Name : Ray Anthony Pranoto

NIM: 00000066655

Class: IS429-DL

**REPOSITORY DATA**

|  |  |
| --- | --- |
| Sumber Dataset | Dari diambil dari Kaggle  <https://www.kaggle.com/datasets/ahmedabbas757/dataset> |
| Deskripsi Dataset | Dataset ini menyajikan informasi yang komprehensif tentang penjualan produk dari berbagai pengecer, wilayah geografis, dan metode penjualan. Setiap entri dalam dataset ini mencakup detail penting seperti tanggal faktur, nama pengecer, lokasi penjualan (baik negara bagian maupun kota), jenis produk yang dijual, harga per unit, jumlah unit yang terjual, total penjualan, serta keuntungan operasional yang dihasilkan. Selain itu, dataset ini juga memberikan informasi tentang metode penjualan yang digunakan, seperti penjualan di toko fisik (In-store) atau melalui platform online (Online), serta tingkat penjualan (Sales\_Level) yang mungkin mencerminkan tingkat permintaan atau popularitas produk tersebut. Data ini dapat digunakan untuk menganalisis kinerja penjualan dari waktu ke waktu, mengidentifikasi tren penjualan di berbagai wilayah atau metode penjualan, memahami preferensi konsumen terhadap produk tertentu berdasarkan lokasi atau metode penjualan, serta mengevaluasi keuntungan operasional yang dihasilkan dari penjualan produk. Dengan informasi yang terperinci ini, stakeholder bisnis dapat membuat keputusan yang lebih baik dalam hal pemasaran, penetapan harga, strategi distribusi, dan manajemen operasional untuk meningkatkan kinerja penjualan dan profitabilitas. |
| Format Dataset | CSV |

Berikut rincian lebih detail mengenai kolom, tipe data, dan pengertian setiap kolom – kolomnya:

Kolom:

|  |  |  |  |
| --- | --- | --- | --- |
| No | Nama Variable | Tipe Data | Deskripsi |
| 1. | Retailer | Varchar | Nama dari retailer yang melakukan penjualan. |
| 2. | Retailer ID | Varchar | ID unik yang mengidentifikasi setiap retailer. |
| 3. | Invoice Date | Varchar | Tanggal terjadinya faktur atau transaksi penjualan tertentu. |
| 4. | Region | Varchar | Wilayah geografis tempat penjualan terjadi (misalnya, Southeast, South, Northeast, West, Midwest). |
| 5. | State | Varchar | Nama negara bagian tempat penjualan dilakukan. |
| 6. | City | Varchar | Nama kota di mana penjualan terjadi. |
| 7. | Product | Varchar | Jenis produk yang terjual (misalnya, Women's Apparel, Men's Street Footwear). |
| 8. | Total Sales | Double | Total penjualan dalam dolar, dihitung dengan mengalikan harga per unit dengan jumlah unit terjual. |
| 9. | Sales Method | Varchar | Metode penjualan yang digunakan (misalnya, Online, Outlet, In-store). |
| 10. | Operating Profit | Double | Keuntungan yang diperoleh pengecer dari operasi bisnis normalnya. |
| 11. | Price per Unit | Double | Biaya atau harga yang terkait dengan satu unit produk. |
| 12. | Units Sold | Double | Kuantitas atau jumlah unit produk tertentu yang terjual selama transaksi penjualan tertentu. |

Link SAS Visualisasi + Modelling : <https://v4e039.vfe.sas.com/links/resources/report?uri=%2Freports%2Freports%2Fedefcf14-724d-465e-ba5a-582e82e635f8&page=vi9739>

1. Question 1: CLO021 Sub-CLO-6 and CLO013 Sub-CLO-7 Weight (40%)

Create your EDA and Data preparation process with the SAS® Data Preparation Application.

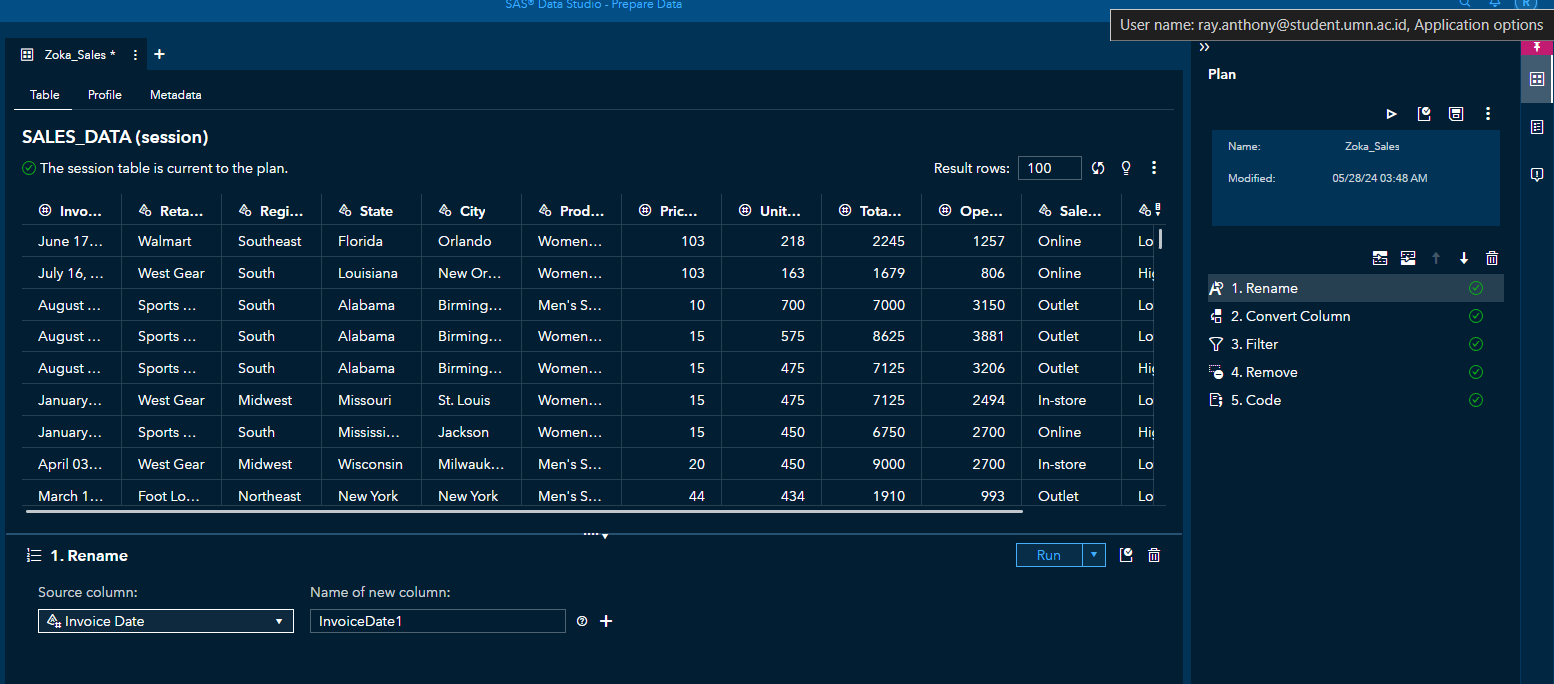




**Gambar 1**. Dataset Zoka

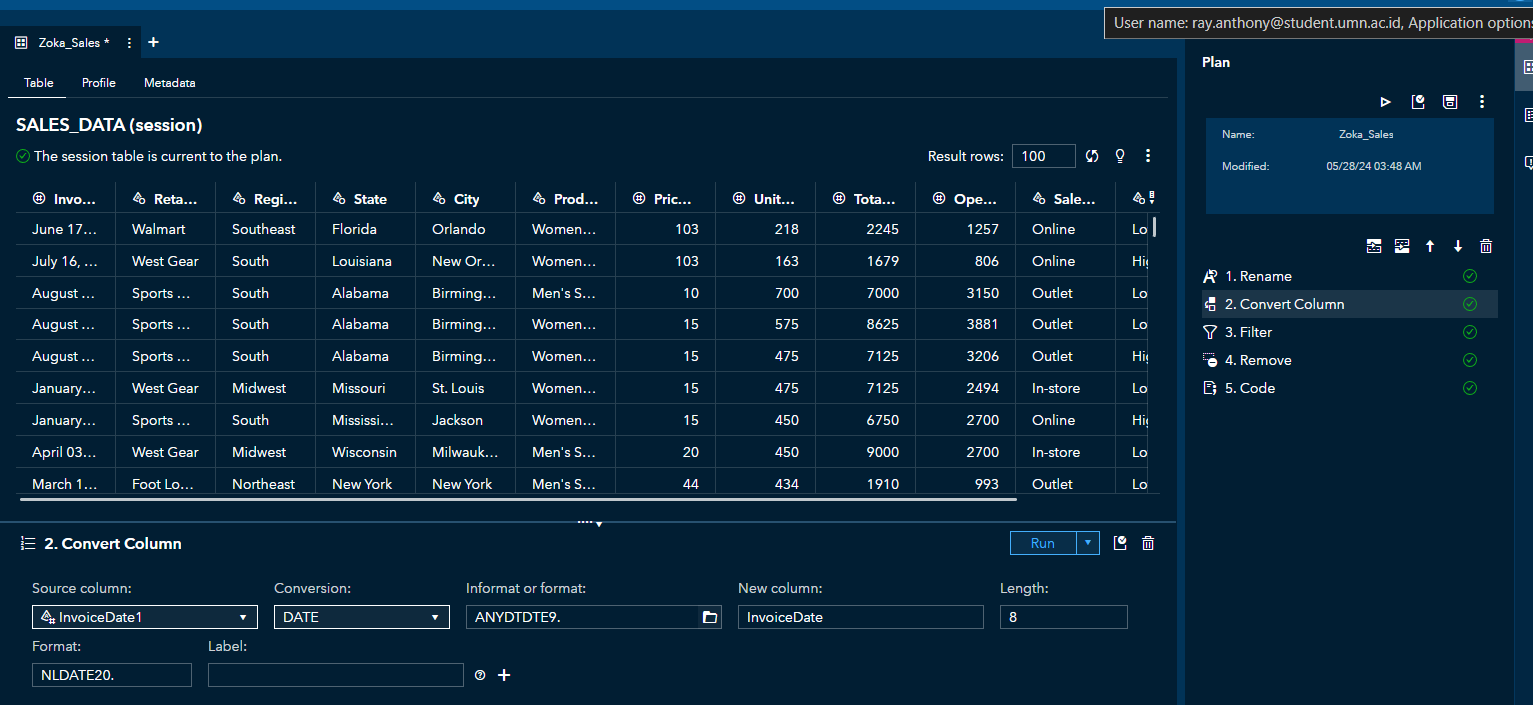
Gambar 1 merupakan dataset dari Zoka Sport Factory Outlet, Zoka menyediakan berbagai produk olahraga mulai dari sepatu, pakaian, hingga alat olahraga dengan harga yang kompetitif. Data penjualan Zoka yang dianalisis mencakup informasi penting seperti jenis produk olahraga, harga, jumlah unit terjual, dan total penjualan. Dengan menganalisis data ini, Zoka dapat menentukan produk mana yang paling laris, memahami tren musiman, dan mengidentifikasi area yang memerlukan perbaikan atau penyesuaian dalam strategi pemasaran mereka. Selain itu, visualisasi data yang dihasilkan membantu Zoka untuk membuat laporan yang informatif dan mudah dipahami, mendukung pengambilan keputusan yang lebih baik. Untuk mencapai semua ini, digunakan alat seperti SAS Visual Studio dan SAS Visual Analytics. SAS Visual Studio memungkinkan pemrosesan data yang efisien dan analisis yang mendalam, sementara SAS Visual Analytics membantu dalam menghasilkan visualisasi data yang interaktif dan intuitif. Melalui penggunaan alat-alat ini, Zoka dapat mengoptimalkan strategi penjualan dan memperkuat posisinya di pasar retail olahraga yang kompetitif.

