Repistory Data

Data yang digunakan diambil dari situs Kaggle.com Dimana Dataset Penjualan Nike mencakup 9.360 transaksi penjualan di berbagai wilayah AS mulai dari 1 Januari 2020 hingga 31 Desember 2021. Terstruktur dalam format CSV, dataset ini mendetailkan jenis produk, metode penjualan, dan informasi pengecer. Variabel kunci meliputi Tanggal Faktur, Produk, Wilayah, Pengecer, Metode Penjualan, Negara Bagian, Harga per Unit, Total Penjualan, dan Unit Terjual. Ideal untuk analisis tren dan riset pasar, dataset ini berfungsi sebagai sumber daya untuk memahami perilaku konsumen, kinerja penjualan regional, dan popularitas produk dalam lanskap ritel Nike.

Sumber : https://www.kaggle.com/datasets/krishnavamsis/nike-sales

Tipe : CSV

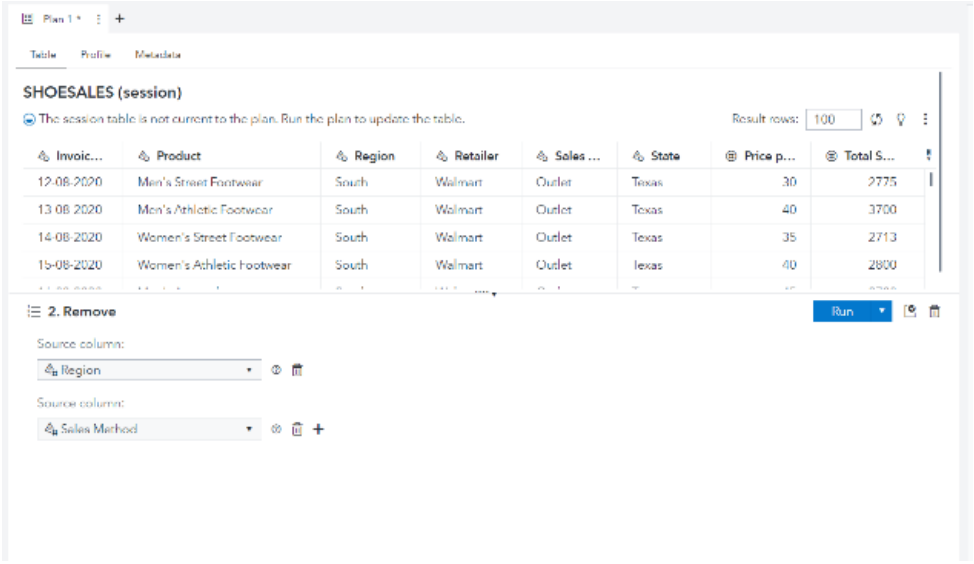
Variable	Tipe	Deskripsi
Invoice Date	Numerical	Merupakan tanggal transaksi dengan format DD-MM-YY
Price per Unit	Numerical	Harga per unit barang dalam dollar
Total Sales	Numerical	Jumlah keseluruhan penjualan untuk barang tersebut dalam dollar
Product	Character	Nama dari kategori produk yang terjual
Units Sold	Numerical	Jumlah keseluruhan unit barang yang terjual
Sales Method	Character	Jenis transaksi yang terjadi untuk penjualan
Region	Character	Daerah dimana transaksi itu dilakukan
Retailer	Character	Toko dari penjualan produk tersebut
State	Character	Negara bagian Dimana transaksi dilakukan.

Note: Dalam ZIP saya masukan 2 dataset (Sebelum dan Sesudah dilakukan Data Preprocessing)

Link SAS:
https://v4e032.vfe.sas.com/links/resources/report?uri=%2Freports%2Freports%2F24575b71-8384-4912-8f99-da480030fadb

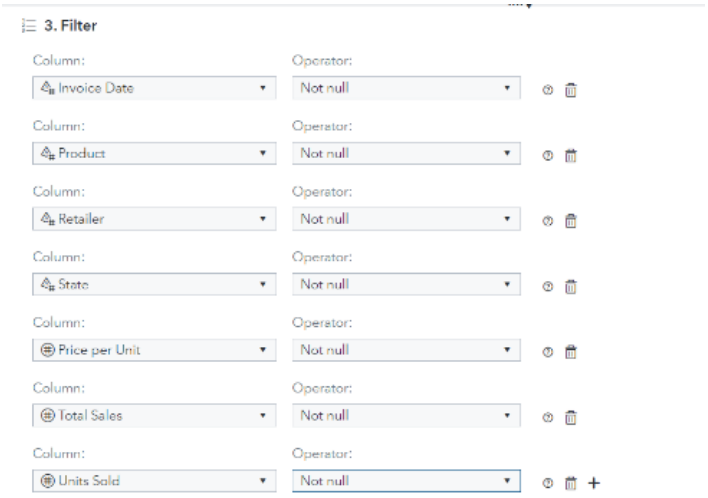
Melakukan Pre processing Data

Penghapusan Kolom



Pertama melakukan fungsi remove dimana, sebelum melakukan pemodelan akan dilakukan penghapusan kolom yang tidak digunakan terlebih dahulu, dalam konteks ini kolom Region dan Sales Method tidak akan digunakan, maka dapat dihapus menggunakan fungsi Remove pada SAS Prepare Data.

Penghilangan Data Null



Selanjutnya melakukan filter data untuk menghilangkan data yang null. Saat dijalankan, data akan tersaring agar semua data yang “Not Null” atau tidak Null akan terkumpul dan data yang Null akan dibuang.

Hasil

SHOESALES (session)

The session table is current to the plan.

