https://www.kaggle.com/datasets/spittman1248/cdc-data-nutrition-physical-activity-obesity

2. Pilih dua dataset diantara dataset berikut dan lakukan Logistic Regression dan tuliskan hasil percobaan anda. Awali dengan membuat deskripsi terkait dataset tersebut sesuai dengan deskripsi dari sumbernya!
   a. https://www.kaggle.com/datasets/nimapourmoradi/abalones-age
   b. https://www.kaggle.com/datasets/nimapourmoradi/adult-incometrain-test-dataset
   c. https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression
   d. https://www.kaggle.com/datasets/vikasukani/loan-eligible-dataset
   e. https://www.kaggle.com/datasets/nareshbhat/health-care-data-set-on-heart-attack-possibility
   f. https://www.kaggle.com/datasets/anyasorc/valentine-dataset
   g. https://www.kaggle.com/datasets/kandij/diabetes-dataset
3. Gunakan data harga labu berikut untuk membuat prediksi menggunakan Polynomial regression dan tuliskan hasil percobaan anda
   https://www.kaggle.com/code/residentmario/pumpkin-price-polynomial-regression/input?select=new-york_9-24-2016_9-30-2017.csv

--- SELAMAT BELAJAR ---