**Data Preparation:**

****

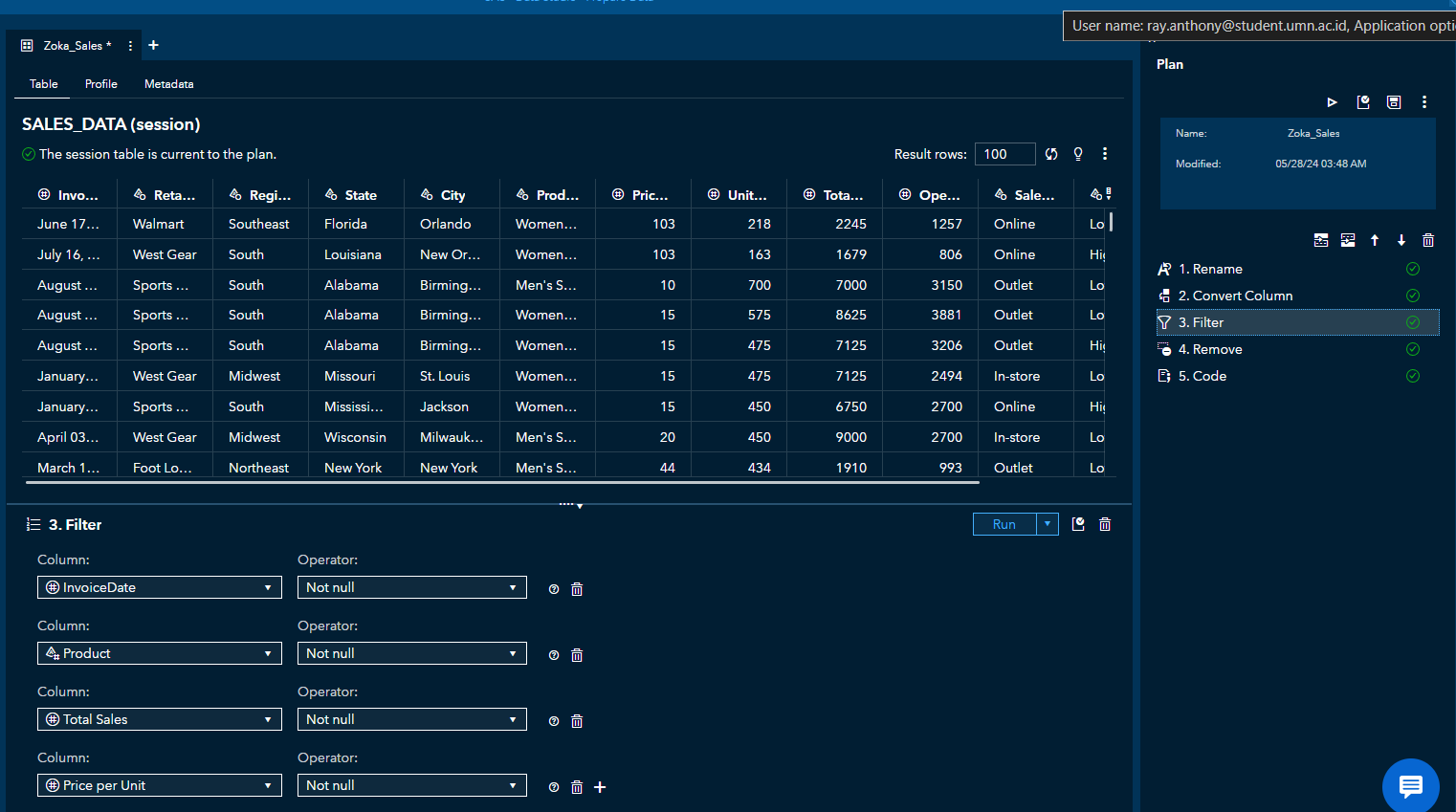
**Gambar 2**. Mengganti Nama Kolom

Pada Gambar 2 di atas, merupakan langkah pertama dalam persiapan data untuk mengubah nama kolom yang sesuai dengan data. Proses pengubahan nama kolom 'Invoice Date' menjadi 'InvoiceDate1' dilakukan. Tujuannya adalah agar proses persiapan data selanjutnya dapat dilakukan tanpa mengkhawatirkan kemungkinan kesalahan karena adanya spasi di antara kata-kata.



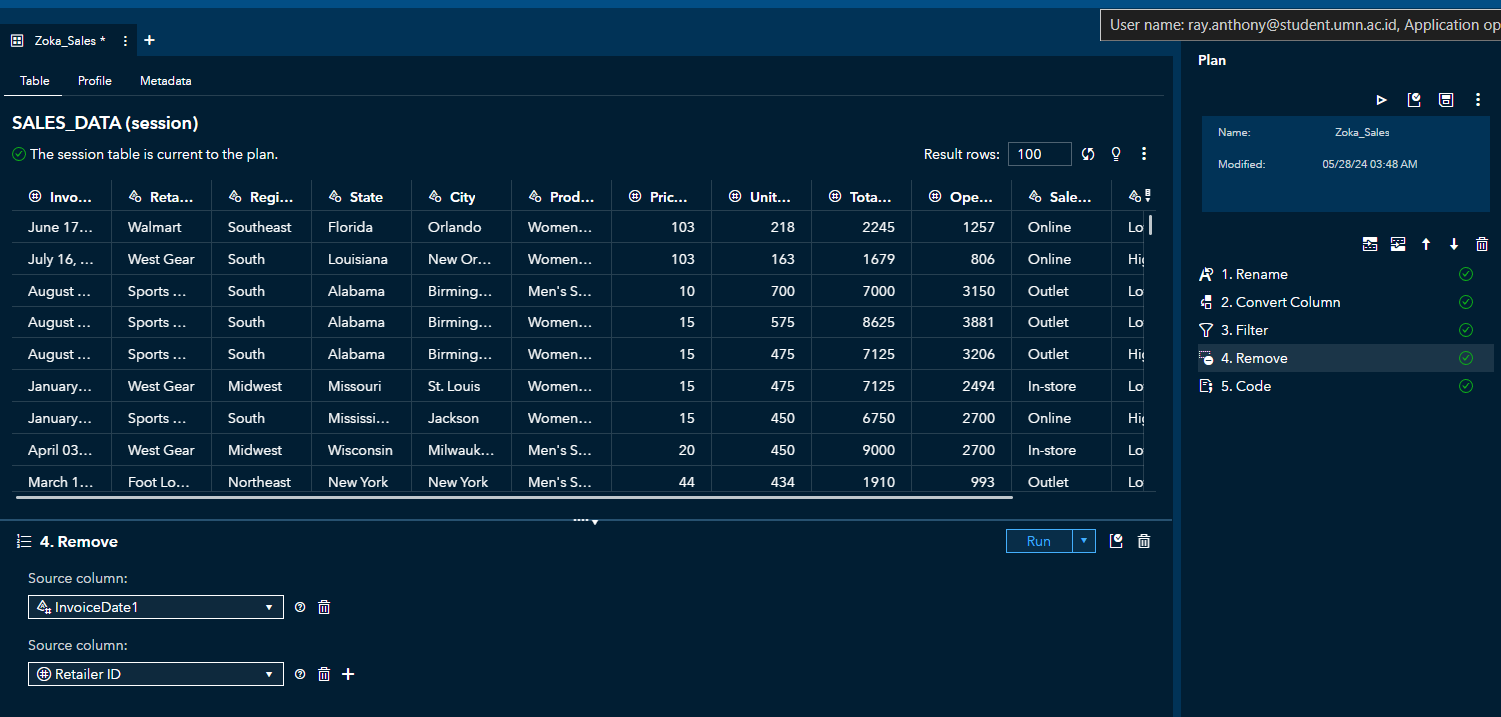
**Gambar 3**. Konversi Kolom

Gambar 3 di atas menunjukkan proses konversi kolom. Konversi kolom ini berguna untuk mengubah tipe data yang tidak sesuai dengan nilainya agar dapat digunakan dalam proses selanjutnya. Proses tersebut mengubah tipe data pada kolom 'InvoiceDate1' dari varchar menjadi Date. Format tanggal yang digunakan adalah ANYDTDTE9., yang mampu mengenali dan memproses berbagai gaya penulisan tanggal seperti "June 17, 2021", "17JUN2021", "06/17/2021", dan sebagainya. Hasil dari konversi tersebut kemudian disimpan dalam kolom 'InvoiceDate' untuk digunakan dalam langkah-langkah berikutnya.



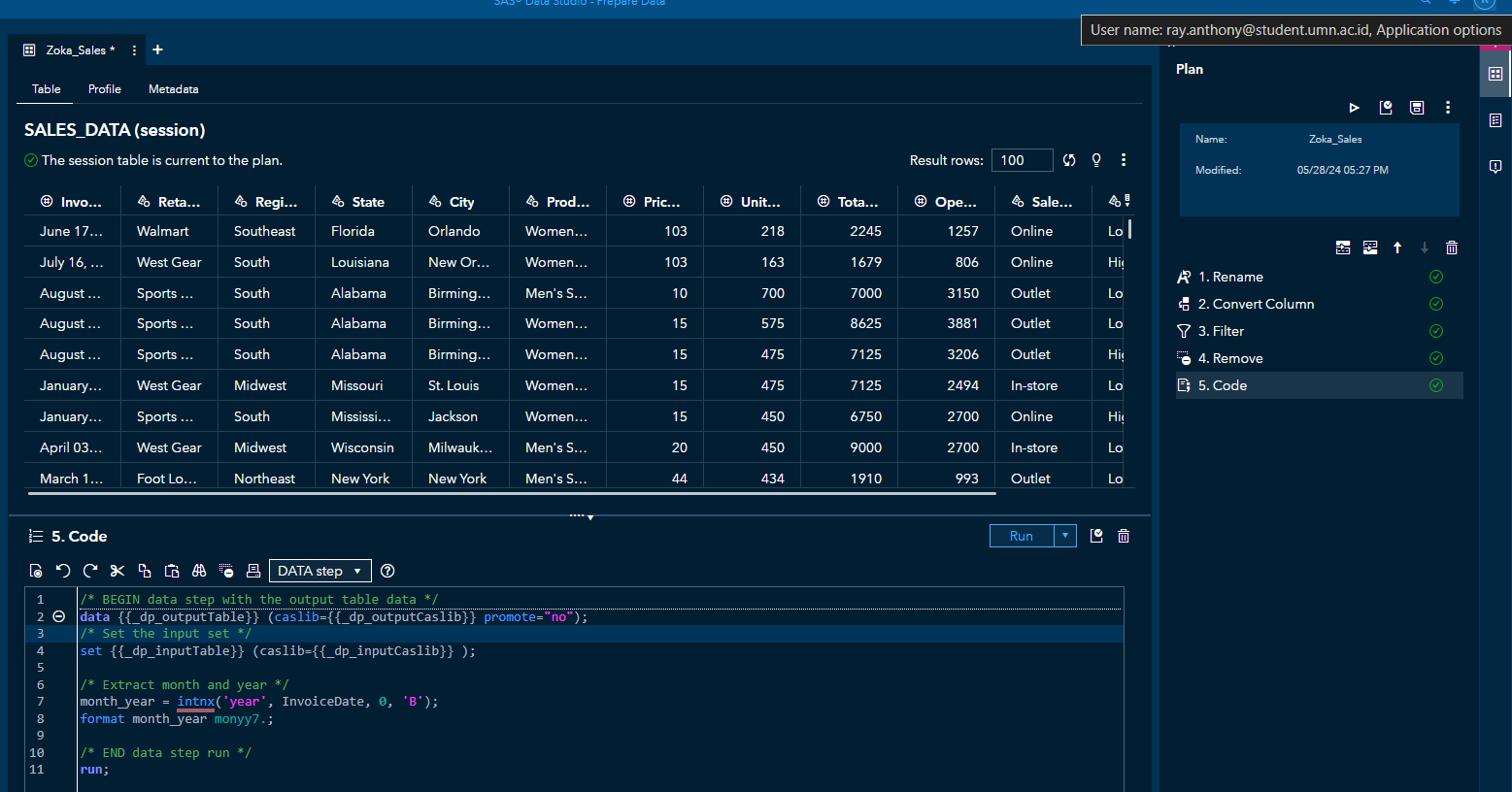
**Gambar 4**. Filter Kolom

Gambar 4 di atas menunjukkan proses penyaringan pada data penjualan Zoka, yang bertujuan untuk menghilangkan seluruh nilai yang hilang (missing values) pada masing-masing kolom. Proses ini memastikan integritas dan kualitas data yang akan digunakan dalam analisis selanjutnya. Kolom-kolom yang ditampilkan pada proses penyaringan meliputi ‘InvoiceDate’ (hasil dari proses pra-pengolahan yang melibatkan penggantian nama dan konversi kolom), ‘Product’ (bertipe Varchar), ‘Total Sales’ (bertipe Double), dan ‘Price per Unit’ (bertipe Double). Dari gambar tersebut, terlihat bahwa proses penghapusan nilai null telah berhasil dilakukan, sehingga dataset Zoka kini bebas dari nilai-nilai null dan siap digunakan untuk langkah-langkah analisis selanjutnya. Ini membantu dalam meningkatkan kualitas dan keandalan hasil analisis yang akan diperoleh dari dataset tersebut.