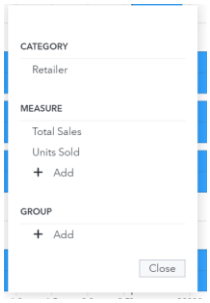
Result rows: 100

Invoice Date	Product	Retailer	State	Price per Unit	Total Sales	Units Sold
25-07-2020	Men's Street Fo...	Sports Direct	Texas	25	2125	85
26-07-2020	Men's Athletic F...	Sports Direct	Texas	35	2975	85
27-07-2020	Women's Street...	Sports Direct	Texas	35	2363	68
28-07-2020	Women's Athle...	Sports Direct	Texas	35	2188	63
29-07-2020	Men's Apparel	Sports Direct	Texas	40	2000	50
30-07-2020	Women's Apparel	Sports Direct	Texas	35	2450	70
31-07-2020	Men's Street Fo...	Sports Direct	Texas	30	2625	88
01-08-2020	Men's Athletic F...	Sports Direct	Texas	40	3500	88
02-08-2020	Women's Street...	Sports Direct	Texas	35	2450	70
03-08-2020	Women's Athle...	Sports Direct	Texas	40	2400	60
04-08-2020	Men's Apparel	Sports Direct	Texas	45	2250	50
05-08-2020	Women's Apparel	Sports Direct	Texas	40	2600	65
06-08-2020	Men's Street Fo...	Sports Direct	Texas	30	2700	90

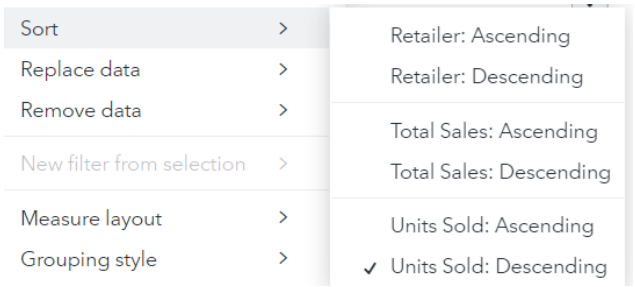
Setelah itu akan disimpan kedalam dataset terbaru bernama “Shoesale New” setelah melalui proses penyaringan untuk menghapus nilai null dan menghapus kolom yang tidak digunakan. Dataset terbaru ini sekarang menggunakan 7 kolom, sedangkan pada dataset sebelumnya memiliki 9 kolom.

Visualisasi Data

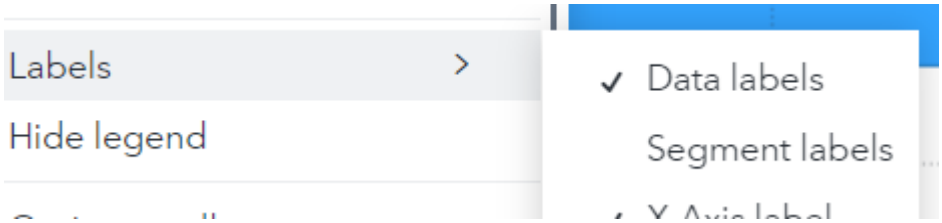
Barchart



Untuk Data roles saya masukan Retailer sebagai Category dan Total Sales dan Unit Sold sebagai Measures.

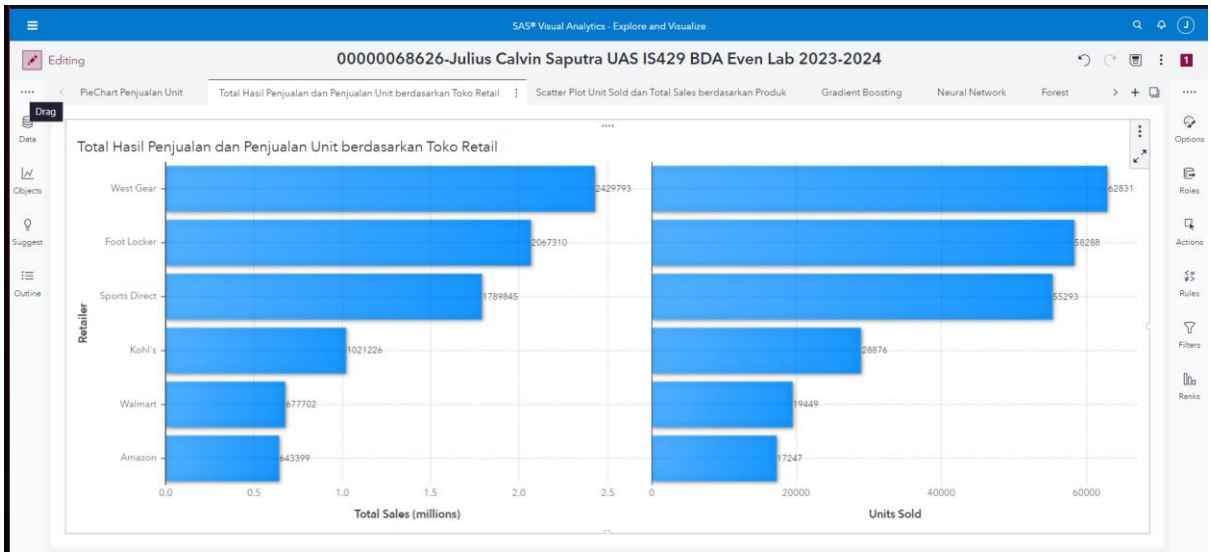


Selanjutnya saya melakukan sort Descending berdasarkan Unit Sold untuk menampilkan barchart dari paling tinggi hingga paling rendah.



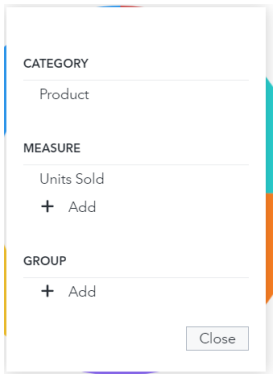
Lalu munculkan Data Labels agar lebih mudah untuk dibaca.

