Analisis Penjualan Produk Manufaktur Adidas Menggunakan SAS

Untuk Mengefektifkan Pemilihan Retailer

Gempar Bambang Godata

Universitas Multimedia Nusantara

Tangerang, Indonesia

[gempar.bambang@student.umn.ac.id](mailto:gempar.bambang@student.umn.ac.id)

***Abstract*—** **The global textile industry has been in the doldrums due to low consumer purchasing power over the past two years. Adidas as one of the market leaders in the apparel and sports equipment manufacturing business has also been affected in terms of high stock but low demand. So Adidas needs all kinds of strategies to maximize existing resources to generate profits along with the improvement of the consumer economy. These strategies can be assisted by machine learning-based programs that can classify the right products to be sold by certain retailers. This research aims to build the best machine learning model with algorithms provided by SAS Visual Analytics. This article provides an in-depth insight into implementing Big Data Analytics applications to provide business solutions for the manufacturer industry.**

**Index Terms—big data analytics; data visualization; decision tree C4.5; random forest; logistic regression; bayesian network; SVM; Adidas; SAS Visual Analitycs.**

# Latar Belakang & Pemahaman Bisnis

Industri tekstil pakaian olahraga bergerak dinamis sesuai dengan perubahan permintaan yang musiman. Lingkungan bisnis ini juga semakin kompetitif yang mana permintaan menurun tetapi merek produsen semakin bertambah baik berskala lokal maupun internasional. Sehingga setiap produsen akan berlomba-lomba dalam merumuskan strategi paling efektif dalam meningkatkan penjualan mereka dan menguasai pasar.

Adidas merupakan nama produsen pakaian dan peralatan olahraga asal Jerman yang telah memiliki popularitas tinggi di dunia. Adidas sendiri telah membangun pabriknya di Indonesia dikarenakan tingginya permintaan ekspor oleh retailer di dunia. Namun, saat pandemi melanda, industri tekstil dan sepatu menunjukkan kelesuan permintaan. Produsen pun membuat berbagai kebijakan untuk melakukan efisiensi biaya operasional termasuk dengan Adidas yang memili banyak stok tersedia tetapi tidak laku. Diawal tahun 2023, Badan Pusat Statistik menunjukkan data bahwa sampai akhir April, ekspor dari industri tekstil meningkat singnifikan sebesar 34,28% dibandingkan selama 3 kuartal di tahun 2022 [1].

Dengan pemulihan kembali permintaan produk industri tekstil, maka Adidas juga membutuhkan sebanyak mungikin wawasan terbaru untuk menyiapkan strategi paling efektif untuk menguasai pasar dengan kondisi ekonomi konsumen belum pulih sepenuhnya. Salah satu strategi yang dapat dilakukan oleh Adidas ialah mengalokasikan stok produk ke retailer yang mampu menjualnya dengan cepat dan untung. Wawasan untuk mengembangkan strategi ini dapat melalui sebuah program berbasis machine learning yang dapat menganalisis dan mengklasifikasikan tren produk yang dijual oleh para retailer produk bermerek Adidas. Adidas perlu mengetahui apa dan bagaimana strategi bisnis yang ingin dibuat, objek bisnis yang diinginkan, bagaimana mendapatkan data, dan menentukan model terbaik untuk mencapai tujuan bisnisnya. Kemudian dari hasil tersebut dapat ditentukan rencana dan strategi untuk mencapai tujuan bisnis dari Adidas.

# Tinjauan Teoritis

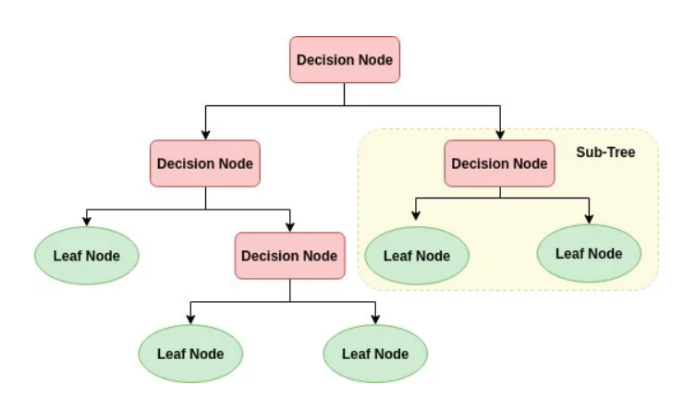
## Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) adalah metode untuk menggambarkan data dengan mengidentifikasi fitur-fitur utamanya dan menyusunnya secara logis. EDA berfokus pada pemeriksaan asumsi yang diperlukan untuk penyesuaian model dan pengujian hipotesis, penanganan nilai-nilai yang hilang, serta pengubahan variabel jika diperlukan [4]. EDA memberikan gambaran sederhana tentang jumlah baris dan kolom, data yang hilang, tipe data, dan tinjauan awal dataset, memperbaiki data yang rusak dan menangani nilai-nilai yang hilang, tipe data yang tidak valid, dan outlier [7].