**Gambar 5**. Menghilangkan Kolom

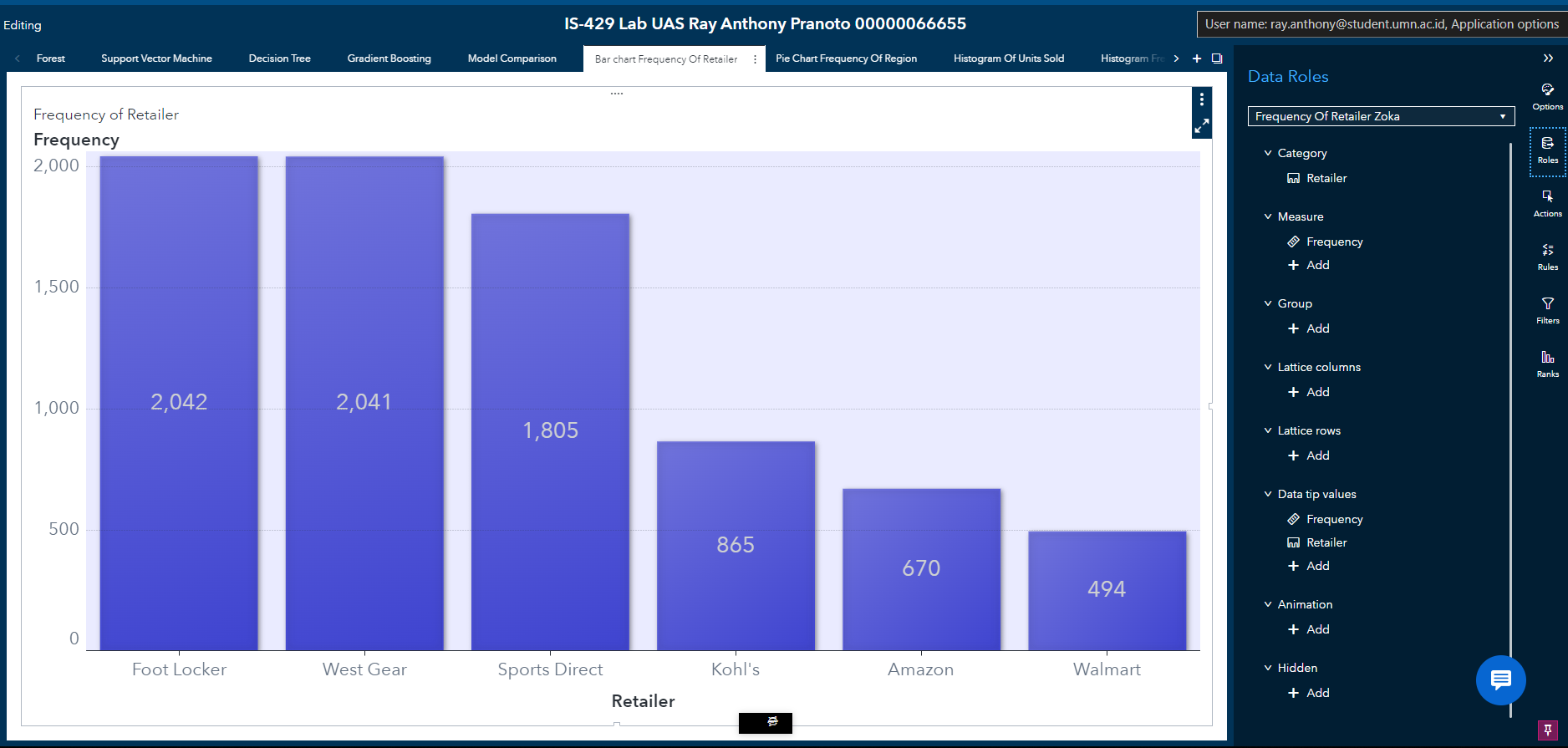
Gambar 5 di atas menunjukkan proses penghapusan kolom-kolom yang tidak diperlukan untuk analisis selanjutnya. Dalam gambar tersebut, kolom ‘InvoiceDate1’ dihapus karena merupakan duplikat dari kolom ‘InvoiceDate’ dan masih memiliki tipe data varchar. Selain itu, kolom ‘RetailerID’ juga dihapus karena tidak akan digunakan dalam proses analisis berikutnya. Penghapusan kolom-kolom ini bertujuan untuk menyederhanakan dataset dan memastikan bahwa hanya data yang relevan dan berkualitas tinggi yang digunakan dalam analisis lebih lanjut.



**Gambar 6**. Menambah Kolom Baru

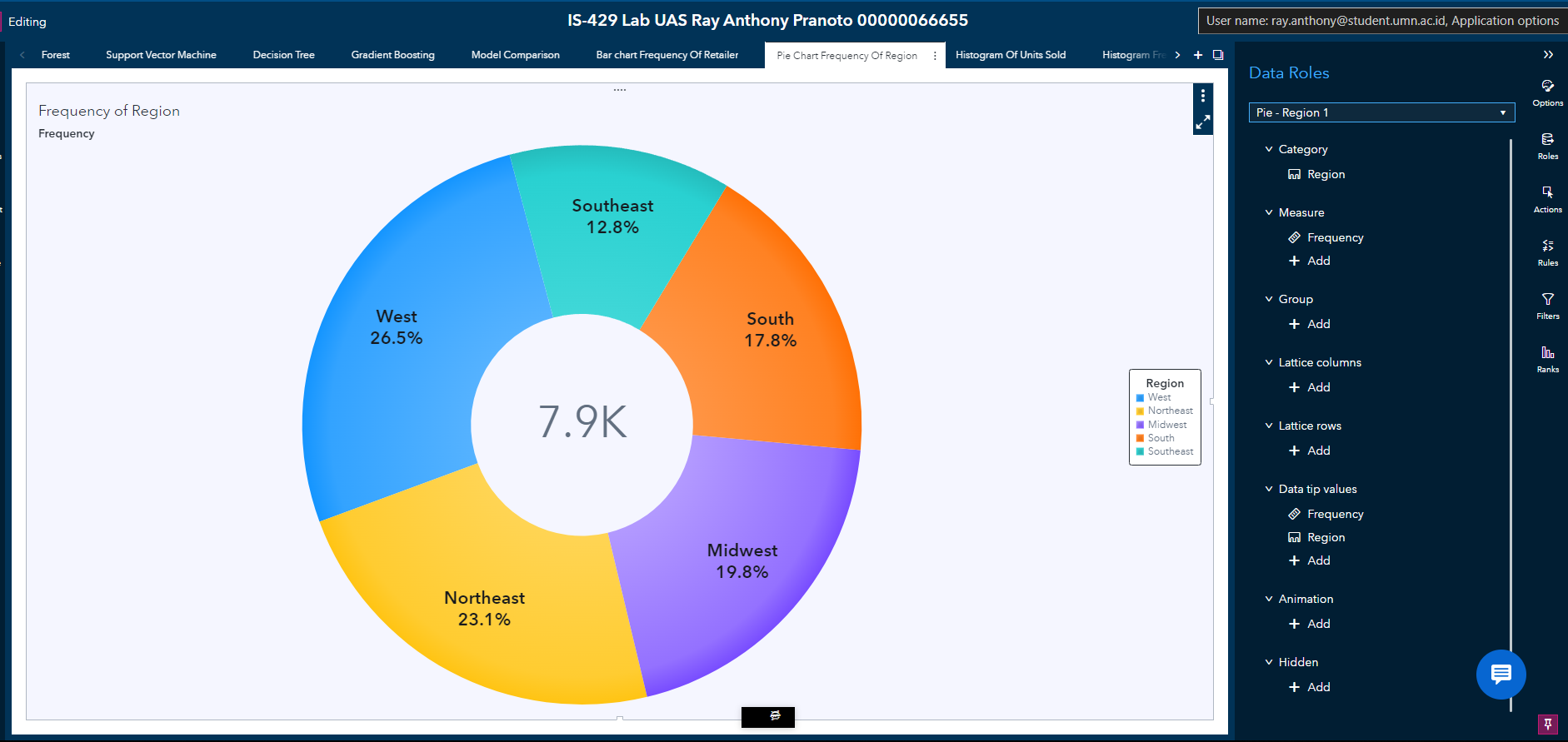
Gambar 6 diatas merupakan kode yang digunakan untuk memproses data dari tabel input dan menyimpannya ke tabel output. Pertama, tabel output `{{\_dp\_outputTable}}` dideklarasikan dalam caslib `{{\_dp\_outputCaslib}}` dengan opsi `promote="no"`. Selanjutnya, data dari tabel input `{{\_dp\_inputTable}}` yang berada dalam caslib `{{\_dp\_inputCaslib}}` dibaca untuk diproses. Kolom baru `month\_year` dibuat dengan menggunakan fungsi `intnx` untuk mengekstrak bulan dan tahun dari kolom `InvoiceDate`, di mana nilai `month\_year` diatur ke awal tahun dari `InvoiceDate`, dan format `monyy7.` diterapkan untuk menampilkan bulan dan tahun dalam format yang mudah dibaca. Perintah `run;` mengeksekusi data step ini, memproses data, dan menghasilkan tabel output dengan kolom tambahan `month\_year` yang berisi informasi bulan dan tahun dari setiap tanggal faktur.

**Exploration Process:**

****

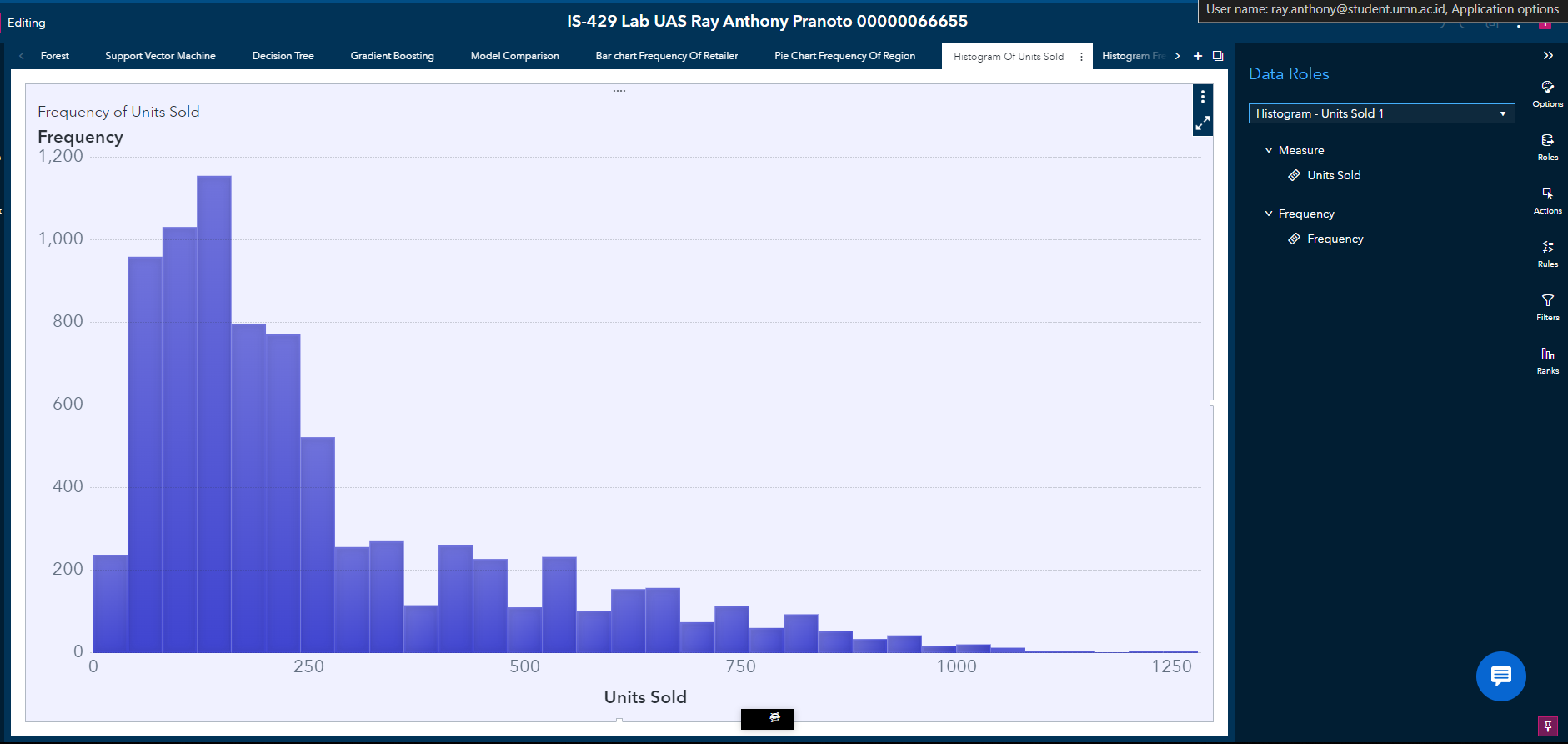
**Gambar 7**. Bar Chart Frekuensi Pengecer

Gambar 7 di atas merupakan visualisasi dalam bentuk barchart yang menghitung frekuensi pada kolom Retailer. Visualisasi ini berfungsi untuk mengetahui retailer mana yang paling sering melakukan transaksi pada dataset tersebut. Set Roles yang pada visualisasi adalah pada category diisi kolom ‘Retailer’, untuk bagian measure diisi dengan frequency dari ‘Retailer’, dan pada bagian data tip values diisi dengan frequency ‘Retailer’ dan kolom ‘retailer’ yang bertipe category. Berdasarkan data yang ditampilkan dalam barchart, Foot Locker adalah retailer yang paling sering melakukan pembelian, dengan total 2042 transaksi selama dua tahun terakhir. West Gear berada di posisi kedua dengan 2041 transaksi, diikuti oleh Sport Gear dengan 1805 transaksi, Kohl’s dengan 865 transaksi, Amazon dengan 670 transaksi, dan yang terakhir adalah Walmart dengan 494 transaksi. Visualisasi ini membantu dalam memahami pola transaksi di antara berbagai retailer.