Hasil

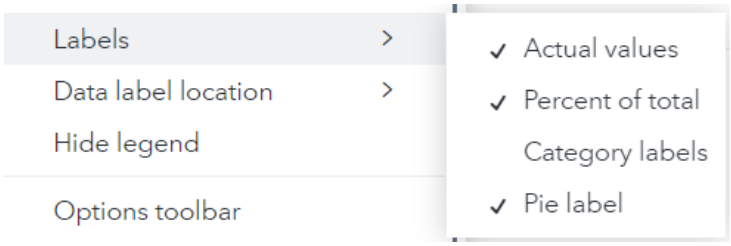


Berdasarkan piechart diatas West Gear memimpin dengan total penjualan sekitar 2,43 juta dolar dan jumlah unit terjual sebanyak 62.831. Foot Locker menyusul dengan total penjualan sekitar 2,07 juta dolar dan 58.288 unit terjual, diikuti oleh Sports Direct dengan penjualan sekitar 1,79 juta dolar dan 55.293 unit terjual. Diikuti oleh Kohl's, Walmart, dan Amazon berada di posisi terakhir dengan penjualan sebesar 643.399 dolar dan 17.247 unit terjual. Dari data ini, dapat disimpulkan bahwa West Gear adalah pemimpin pasar dalam hal total penjualan dan jumlah unit terjual, menunjukkan bahwa strategi mereka lebih efektif dalam menarik pelanggan. Di sisi lain, meskipun Amazon adalah platform besar, penjualannya relatif rendah dibandingkan dengan pesaing lainnya.

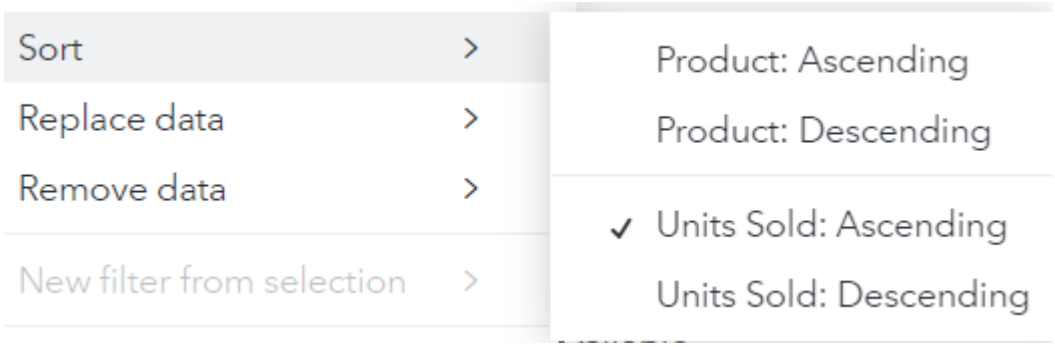
Piechart



Untuk Piechart saya memasukan Product sebagai category dan Unit Sold sebagai measures.

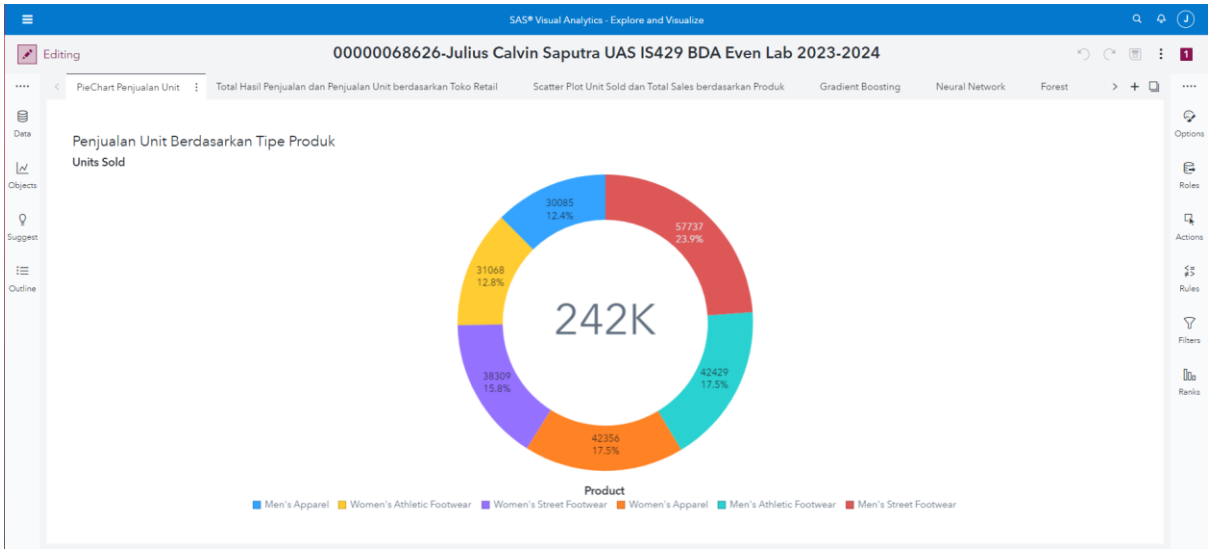


Lalu menampilkan Label berupa presentase (Percent of total) dan jumlah unit penjualan tiap produk (Actual Values).



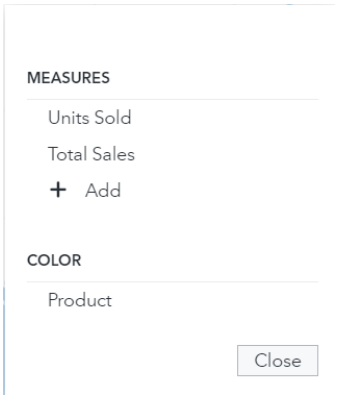
Selanjutnya saya urutkan piechart dari kategori produk dengan penjualan terendah berada di kiri hingga tertinggi berada di kanan piechart (Ascending).