Selain itu, EDA menggunakan berbagai jenis visualisasi seperti diagram batang, histogram, box plot, dan berbagai metode lainnya untuk menggambarkan sebaran data. EDA juga dapat menemukan kekuatan korelasi atau hubungan antar variabel dan memvisualisasikannya dalam bentuk heat map [7].

## Decision Tree C4.5

Salah satu model klasifikasi yang cukup terkenal adalah Decidion Tree. Metode ini sangat populer karena dapat dengan mudah dijelaskan oleh manusia. Model klasifikasi menggunakan pohon keputusan yang terdiri dari deskripsi dan prediksi. Dalam menggunakan metode pohon keputusan, terdapat tiga komponen utama yang perlu diketahui, yaitu akar (root node), simpul internal (internal node), dan simpul daun (leaf node). Akar adalah titik awal dalam pengambilan keputusan pada pohon keputusan. Simpul internal adalah cabang-cabang yang terbentuk berdasarkan kriteria-kriteria selanjutnya. Sedangkan simpul daun mewakili keputusan yang dihasilkan [14].



*Gambar Ilustrasi Algoritma Decision Tree*

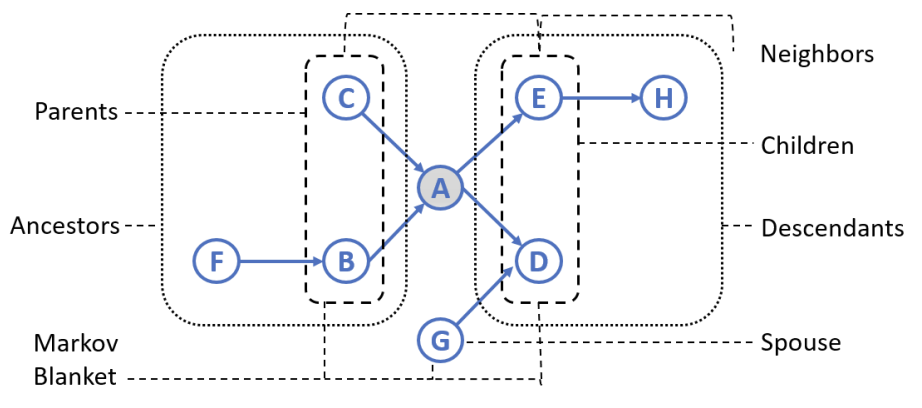
## Random Forest Classification

Random Forest Classifier (RFC) merupakan algoritma pada machine learning yang digunakan untuk melakukan klasifikasi dataset untuk jumlah yang besar. Hal ini dikarenakan RFC bisa digunakan pada banyak dimensi dengan berbagai skala dan performa yang tinggi. Klasifikasi ini dibangun dengan menggunakan penggabungan tree dalam decision tree dengan melakukan training terlebih dahulu pada dataset yang ingin digunakan [10].

RFC sendiri dapat dikategorikan sebagai algoritma terbaik dalam machine learning, algoritma ini menggunakan decision tree atau pohon keputusan dalam melakukan proses seleksi. Tree atau pohon pada decision tree akan dibagi secara rekursif berdasarkan data di kelas yang sama. Penggunaan tree yang semakin banyak pada algoritma ini akan memengaruhi nilai akurasi sehingga lebih maksimal. Klasifikasi ditentukan berdasarkan hasil voting dan tree yang terbentuk.

Seperti yang telah dikatakan sebelumnya, RFC ini bekerja dengan membangun decision tree dan melakukan penggabungan untuk mendapatkan prediksi yang stabil dan akurat. Kumpulan decision tree, biasanya disebut sebagai ‘hutan’, seringkali dibangun dan dilatih dengan metode bagging. Metode bagging sendiri merupakan kombinasi model pembelajaran untuk meningkatkan hasil kesuluruhan. Selain membangun tree, algoritma ini juga meningkatkan keacakan pada model untuk digunakan dalam mencari fitur diantara subset fitur ini. Sehingga algoritma ini menghasilkan keragaman yang sangat luas, sehingga biasanya menghasilkan model yang lebih baik [11].