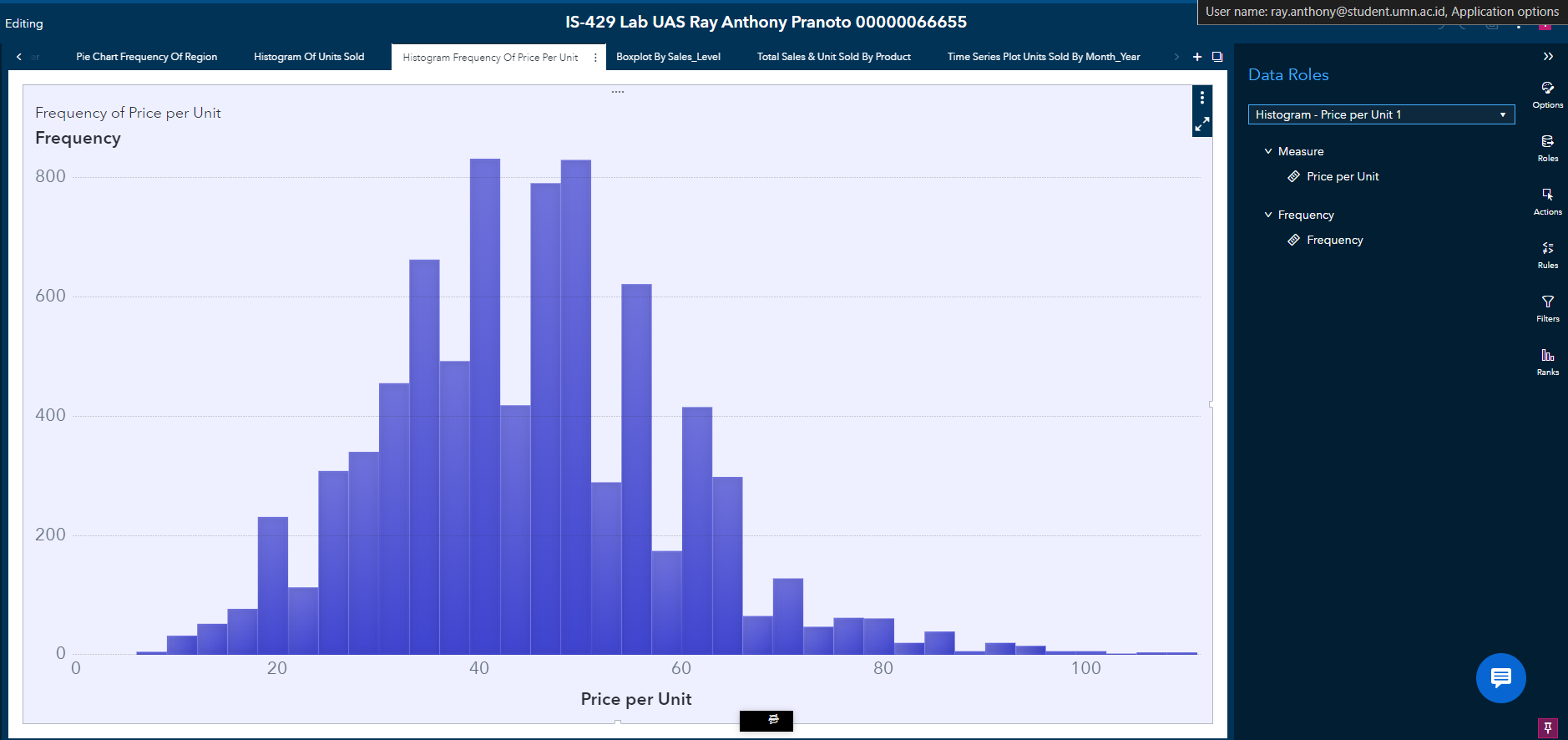
**Gambar 8**. Diagram Lingkaran Frekuensi Wilayah

Gambar 8 di atas merupakan diagram lingkaran (pie chart) yang menggambarkan frekuensi transaksi berdasarkan wilayah (Region) dalam dataset penjualan Zoka. Set roles pada pie chart ini adalah pada bagian category diisi dengan kolom ‘Region’, kemudian pada measure diisi dengan menghitung nilai frequency dari masing masing region, kemudian untuk bagian data tip values diisi dengan frequency dari ‘Region’ dan kolom ‘Region’. Diagram ini menunjukkan persentase kontribusi masing-masing wilayah terhadap total transaksi yang berjumlah 7.9K (7900 transaksi). Wilayah West memiliki frekuensi transaksi tertinggi dengan 26.5%, menunjukkan bahwa lebih dari seperempat transaksi berasal dari wilayah ini. Northeast mengikuti dengan 23.1% dari total transaksi, disusul oleh Midwest dengan 19.8%, South dengan 17.8%, dan Southeast dengan 12.8%. Visualisasi ini membantu dalam memahami distribusi geografis dari transaksi penjualan dan menunjukkan bahwa wilayah West dan Northeast merupakan kontributor terbesar dalam dataset ini. Dengan informasi ini, perusahaan dapat mengidentifikasi wilayah dengan aktivitas penjualan tertinggi dan merancang strategi pemasaran yang lebih efektif untuk setiap wilayah.



**Gambar 9**. Histogram Frekuensi Unit Terjual

Gambar 9 di atas merupakan visualisasi menggunakan diagram batang yang menampilkan distribusi frekuensi penjualan unit suatu produk. Set roles pada visualisasi tersebut adalah pada bagian measure diisi dengan kolom ‘Units Sold’ yang bertipe numeric dan pada frequency diisi dengan nilai frequency. Pada sumbu horizontal, terdapat label “Unit Terjual,” sedangkan pada sumbu vertikal, terdapat label “Frekuensi Unit Terjual.” Rentang sumbu horizontal berkisar antara 0 hingga 1250 dengan peningkatan sebesar 250, sementara sumbu vertikal berkisar antara 0 hingga 1200 dengan peningkatan sebesar 200. Batang-batang dalam diagram tersebut mewakili frekuensi penjualan untuk berbagai jumlah unit yang terjual. Dari visualisasi ini, distribusi miring ke kanan, sebagian besar titik data terkonsentrasi di sisi kiri kurva distribusi, dengan ekor yang lebih kecil memanjang ke kanan. Artinya, terdapat lebih banyak titik data dengan nilai lebih rendah dibandingkan titik data dengan nilai lebih tinggi. Visualisasi ini memberikan gambaran yang jelas tentang distribusi penjualan unit dan membantu mengidentifikasi jumlah unit yang paling sering terjual.



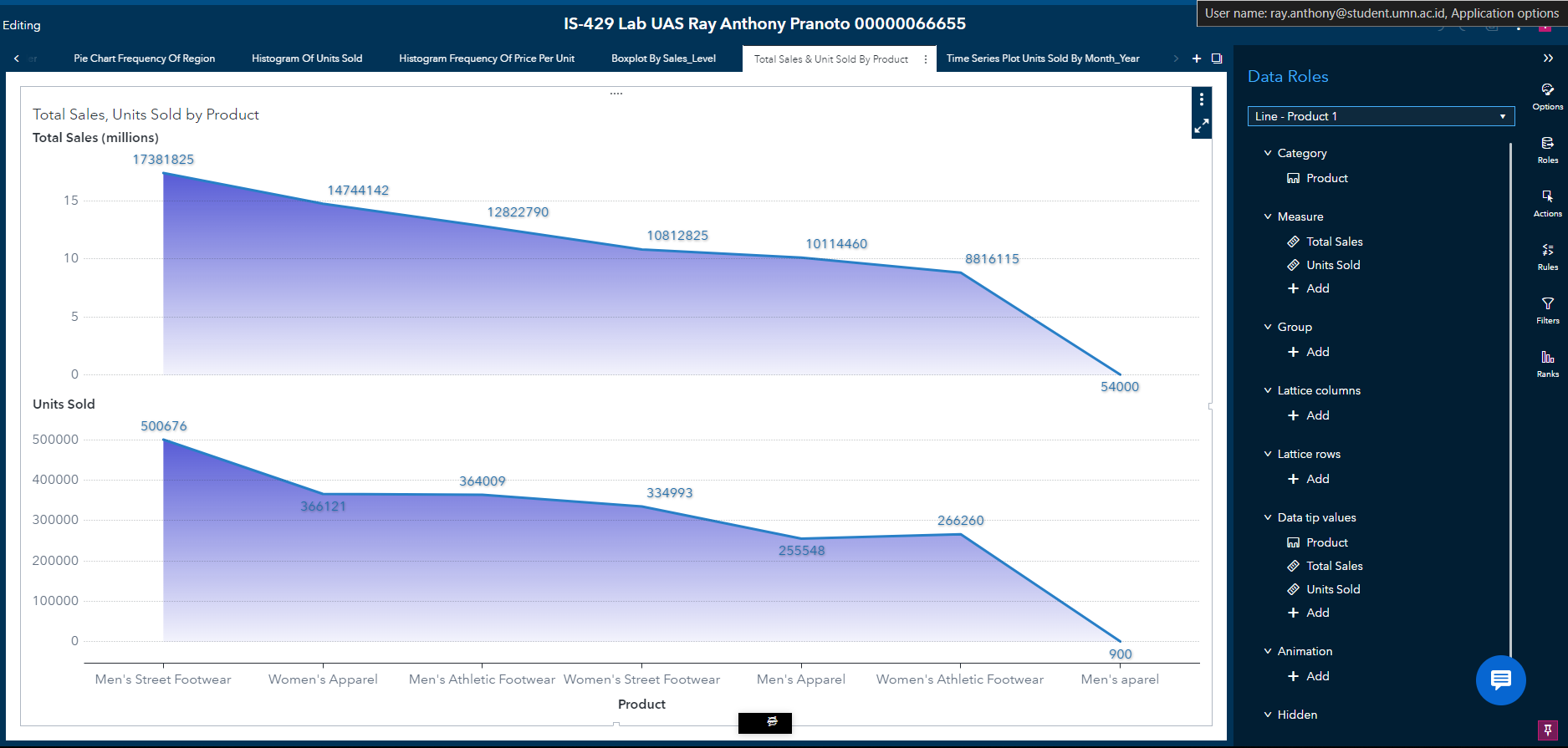
**Gambar 10**. Histogram Frekuensi Harga per Unit

Gambar 10 di atas adalah histogram yang menunjukkan distribusi frekuensi harga per unit dari suatu produk. Set roles pada visualisasi diatas adalah pada bagian measure diisi dengan kolom ‘Price per Unit’ dan pada bagian frequency diisi dengan nilai frequency. Sumbu horizontal (x-axis) merepresentasikan rentang harga per unit, mulai dari 0 hingga lebih dari 100. Sumbu vertikal (y-axis) menunjukkan frekuensi atau jumlah kejadian untuk setiap rentang harga yang ditentukan. Dari histogram ini, terlihat bahwa sebagian besar harga per unit berkisar antara 20 hingga 60. Rentang harga dengan frekuensi tertinggi berada di sekitar 40 hingga 50, dengan puncak yang jelas di antara 40 dan 45. Frekuensi harga per unit mulai menurun secara signifikan setelah 60, dengan beberapa harga mencapai di atas 80 tetapi dengan frekuensi yang jauh lebih rendah. Harga di bawah 20 dan di atas 100 jarang terjadi, menunjukkan bahwa sebagian besar produk dijual dalam rentang harga tengah tersebut. histogram ini menggambarkan distribusi harga yang cenderung normal dengan sedikit skewness ke kanan, mengindikasikan ada beberapa harga yang sangat tinggi namun jarang terjadi.