Hasil



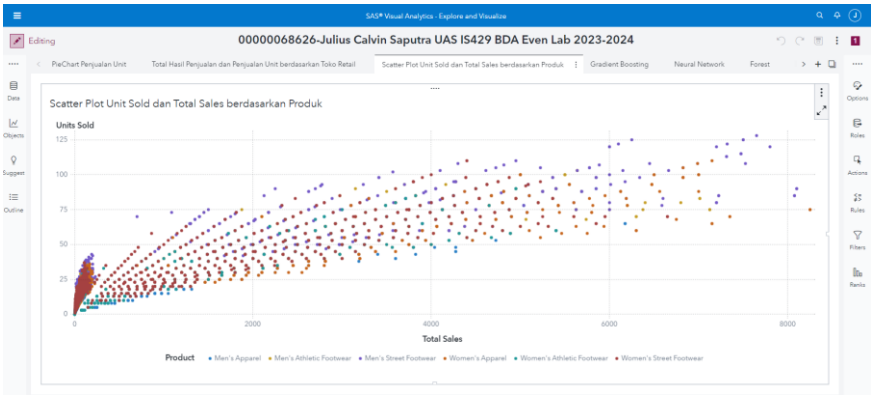
Hasil dari piechart menunjukan penjualan terbanyak dipegang oleh produk Men's Street Footwear dengan 57.737 barang yang terjual, diikuti oleh Men's athletic Footwear sebanyak 42.429 terjual, Women's apparel sebanyak 42.356 terjual dan seterusnya. Penjualan yang paling sedikit dimiliki oleh Men's apparel yang hanya terdapat 30.085 unit yang terjual. Dapat disimpulkan bahwa produk Sepatu pria dan pakaian Wanita merupakan produk yang paling banyak terjual.

Scatterplot



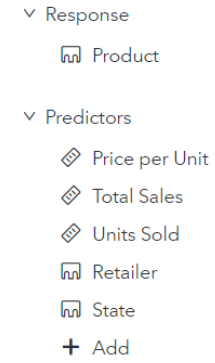
Untuk Scatterplot saya gunakan Unit Sold dan Total Sales sebagai measures dan Produk sebagai Color (Yang menentukan warna pada setiap plot).

Hasil



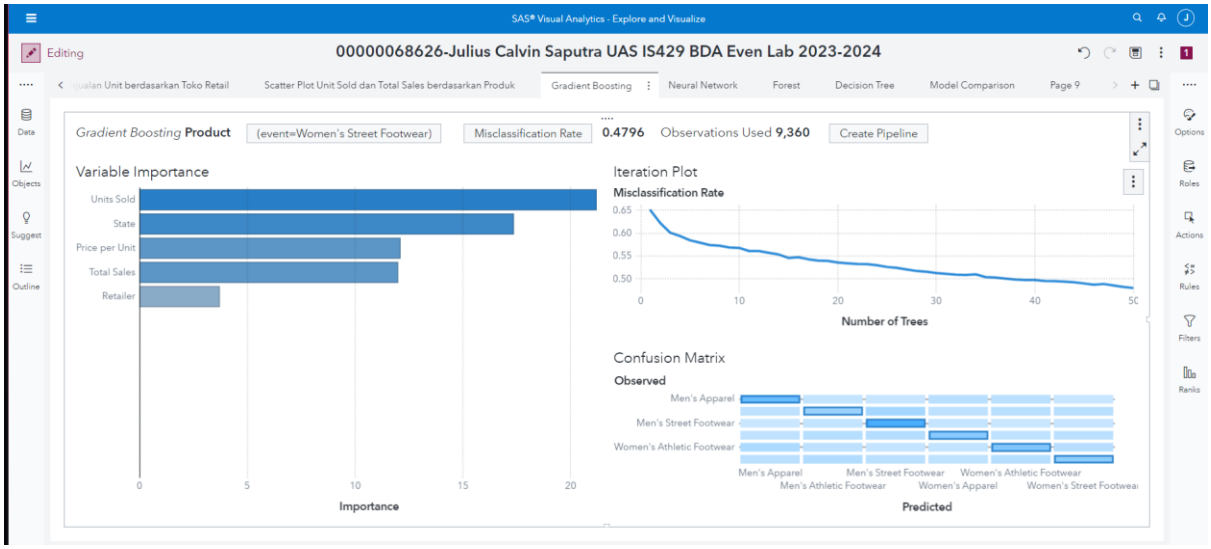
Scatterplot menunjukan visualisasi yang memberikan wawasan utama tentang kategori produk, yang masing-masing dibedakan per produk dengan warna Pakaian Pria (biru), Sepatu Atletik Pria (kuning), Sepatu Jalan Pria (ungu), Pakaian Wanita (merah), Sepatu Atletik Wanita (hijau), dan Sepatu Jalan Wanita (oranye). Tren umum menunjukkan korelasi positif antara total penjualan dan unit yang terjual, menunjukkan bahwa seiring meningkatnya total penjualan, jumlah unit yang terjual juga meningkat. Banyak titik data terkumpul di ujung bawah kedua sumbu, yang menunjukkan bahwa sebagian besar produk memiliki total penjualan dan unit yang terjual lebih rendah. Namun, terdapat beberapa outlier dengan penjualan dan unit terjual yang sangat tinggi. Secara khusus, Sepatu Jalan Pria (ungu) menunjukkan beberapa total penjualan tertinggi, dengan banyak titik yang tersebar di ujung kanan plot.

Modeling
Gradient Boosting



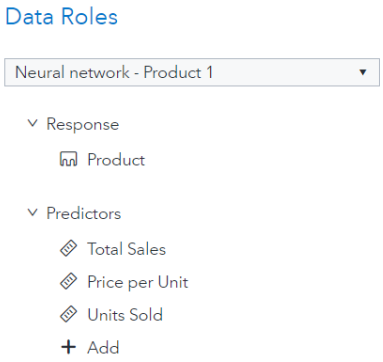
Pada model Gradient Boosting, saya menggunakan Product sebagai response dan Price per Unit, Total Sales, Unit Sold, Retailer, dan State sebagai predictors.

Hasil



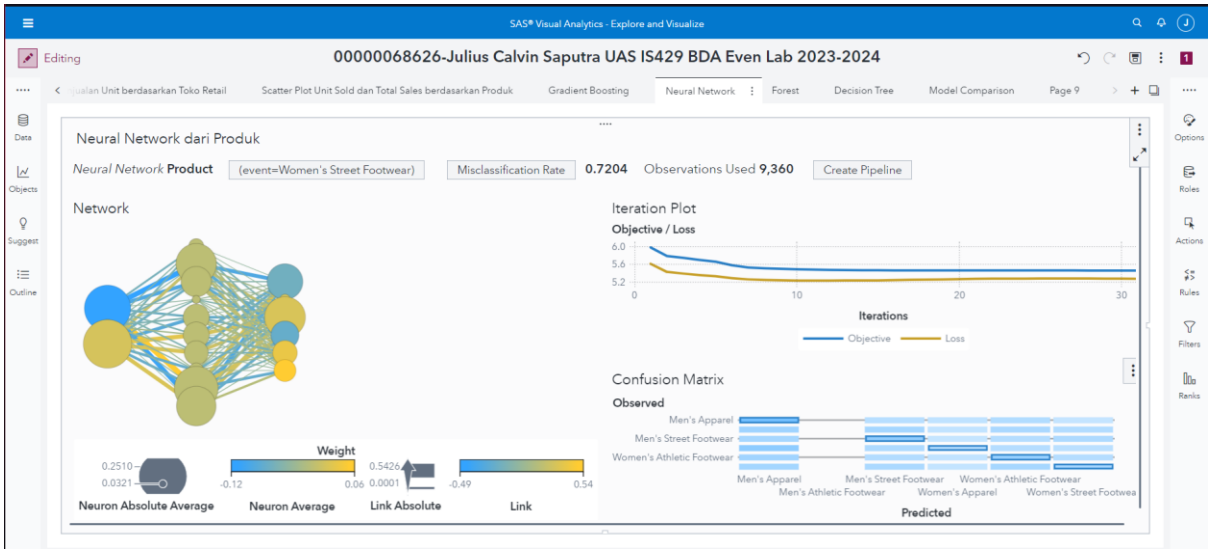
Pemodelan Gradient Boostin dalam analisis ini, model Gradient Boosting mencapai akurasi sebesar 52,04%, dengan Tingkat Kesalahan Klasifikasi sebesar 0,4796. Grafik Variabel Penting mengungkapkan bahwa "Unit Terjual" memainkan peran paling signifikan dalam prediksi model, diikuti oleh "State," "PRice per Unit," "Total Sales," dan "Retailer." Ini menekankan pentingnya volume penjualan dalam memengaruhi proses pengambilan keputusan model. Selain itu, Plot Iterasi menunjukkan bahwa peningkatan jumlah pohon dalam model meningkatkan akurasi, meskipun dengan pengembalian yang berkurang setelah mencapai titik tertentu, sekitar 50 pohon. Terakhir, Matriks Konfusi memberikan wawasan tentang performa model dalam berbagai kategori produk, dimana memperlihatkan area-area akurasi dan area-area yang memerlukan perbaikan, seperti mengklasifikasikan Pakaian Pria dengan benar tetapi kadang-kadang salah mengklasifikasikannya sebagai Sepatu Olahraga Pria.

Neural Network



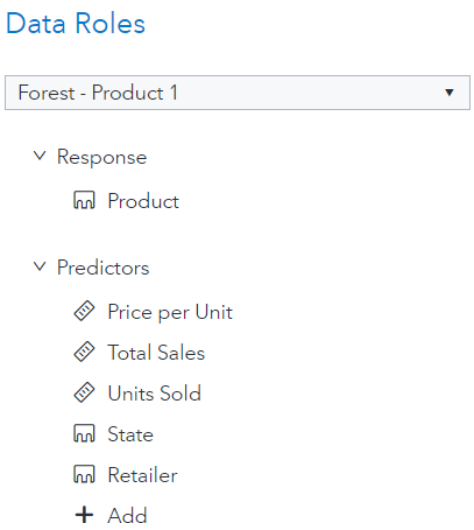
Untuk pemodelan Neural Network, saya menggunakan Product sebagai response dan Total Sales, Price per Unit, dan Unit Sold sebagai predictornya.

Hasil



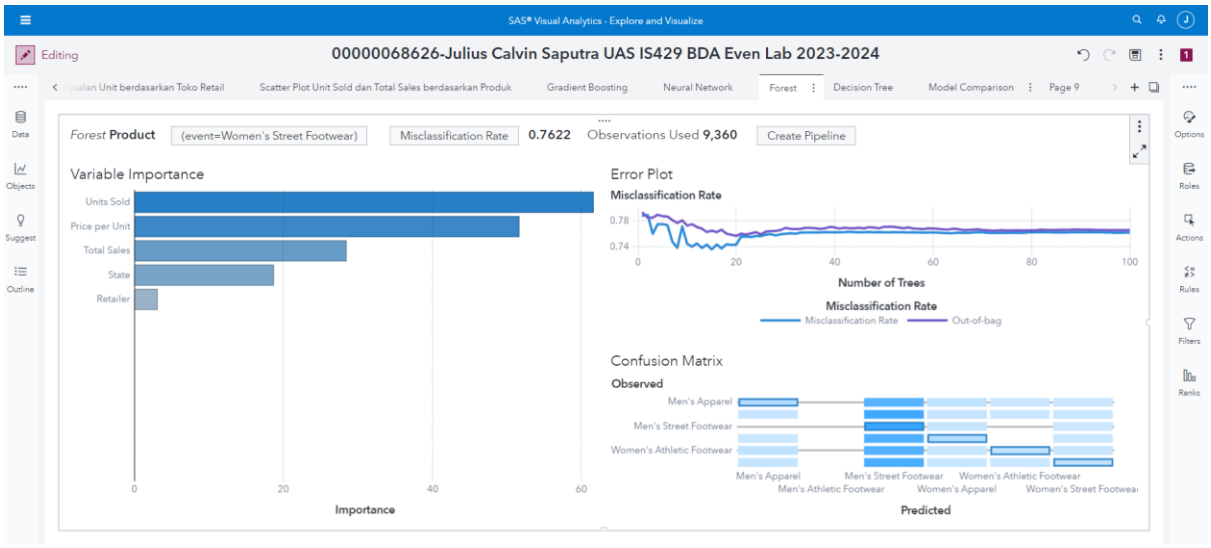
Hasil pemodelan Neural Network menampilkan struktur, dengan ukuran dan warna neuron serta koneksi (link) yang menunjukkan bobot dan signifikansinya. Neuron dan koneksi yang lebih besar dan lebih cerah menunjukkan pentingnya dan pengaruh yang lebih kuat pada prediksi model. Plot iterasi menggambarkan perubahan fungsi objektif/kerugian selama iterasi. Garis biru mewakili tujuan model, kemungkinan ukuran akurasi prediksi atau kesalahan, sedangkan garis kuning mewakili kesalahan dalam prediksi model. Selama sekitar 30 iterasi, baik tujuan maupun kerugian menurun, menunjukkan bahwa model ini belajar dan meningkatkan kinerjanya. Matriks konfusi (Confusion Matrix) membandingkan kategori yang diprediksi dengan kategori yang sebenarnya, memperlihatkan seberapa baik model mengklasifikasikan berbagai jenis produk. Nilai positif menunjukkan korelasi positif, sementara nilai negatif menunjukkan korelasi negatif. Model Neural Network juga menunjukan Misclassification Rate sebesar 0.7204, dimana menunjukan akurasi model sebesar 27.96%.