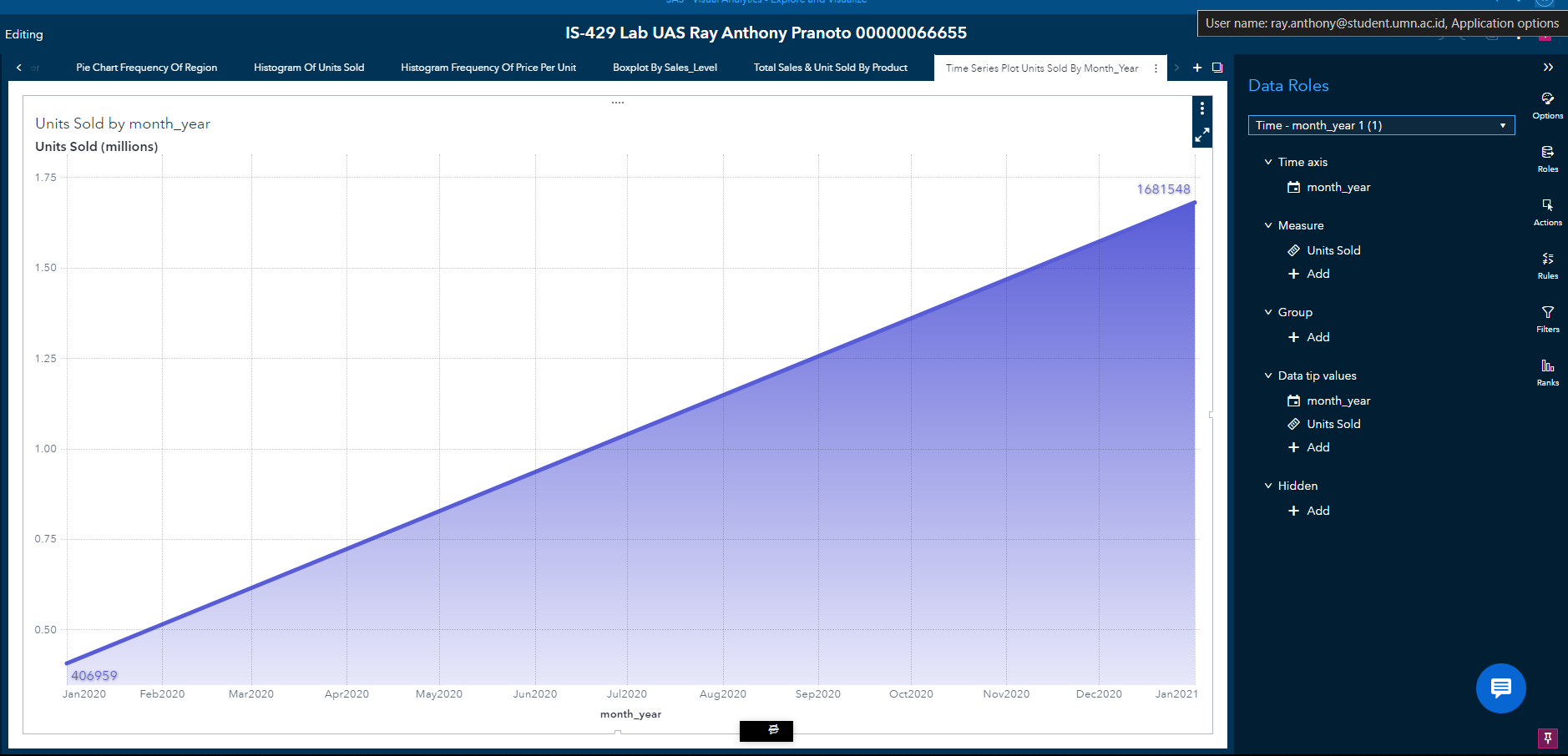
**Gambar 11.** Boxplot Laba Operasional & Total Penjualan Berdasarkan Tingkat\_Penjualan

Gambar 11 diatas merupakan Boxplot yang berguna untuk melihat perbandingan antara Operating Profit dan Total Sales berdasarkan dua kategori tingkat penjualan: "High" dan "Low". Set roles pada visualisasi diatas adalah pada bagian category diisi dengan kolom ‘sales\_level’ dan pada bagian measures diisi dengan kolom ‘Operating Profit’ dan ‘Total Sales’. Batang dalam diagram ini diwarnai dengan gradasi biru, di mana Operating Profit ditandai dengan warna biru tua dan Total Sales dengan warna biru muda. Pada kategori penjualan "Low", Total Sales mencapai hingga sekitar 80.000, menunjukkan rentang yang sangat luas. Sebaliknya, Operating Profit pada kategori ini jauh lebih kecil, dengan rentang sekitar 0 hingga sedikit di atas 20.000. Ini mengindikasikan bahwa meskipun penjualan total cukup tinggi, keuntungan operasional yang diperoleh relatif rendah. Kemudian pada kategori penjualan "High", baik Total Sales maupun Operating Profit menunjukkan rentang yang jauh lebih kecil dibandingkan dengan kategori "Low". Total Sales untuk kategori ini berkisar antara 0 hingga sekitar 10.000, sedangkan Operating Profit berkisar antara 0 hingga sedikit di atas 5.000. Ini menunjukkan bahwa pada tingkat penjualan yang tinggi, baik penjualan total maupun keuntungan operasional tetap berada pada tingkat yang lebih rendah dibandingkan dengan kategori penjualan "Low". Kemudian didalam boxplot juga terdapat informasi detail boxplot yang digunakan untuk melihat minimum average, Median, maximum, Std Dev, First Quartile, Third Quartile, dan Count pada masing masing Total Sales dan Operating Profit berdasarkan pada Sales\_Level.



**Gambar 12**. Grafik Garis Total Penjualan & Unit Terjual Berdasarkan Produk

Gambar 12 di atas merupakan Line Chart yang menggambarkan “Total Penjualan dan Jumlah Unit Terjual Berdasarkan Produk”. Set roles pada visualisasi diatas pada bagian category diisi dengan kolom ‘Product’, kemudian pada bagian measure diisi dengan kolom ‘Total Sales’ dan ‘Units Sold’ dan pada bagian data tip values diisi dengan kolom ‘Product’, ‘Total Sales’, dan ‘Units Sold’Berdasarkan visualisasi tersebut, 3 produk tertinggi terdapat pada produk Men’s Street Footwear, dengan jumlah penjualan mencapai $17.381.825 USD dan total unit terjual sebanyak 500.676 unit. Produk berikutnya adalah Women’s Apparel, dengan total penjualan sebesar $14.744.142 USD dan jumlah unit terjual sebanyak 366.121 unit. Kemudian Men’s Athletic Footwear, dengan total penjualan $12.822.790 USD dan jumlah unit terjual sebanyak 364.009 unit. Produk dengan total penjualan terendah adalah Men’s Apparel, dengan total penjualan hanya $54.000 USD dan jumlah unit terjual sebanyak 900 unit. Visualisasi ini memberikan gambaran yang jelas mengenai kinerja penjualan dari berbagai produk, menunjukkan produk mana yang paling sering dibeli dan mana yang kurang diminati oleh konsumen.



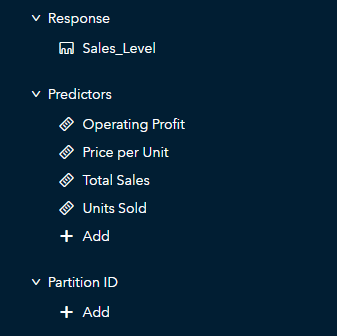
**Gambar 13**. Pertumbuhan Barang yang Terjual Pada 2020

Berdasarkan gambar 13 di atas tersebut menunjukkan pertumbuhan barang yang terjual sepanjang tahun 2020. Set roles pada visualisasi diatas adalah pada bagian time axis diisi dengan kolom ‘month\_year’ yang bertipe date, kemudian pada bagian measure diisi dengan kolom ‘Units Sold’, dan pada bagian data tip values diisi dengan kolom ‘month\_year’ dan ‘Units Sold’. Pada Januari 2020, terdapat 406.959 item yang terjual, dan angka ini terus meningkat setiap bulan, hingga mencapai 1.681.548 item pada Januari 2021. Fungsi dari visualisasi ini adalah untuk memantau pertumbuhan penjualan, yang berguna untuk memahami perkembangan pasar produk Zoka. Visualisasi tersebut memberikan wawasan penting mengenai tren penjualan dan membantu dalam pengambilan keputusan strategis untuk bisnis.

1. Question 2: CLO021 Sub-CLO-11, CLO021 Sub-CLO-12 and CLO013 Sub-CLO-14 Weight (60%)

Build and manage your Analytics model using SAS® Software

Sebelum membangun model yang akan digunakan untuk memproses dataset pada Perusahaan Zoka diperlukan adanya pengelompokkan data apa saja yang akan digunakan sebagai variable independent dan variable dependen agar proses modelling pada data Zoka dapat diproses.



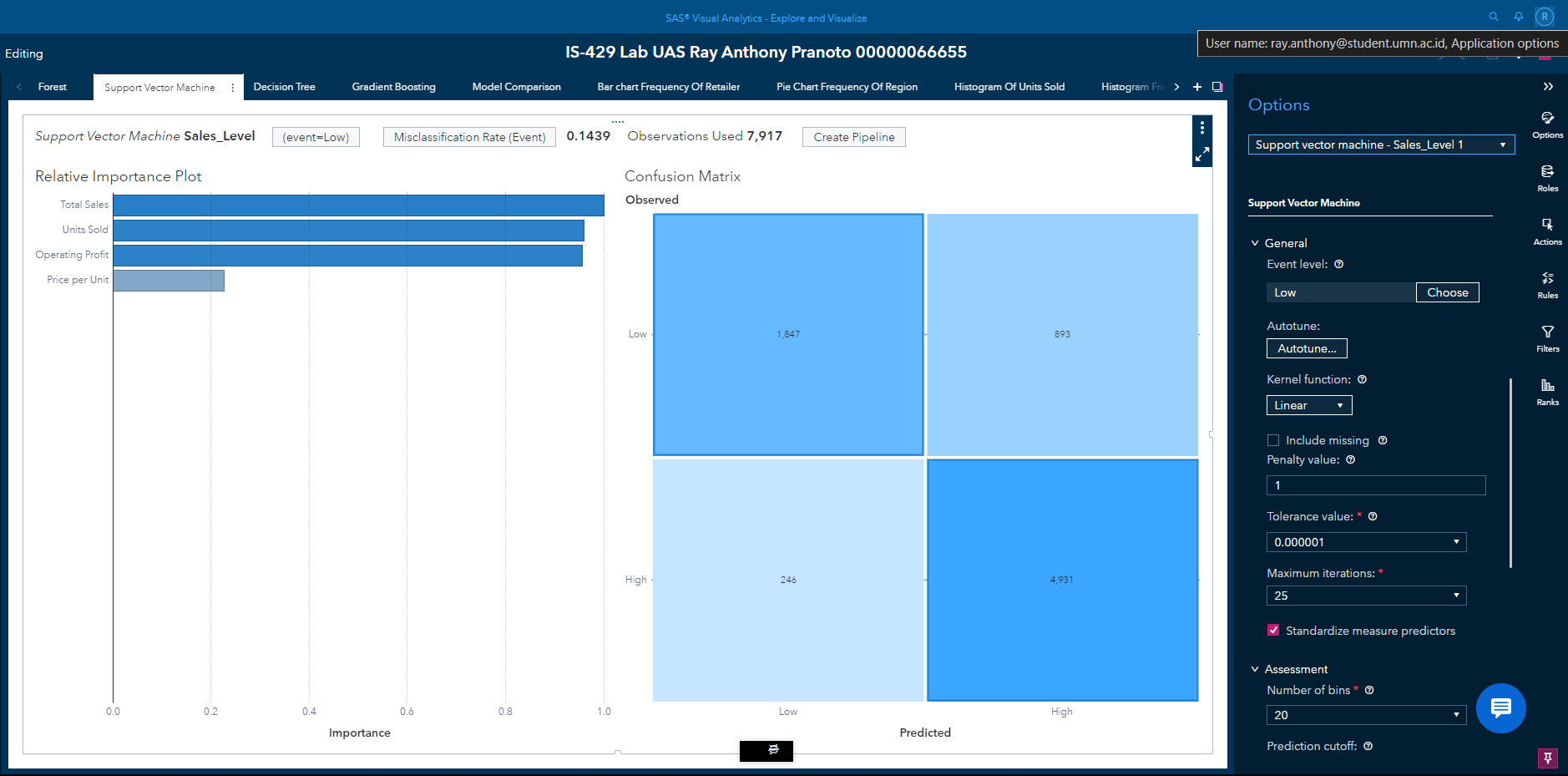
**Gambar 14. Data Roles Modelling**

Pada gambar 14 diatas merupakan proses mengelompokkan kolom kolom kedalam response dan predictors agar data dapat di modelkan. Pada gambar diatas response diisi dengan kolom ‘Sales\_Level’ dan pada bagian predictors terdapat 4 kolom bertipe numeric yaitu ‘Operating Profit’, ‘Price per Unit’, ‘Total Sales’, dan terakhir adalah kolom ‘Units Sold’. Pengelompokkan diatas berlaku untuk seluruh proses modelling yang akan digunakan yaitu Forest, Decision Tree, SVM, dan Gradient Boosting.



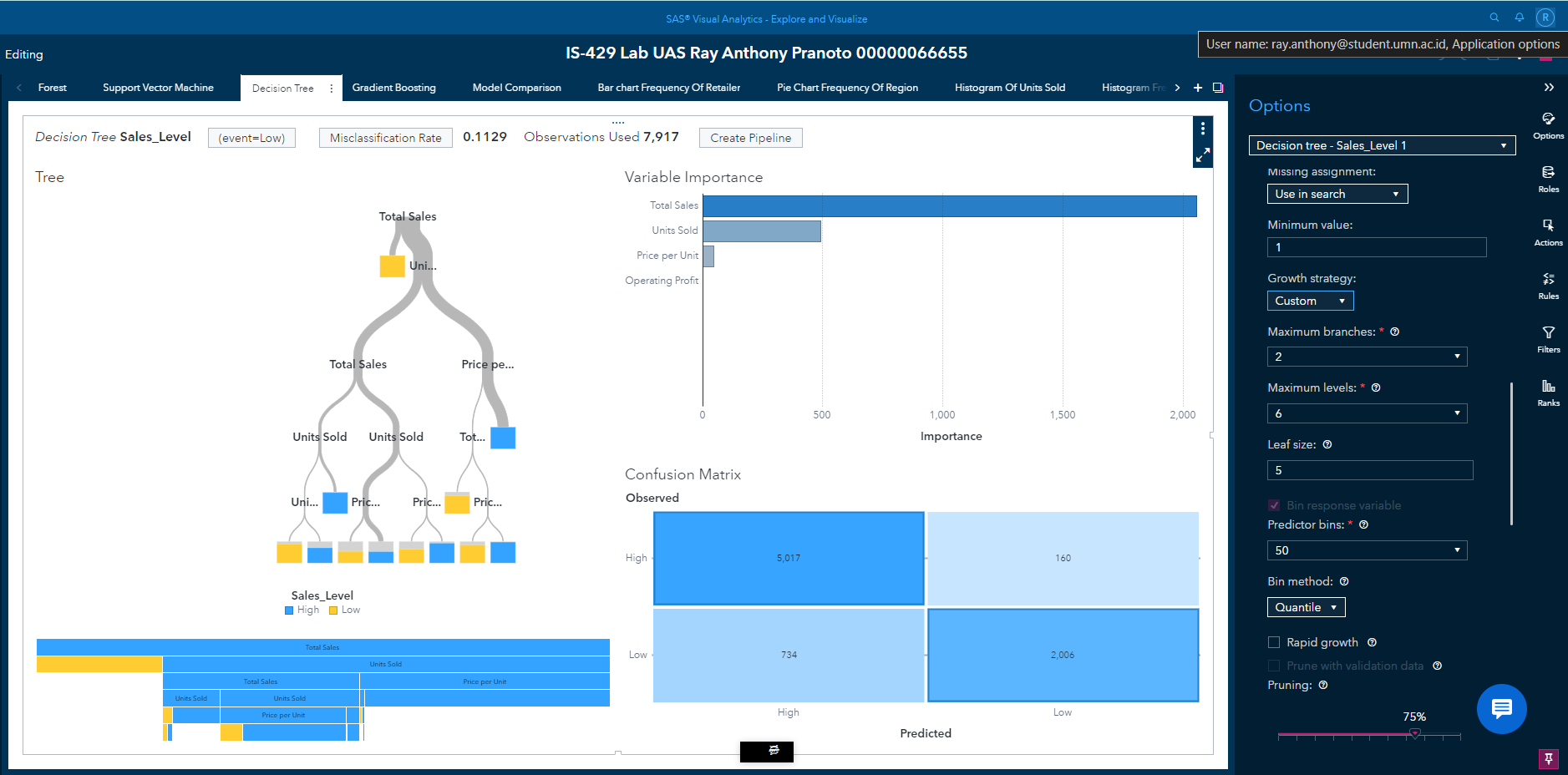
**Gambar 15**. Model Forest

Gambar 15 di atas menunjukkan pemodelan menggunakan algoritma Forest untuk membantu dalam pengambilan keputusan berbasis klasifikasi. Model diatas menggunakan jumlah pohon sebanyak 100 dan boostrap sebesar 0,6 dengan vote majority dan splitting criterion yang digunakan menggunakan information gain ratio. Kemudian Pada gambar tersebut, variabel respons diisi dengan 'Sales\_Level' sebagai variabel dependen untuk mengklasifikasikan apakah produk tersebut termasuk dalam kategori penjualan tinggi atau rendah. Variabel independen (predictors) yang digunakan adalah 'Total Sales', 'Operating Profit', 'Units Sold', dan 'Price per Unit'. Hasil pemodelan menunjukkan bahwa variabel yang paling mempengaruhi analisis model Forest ini adalah 'Total Sales', diikuti oleh 'Operating Profit', 'Units Sold', dan terakhir 'Price per Unit'. Tingkat kesalahan klasifikasi (Misclassification Rate) pada model Forest ini sebesar 0.1148, dengan skor F1 yang cukup tinggi yaitu 0.802. Confusion matrix juga menunjukkan bahwa prediksi pada level penjualan tinggi (high) memiliki angka yang paling tinggi, mengindikasikan akurasi yang baik dalam klasifikasi untuk kategori tersebut.



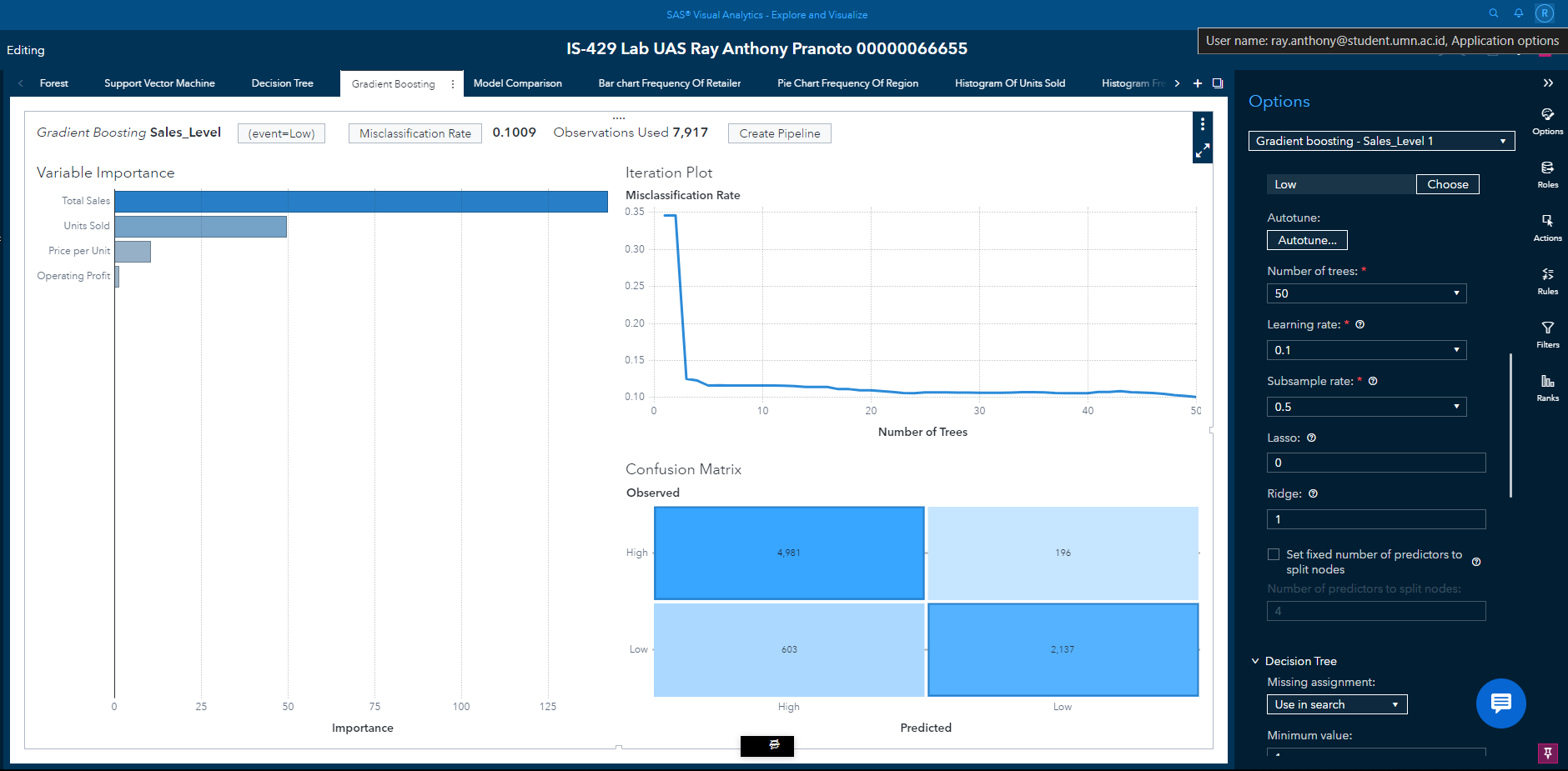
**Gambar 16**. Model Support Vector Machine (SVM)

Gambar 16 di atas menunjukkan pemodelan menggunakan algoritma Support Vector Machine (SVM) untuk Memisahkan dua kelas berbeda, Model diatas fungsi kernel yang digunakan adalah SVM linear, kemudian tolerance value sebesar 0.000001 dan nilai max iterations adalah 25. contoh pada data diatas adalah sales level Tingkat tinggi dan rendah. Pada gambar tersebut, variabel respons diisi dengan 'Sales\_Level' sebagai variabel dependen untuk mengklasifikasikan apakah produk tersebut termasuk dalam kategori penjualan tinggi atau rendah. Variabel independen (predictors) yang digunakan adalah 'Total Sales', 'Operating Profit', 'Units Sold', dan 'Price per Unit'. Hasil pemodelan menunjukkan bahwa variabel yang paling mempengaruhi analisis model Forest ini adalah 'Total Sales', diikuti oleh 'Operating Profit', 'Units Sold', dan terakhir 'Price per Unit'. Tingkat kesalahan klasifikasi (Misclassification Rate) pada model SVM ini sebesar 0.1439, dengan skor F1 yang cukup tinggi yaitu 0.764. Confusion matrix juga menunjukkan bahwa prediksi pada level penjualan tinggi (high) memiliki angka yang paling tinggi, mengindikasikan akurasi yang baik dalam klasifikasi untuk kategori tersebut.