Forest



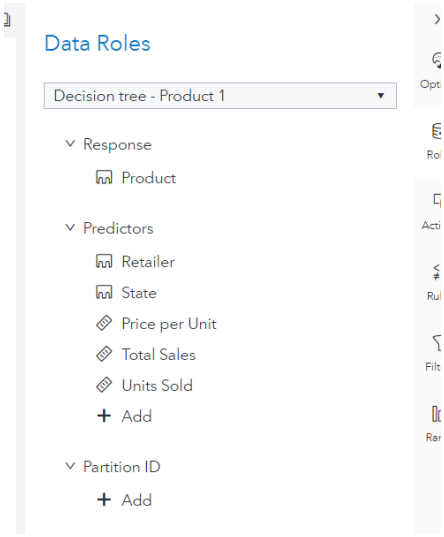
Pada model Forest, saya menggunakan Product sebagai response dan Price per Unit, Total Sales, Unit Sold, Retailer, dan State sebagai predictors.

Hasil



Hasil pemodelan Forest mencapai tingkat kesalahan klasifikasi (Misclassification Rate) sebesar 0.7622, yang berarti tingkat akurasinya sebesar 23.78%. Diagram pentingnya variabel menunjukkan bahwa "Units Sold" dan "Price per Unit" adalah fitur paling penting dalam menentukan prediksi model, diikuti oleh "Total Sales", "State", dan "Retailer". Plot kesalahan menggambarkan perubahan tingkat kesalahan klasifikasi saat jumlah pohon dalam model meningkat. Garis biru mewakili tingkat kesalahan klasifikasi, sedangkan garis ungu mewakili kesalahan out-of-bag, dengan keduanya menunjukkan penurunan seiring bertambahnya jumlah pohon, yang berarti model belajar dan meningkatkan kinerjanya

Decision Tree

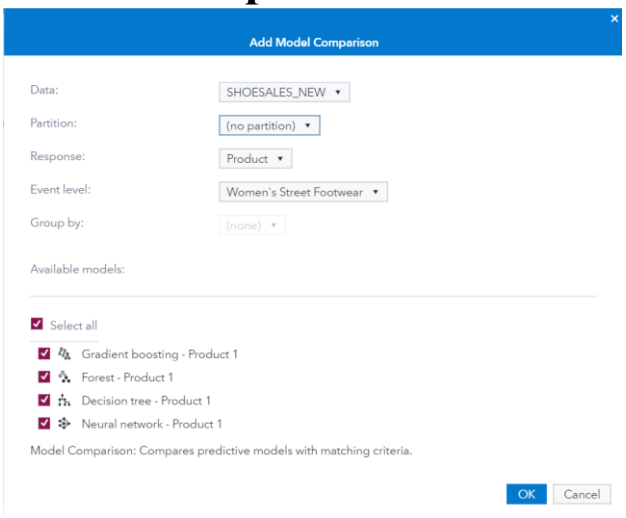


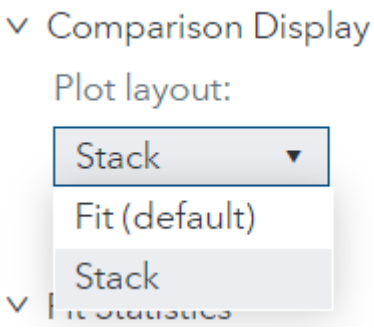
Pada model Decision Tree, saya menggunakan Product sebagai response dan Price per Unit, Total Sales, Unit Sold, Retailer, dan State sebagai predictors.

Hasil

Hasil pemodelan dengan Decision Tree mempunyai hasil dengan tingkat kesalahan klasifikasi (Misclassification Rate) sebesar 0.6752, yang berarti akurasinya 32.48%. Pohon keputusan ini memulai pemisahan dengan "Units Sold", diikuti oleh "Price per Unit", "State", "Total Sales", dan "Retailer". Diagram pentingnya variabel menegaskan bahwa "Units Sold" dan "Price per Unit" adalah fitur utama dalam prediksi, diikuti oleh "State", "Total Sales", dan "Retailer". Visualisasi lift menunjukkan kinerja model dalam membedakan kategori produk, dengan cumulative lift yang tinggi pada percentil atas, namun menurun pada percentil bawah.

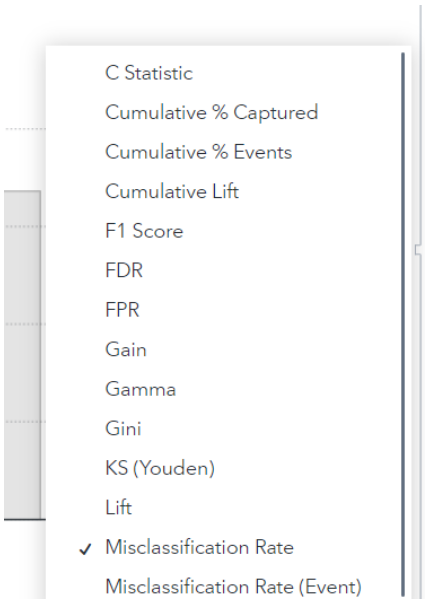
Model Comparison





Saya select Semua Model Comparison dan memilih "Stack" sebagai Plot Layout, Dimana ini akan memudahkan kita untuk melihat visualisasi dari setiap Model Comparison satu per satu.

Fit Statistic (Misclassification Rate)

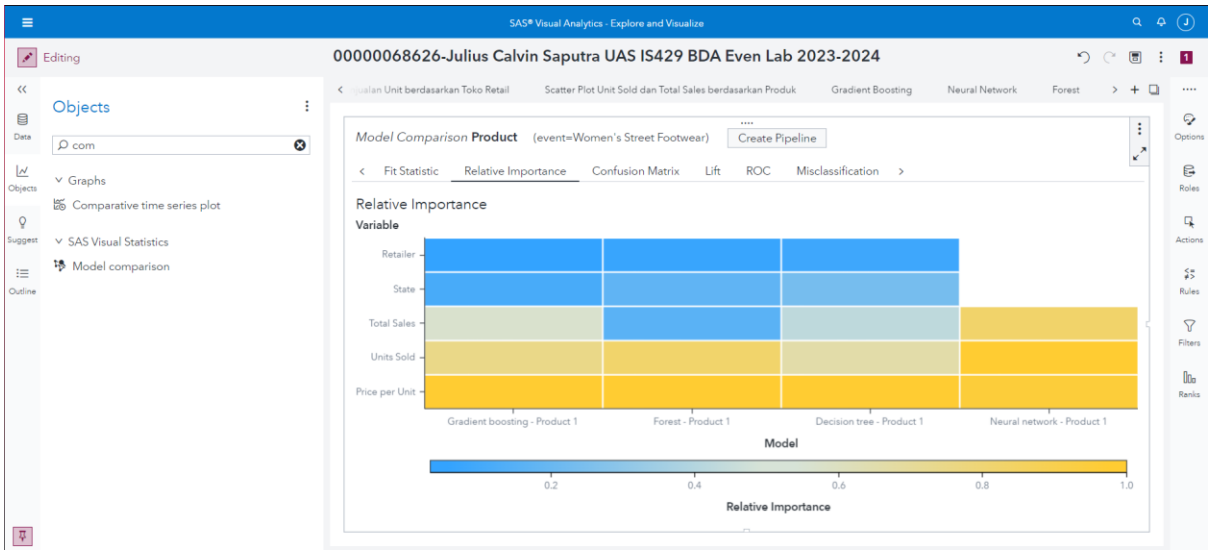


Memilih Misclassification Rate (Default berupa KS (Youden).

Hasil

Berdasarkan hasil Fit Statistic yang ditunjukan Gambar 16 untuk Misclassification Rate, model dengan tingkat kesalahan klasifikasi (Misclassification) terendah merupakan Gradient Boosting dengan nilai 0.4796. Ini berarti model tersebut memiliki tingkat akurasi tertinggi, yaitu sebesar 52,04%. Oleh karena itu, hasil dari perbandingan model ini menunjukkan bahwa Gradient Boosting adalah model yang terpilih. Tingkat kesalahan klasifikasi yang lebih kecil mempunyai arti Tingkat akurasi yang lebih besar, dan juga bekerja sebaliknya.

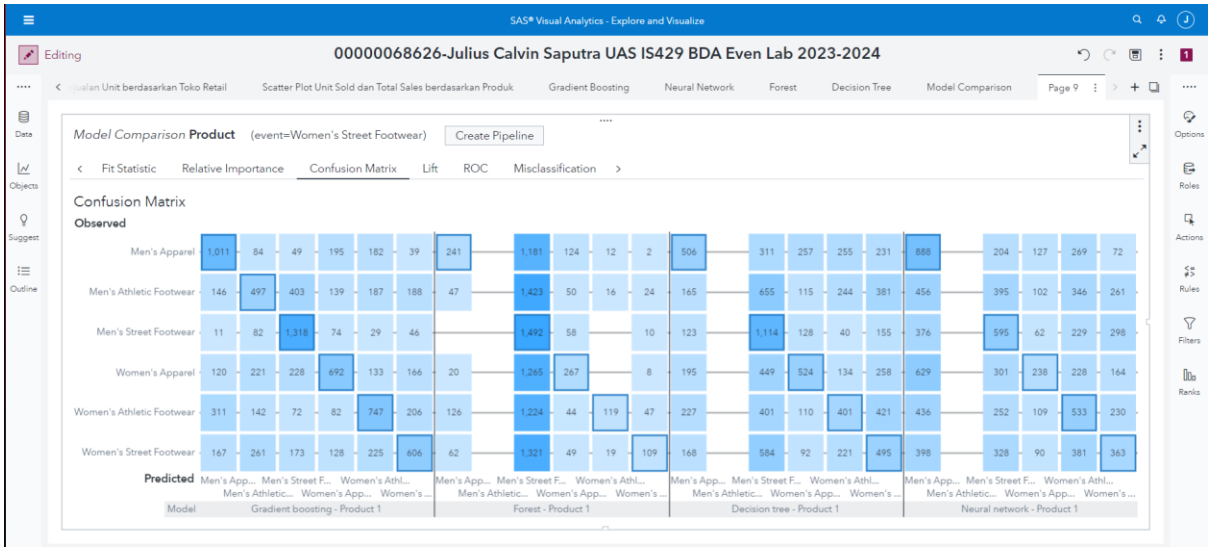
Relative Importance
Hasil



Matriks Relative Importance, dimana warna biru menunjukan dekat ke angka 0 (Rendah) dan warna kuning menunjukan dekat ke angka 1 (Tinggi). Dapat dilihat bahwa variable yang menunjukan paling banyak warna kuning mempunyai arti variable yang paling penting. Sehingga variable paling penting adalah “Price per Unit” Category yang dimana memiliki korelasi tinggi dengan Gradient Boosting, Forest, Decision Tree, dan Neural Network.