**Gambar 17**. Model Decision Tree

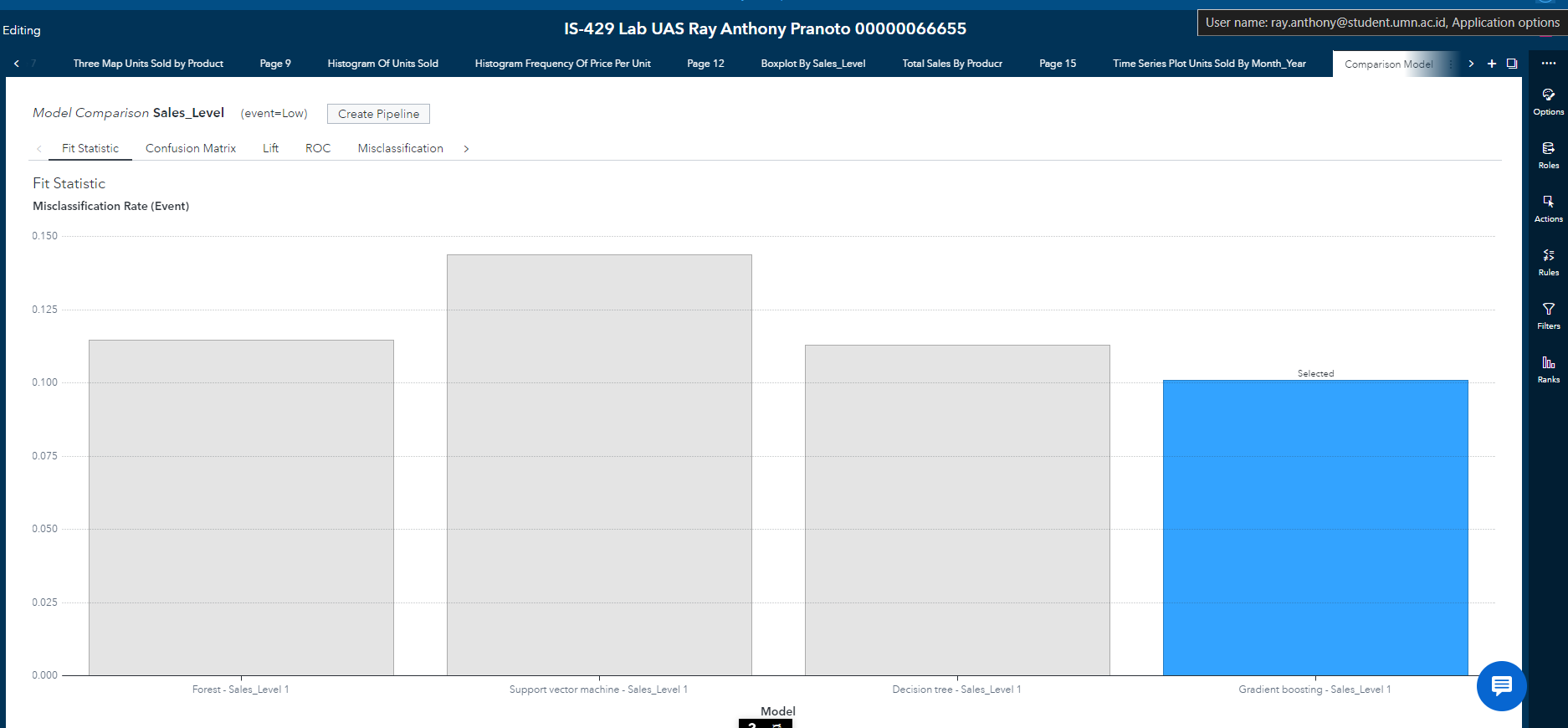
Gambar 17 di atas menunjukkan hasil pemodelan menggunakan algoritma Decision Tree untuk memprediksi 'Sales\_Level' yang dikategorikan menjadi dua kelas: tinggi (High) dan rendah (Low). Model ini mengatur maksimal cabang sebesar 2 dengan maksimum level sebesar 6 dan ukuran untuk setiap leaf pada decision tree tersebut sebear 5. Model ini juga menggunakan variabel independen 'Total Sales', 'Operating Profit', 'Units Sold', dan 'Price per Unit' untuk memprediksi apakah penjualan suatu produk tergolong tinggi atau rendah. Bagian kiri gambar menampilkan struktur pohon keputusan (decision tree) yang dimulai dengan 'Total Sales' sebagai akar pohon, menunjukkan bahwa 'Total Sales' adalah variabel paling penting dalam menentukan 'Sales\_Level'. Cabang-cabang berikutnya dipecah berdasarkan variabel 'Units Sold' dan 'Price per Unit', menandakan urutan kepentingan mereka dalam proses klasifikasi. Di sisi kanan atas, grafik "Variable Importance" menunjukkan bahwa 'Total Sales' memiliki pengaruh terbesar dalam model, diikuti oleh 'Units Sold', 'Price per Unit', dan terakhir 'Operating Profit', yang memiliki pengaruh minimal. Kemudian misclassification rate pada model decision tree menunjukan yang kecil 0.1129 dengaan f1 score yang cukup tinggi 0.818 yang menandakan persentase klasifikasi pada model ini memiliki tingkat keberhasilan yang tinggi dalam memisahkan kelas penjualan tinggi dan rendah. Di bagian kanan bawah, terdapat confusion matrix yang mengilustrasikan kinerja model. Matrix ini menunjukkan bahwa dari 5,177 prediksi untuk kelas tinggi (High), model berhasil memprediksi dengan benar 5,017 dan salah 160. Untuk kelas rendah (Low), dari 2,740 prediksi, model benar dalam 2,006 kasus dan salah dalam 734 kasus.



**Gambar 18**. Model Gradient Boosting

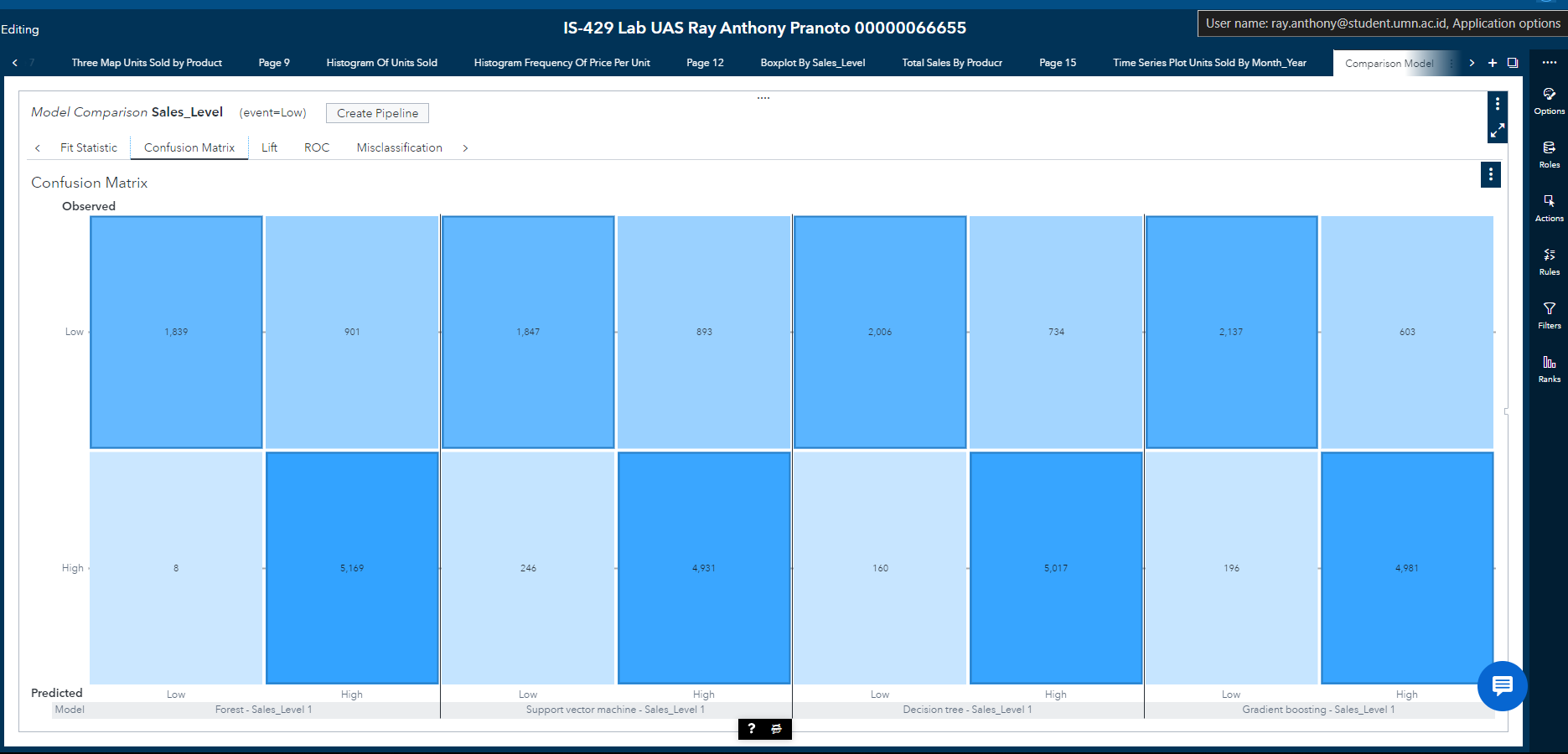
Gambar 18 diatas menunjukan hasil pemodelan menggunakan Gradient Boosting yang berfungsi untuk menggabungkan beberapa model prediktif yang lebih lemah (biasanya decision tree sederhana) untuk membentuk model prediktif yang lebih kuat. Hal ini dilakukan dengan cara membangun model secara bertahap, di mana setiap model baru berusaha untuk memperbaiki kesalahan dari model sebelumnya. Pada model ini digunakan untuk memprediksi tingkat penjualan (Sales\_Level) dengan dua kategori: tinggi (High) dan rendah (Low). Model ini melakukan pengaturan nilai pada pohon sebesar 50 dengan learning rate nya 0.1, dan subsample rate nya sebesar 0.5. Dalam model ini, total terdapat 7.917 observasi yang digunakan dengan tingkat misclassifikasi sebesar 0,1009 dengan f1 score yang cukup tinggi yaitu 0.842 yang menandakan bahwa persentase klasifikasi pada model ini memiliki tingkat keberhasilan yang tinggi dalam memisahkan kelas penjualan tinggi dan rendah. Bagian "Variable Importance" menampilkan pentingnya masing-masing variabel dalam model. Terlihat bahwa "Total Sales" adalah variabel yang paling berpengaruh, diikuti oleh "Units Sold", "Price per Unit", dan "Operating Profit". Grafik "Iteration Plot" menunjukkan bagaimana tingkat misclassifikasi berubah seiring dengan bertambahnya jumlah pohon dalam model Gradient Boosting. Misclassifikasi awalnya tinggi, namun menurun drastis dan stabil setelah sekitar 10 pohon, menunjukkan bahwa penambahan pohon lebih lanjut tidak banyak mengurangi tingkat kesalahan. Matriks kebingungan (Confusion Matrix) memperlihatkan performa model dalam mengklasifikasikan data observasi ke dalam kategori tinggi dan rendah. Dari 4.981 observasi yang sebenarnya memiliki penjualan tinggi, model berhasil mengklasifikasikan 4.981 dengan benar dan salah mengklasifikasikan 196. Sebaliknya, dari 2.740 observasi dengan penjualan rendah, 2.137 diklasifikasikan dengan benar sementara 603 salah klasifikasi sebagai tinggi.