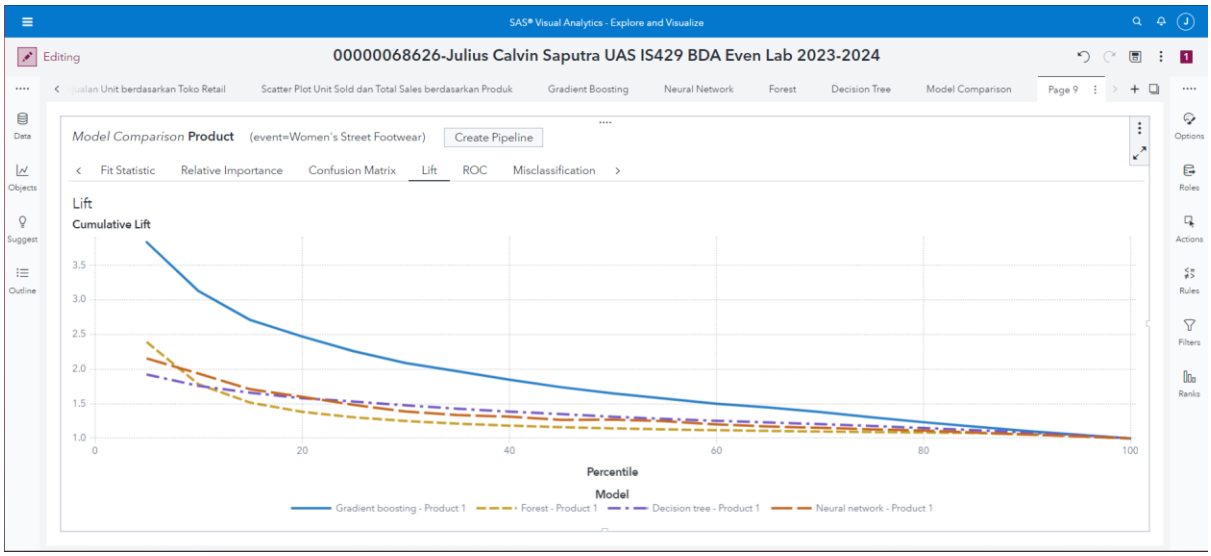
Confusion Matrix

Hasil



Confusion Matrix perbandingan kelas aktual dan prediksi di berbagai model, yang berfokus pada empat kategori: "Men's Apparel", "Men's Athletic Footwear", "Women's Apparel", dan "Women's Athletic Footwear". Baris menunjukkan kelas aktual, sementara kolom menunjukkan prediksi yang dibuat oleh model Gradient Boosting, Random Forest, Decision Tree, dan Neural Network. Setiap sel dalam matriks menunjukkan jumlah observasi yang diprediksi berada dalam kelas tertentu oleh model tertentu. Ssel untuk "Men's Apparel"di bawah Gradient Boosting menunjukkan 348 prediksi yang benar dan 47 instance yang salah diklasifikasikan sebagai "Men's Athletic Footwear". Sel diagonal dari kiri atas ke kanan bawah menyoroti prediksi benar di mana kelas yang diamati sesuai dengan kelas yang diprediksi. Sebaliknya, sel off-diagonal menunjukkan mis-klasifikasi, yang mencerminkan ketidakcocokan antara kelas aktual dan kelas yang diprediksi.

Lift

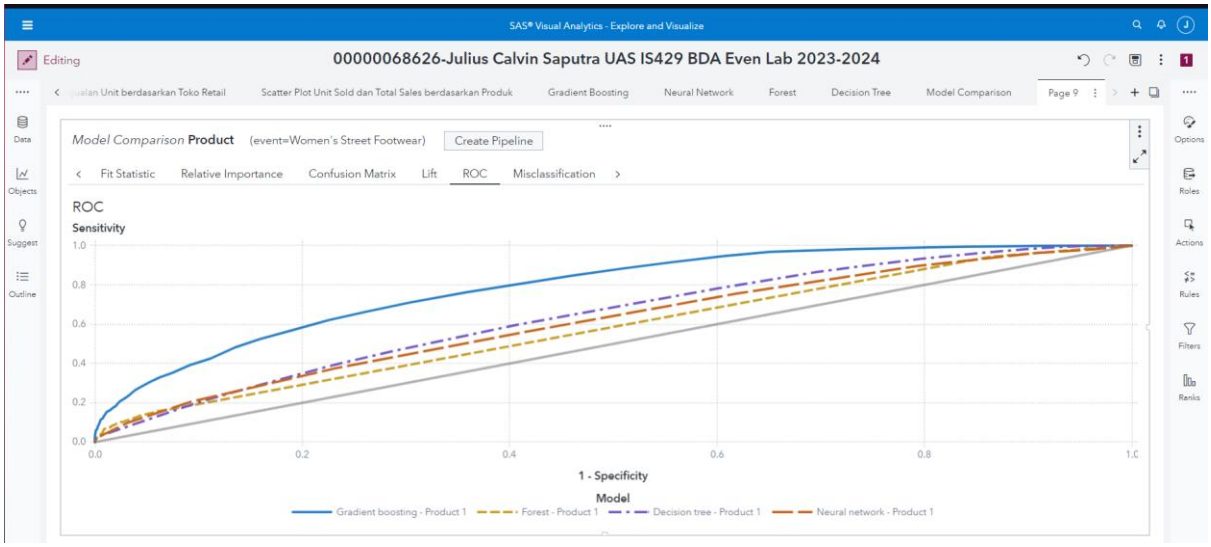


Hasil Lift menunjukan dimana warna biru menunjukan Model Gradient Boosting, warna kuning menunjukan Forest, warna ungu menunjukan Decision Tree, dan warna oranye menunjukan Neural Network. Dapat dilihat bahwa warna biru (Gradient Boosting) mempunyai nilai cumulative lift yang

tertinggi. Nilai cumulative lift yang tinggi menunjukkan bahwa model tersebut secara konsisten mampu memberikan prediksi yang lebih akurat di seluruh percentil data

Kurva ROC

Hasil



Kurva ROC menghasilkan warna biru yang menunjukan Model Gradient Boosting, warna kuning menunjukan Forest, warna ungu menunjukan Decision Tree, dan warna oranye menunjukan Neural Network. Dapat dilihat bahwa warna biru (Gradient Boosting) mempunyai nilai ROC yang paling tinggi. Berarti penggunaan model Gradient Boosting memiliki tingkat True Positive yang tinggi dan False Positive yang rendah pada berbagai threshold. Ini berarti model tersebut lebih efektif dalam mengidentifikasi kejadian yang benar (positif) sambil meminimalkan kesalahan identifikasi (false positives).