**Model Comparison:**



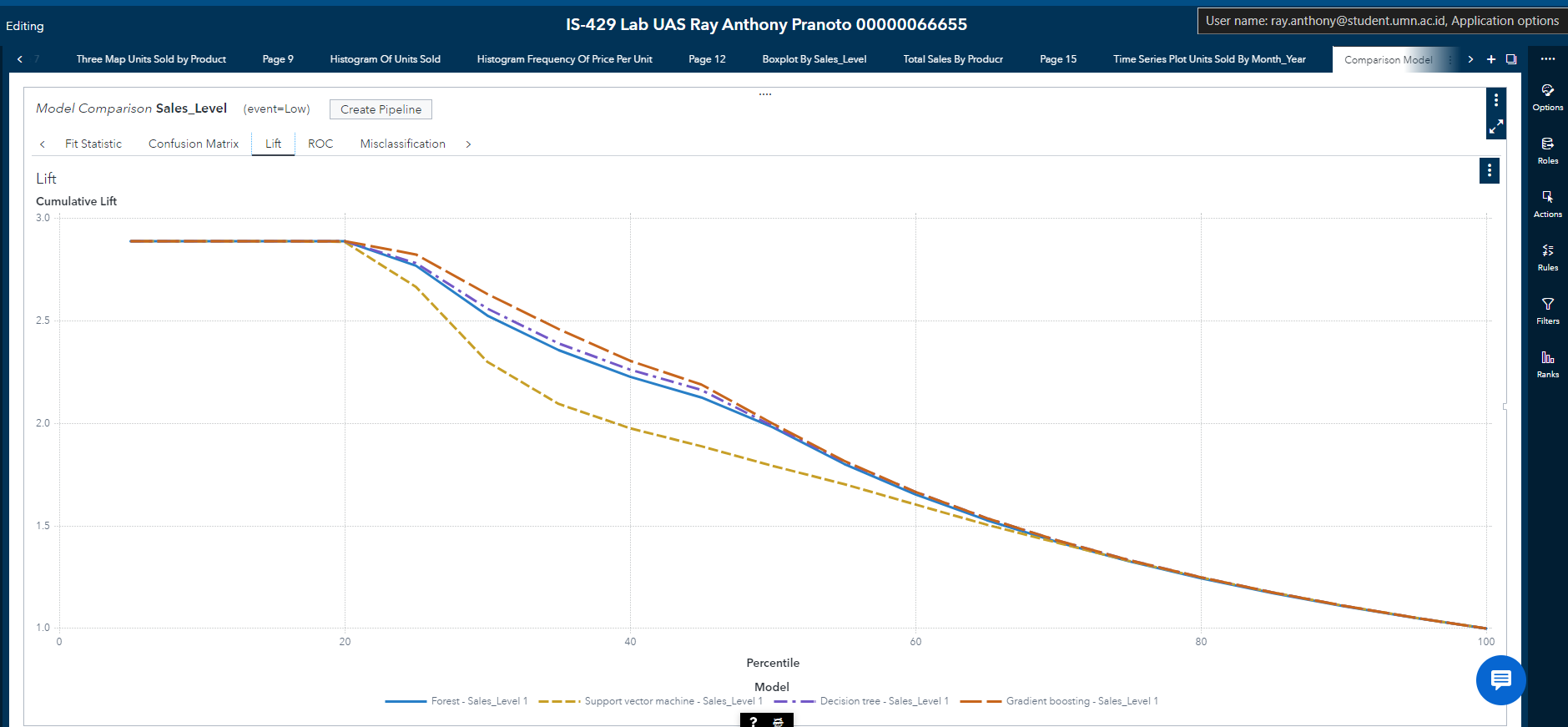
**Gambar 19**. Fit Stastistic Model Comparison

Gambar 19 diatas merupakan hasil dari beberapa model yang telah dibuat terdapat model Forest, Support Vector Machine, Decision Tree, dan Gradient Boosting untuk di compare mengenai tingkat keberhasilan tertinggi sehingga ditemukan model yang cocok digunakan untuk dataset tersebut. Berdasarkan hasil dari gambar diatas fit statistic diukur menggunakan misclassification rate pada masing masing model, sehingga makin kecil misclassification rate pada masing masing model maka semakin bagus model tersebut untuk data Zoka. Hasil dari fit statistic yang dipilih adalah model Gradient Boosting dengan misclassification rate 0.1009 dengan F1 score yang lebih tinggi dibandingkan model lainnya adalah 0.842, kemudian diikuti oleh Decision Tree dengan misclassification rate 0.1129 dengan F1 score 0.818, kemudian Forest model dengan misclassification rate 0,1148 dengan F1 score 0.802, dan terakhir adalah SVM model dengan misclassification rate 0,1439 dengan F1 score 0,764.



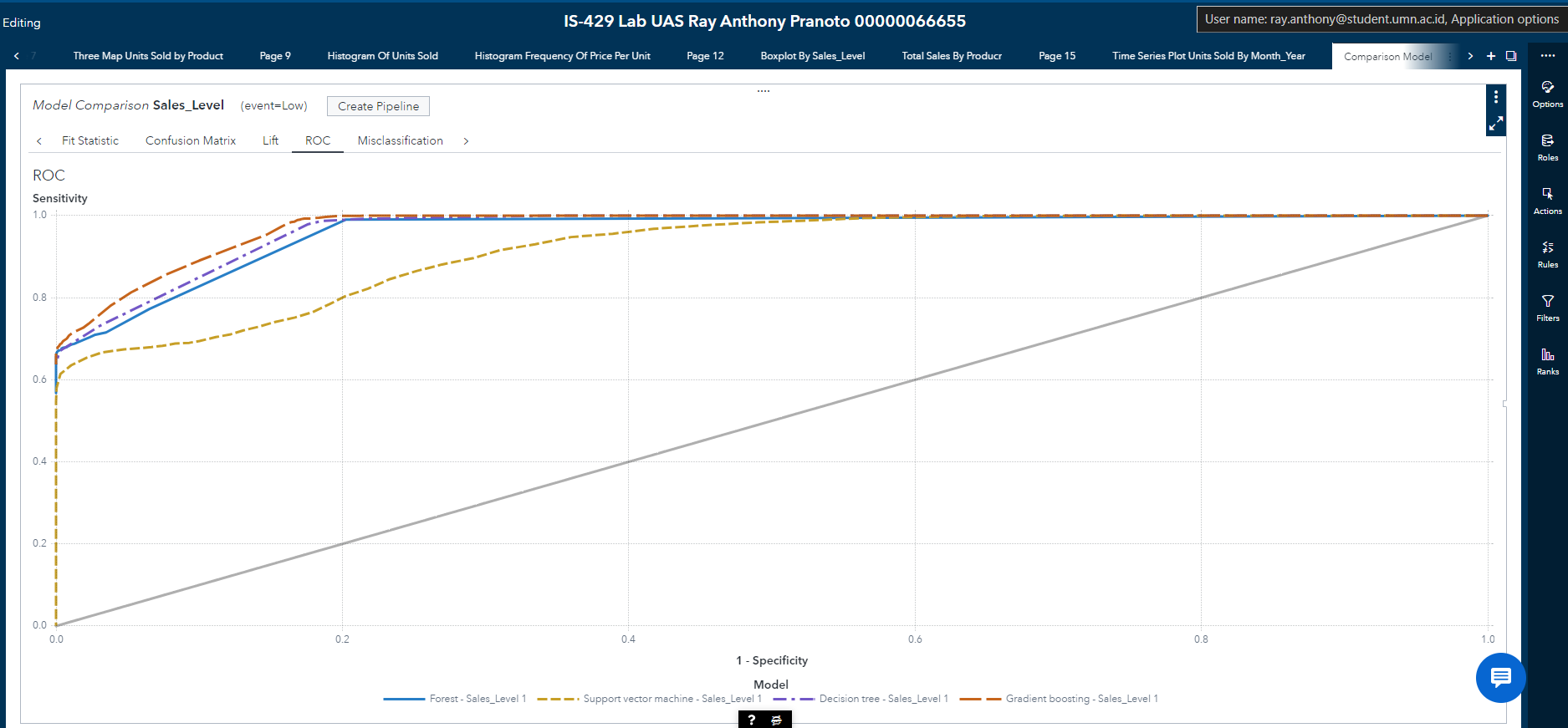
**Gambar 20**. Confusion Matrix Comparison Model

Berdasarkan gambar 20 di atas, confusion matrix dari masing-masing model, dapat dilihat bahwa secara rata-rata model memiliki tingkat keberhasilan yang tinggi dalam memisahkan kelas penjualan tinggi dan rendah. Hal ini ditunjukkan oleh nilai true positif dan true negatif yang lebih tinggi dibandingkan dengan false positif dan false negative. Model Gradient Boosting, secara khusus, memiliki nilai true positif sebesar 2.137 dan true negative sebesar 4.981, yang lebih tinggi dibandingkan dengan model lainnya. Selain itu, nilai false negative sebesar 196 dan false positif sebesar 603 lebih rendah dibandingkan dengan model-model lain. Oleh karena itu, berdasarkan nilai-nilai ini, serta misclassification rate dan F1 score, dapat disimpulkan bahwa model Gradient Boosting adalah model terbaik untuk klasifikasi data dalam memisahkan kelas penjualan tinggi dan rendah di antara decision tree, SVM, forest, dan Gradient Boosting.



**Gambar 21**. Lift Comparison Model

Gambar 21 di atas menunjukkan grafik cumulative lift yang digunakan untuk membandingkan kinerja beberapa model dalam mengklasifikasikan data berdasarkan level penjualan (Sales\_Level). Model yang dibandingkan meliputi Forest, Support Vector Machine (SVM), Decision Tree, dan Gradient Boosting. Grafik cumulative lift ini memplot lift kumulatif terhadap persentil data yang diurutkan berdasarkan prediksi dari model. Dari grafik tersebut, terlihat bahwa model Gradient Boosting (garis oranye) memiliki performa yang bagus dan menunjukkan lift yang lebih tinggi di awal persentil dibandingkan dengan model lain. Model Forest (garis biru), Decision Tree juga menunjukkan performa yang baik, namun sedikit di bawah Gradient Boosting. Model SVM (garis kuning putus-putus) terlihat memiliki performa yang paling rendah di antara keempat model, terutama di bagian awal persentil. Sehingga secara keseluruhan, grafik ini menunjukkan bahwa model Gradient Boosting memiliki kemampuan yang lebih baik dalam memisahkan kelas penjualan tinggi dan rendah dibandingkan dengan model lainnya.



**Gambar 22**. ROC Comparison Model

Gambar 22 di atas adalah grafik ROC (Receiver Operating Characteristic) yang digunakan untuk membandingkan performa beberapa model klasifikasi dalam memprediksi "Sales\_Level" dengan label "Low". Grafik ROC menunjukkan hubungan antara Sensitivitas (True Positive Rate) dan Spesifisitas (False Positive Rate) untuk berbagai threshold keputusan. Garis diagonal abu-abu menunjukkan performa model yang acak, di mana tidak ada kemampuan untuk membedakan antara dua kelas. Semakin dekat kurva ROC ke sudut kiri atas, semakin baik performa model dalam membedakan antara kelas "Low" dan kelas lainnya. Dari visualisasi ini, terlihat bahwa model Gradient Boosting berwarna orange memiliki performa yang terbaik karena kurvanya paling dekat dengan sudut kiri atas, menunjukkan kombinasi sensitivitas dan spesifisitas yang tinggi. Kemudian Model Forest berwarna ungu menunjukkan performa yang baik, meskipun sedikit lebih rendah dibandingkan model Gradient Boosting. Model Decision Tree dan Support Vector Machine memiliki performa yang kurang dibandingkan dua model sebelumnya, dengan Support Vector Machine menunjukkan performa terendah di antara keempat model. Sehingga model yang terbaik adalah model Gradient Boosting.



**Gambar 23**. Misclassification Model Comparison

Gambar 23 di atas adalah diagram batang yang menunjukkan frekuensi klasifikasi benar (biru) dan salah (kuning) untuk berbagai model prediksi—Forest, Support Vector Machine, Decision Tree, dan Gradient Boosting—berdasarkan dua tingkat penjualan: "Low" dan "High". Pada kategori "Low", semua model menunjukkan frekuensi klasifikasi salah yang lebih tinggi dibandingkan kategori "High". Model Gradient Boosting menunjukkan performa terbaik dengan kesalahan klasifikasi yang relatif lebih rendah dibandingkan model Forest, Support Vector Machine, dan Decision Tree pada kategori "Low". Pada kategori "High", semua model menunjukkan frekuensi klasifikasi benar yang sangat dominan, tetapi Gradient Boosting tetap menonjol dengan konsistensi dan akurasi yang lebih baik secara keseluruhan. Diagram ini mengindikasikan bahwa model Gradient Boosting adalah yang paling bagus dalam mengklasifikasikan tingkat penjualan, baik pada kategori "Low" maupun "High".

Dari semua itu dapat disimpulkan bahwa Pemodelan yang dilakukan dengan beberapa model statistik untuk menentukan yang terbaik dalam memprediksi penjualan, Pemodelan Gradient Boosting terpilih sebagai model terbaik. Model ini memiliki misclassification rate terendah (0.1009) dan F1 score tertinggi (0.842), menunjukkan performa yang baik dalam memisahkan kelas penjualan tinggi dan rendah, menunjukkan akurasi tinggi dan tingkat kesalahan klasifikasi rendah. Implementasi hasil analisis ini memudahkan Zoka untuk memprediksi permintaan produk, mengelola persediaan secara efisien, dan meningkatkan efisiensi dalam tahap produksi dan distribusi. Optimalisasi rantai pasokan dan manajemen stok juga diperlukan untuk mengurangi biaya dan meningkatkan efisiensi operasional. penelitian ini juga merekomendasikan Zoka untuk memfokuskan strategi penjualan pada produk-produk yang diklasifikasikan sebagai 'Sales Level High' untuk meningkatkan profitabilitas, mengembangkan strategi khusus untuk meningkatkan penjualan produk-produk yang diklasifikasikan sebagai 'Sales Level Low', dan memanfaatkan insights dari data penjualan untuk merancang strategi pemasaran dan promosi yang lebih efektif dan tertarget.