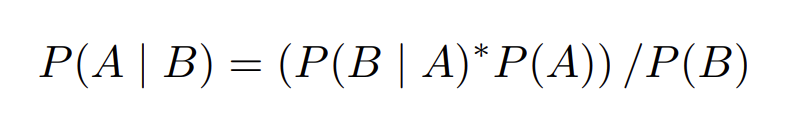
## Bayesian Network



*Gambar Struktur Model Bayesian Network*

*Sumber: Wang*

Algoritma yang digunakan pada struktur Bayesian Network SAS ada empat. Pertama ialah Naïve Bayes, algoritma Naïve Bayes adalah sebuah algoritma yang digunakan dalam bidang data mining dan machine learning untuk menghitung probabilitas suatu kejadian berdasarkan data yang telah dikumpulkan [12]. Algoritma ini menggunakan teori probabilitas dan statistik untuk menghitung probabilitas suatu kejadian, dengan asumsi bahwa setiap fitur (variabel) yang digunakan dalam analisis independen terhadap fitur lainnya [13]. SAS mendefinisikan Naïve sebagai pemanjang dari sebuah Bayesian Network dengan cara mengizinkan predictor node independen membentuk Bayesian Network baru.



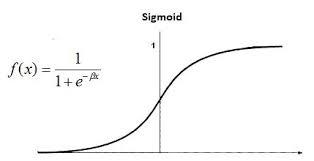
*Gambar Rumus Algoritma Naïve Bayes*

1. P(A | B) adalah probabilitas kejadian A terjadi (target variable) jika kejadian B terjadi (predictor variable)
2. P(B | A) adalah probabilitas kejadian B terjadi jika kejadian A terjadi
3. P(A) adalah probabilitas kejadian A terjadi
4. P(B) adalah probabilitas kejadian B terjadi

Algoritma Bayesian kedua ialah Tree-augmented naïve, yang didefinisikan SAS sebagai penumbuh pohon yang menghubungkat langsung variabel respon kepada setiap predictor. Ketiga ialah algoritma Parent-child Bayesian sebagai perpanjangan node yang berfungsi mengizinkan predictor menjadi parent, menghubungkan childern dari response variable. Keempat ialah algoritma Markov Blanket yang didefinisikan sebagai subset minimum yang diperlukan nodes untuk memprediksi pola dirinya sendiri [12]. Keempat pembangun struktur Bayesian Network ini menjadikannya mampu menggabungkan informasi sebelumnya, pengetahuan, dan segala hal yang berhubungan dengan variabel subjek tanpa data yang sesuai dengan interpretasi sebab-akibat daripada hanya berbasis korelasi antar variabel.

## Logistic Regression

Logistic Regression adalah metode statistik yang populer digunakan untuk memprediksi probabilitas terjadinya suatu kejadian biner atau kategori berdasarkan serangkaian variabel independen. Model ini banyak dipakai karena kesederhanaannya dalam menginterpretasi hasil, dan kemampuannya yang tinggi dalam menghadapi masalah klasifikasi biner [8]. Logistic Regression didasarkan pada fungsi logistik atau sigmoid yang mengubah nilai prediksi linier menjadi probabilitas antara 0 dan 1. Fungsi logistik memodelkan hubungan antara variabel independen (biasanya disebut sebagai fitur atau prediktor) dan probabilitas terjadinya kejadian. Model Logistic Regression mengestimasi koefisien (bobot) untuk setiap variabel independen untuk mengukur pengaruhnya terhadap probabilitas kejadian [16].



*Gambar Ilustrasi Kurva Logistic Regression*

*Sumber:* [*https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/*](https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/)

Proses pelatihan Logistic Regression melibatkan penyesuaian koefisien model dengan menggunakan metode Maximum Likelihood Estimation untuk memaksimalkan kemungkinan observasi yang diamati. Dalam tahap ini, variabel independen dan label yang diketahui digunakan untuk mempelajari hubungan antara variabel dan probabilitas kejadian. Hasilnya adalah model logistik yang dapat digunakan untuk memprediksi probabilitas kejadian yang baru berdasarkan nilai variabel independen yang diberikan.

Logistic Regression dapat digunakan untuk berbagai tujuan, seperti klasifikasi biner, prediksi peluang, analisis faktor risiko, dan pengendalian variabel. Hasil dari Logistic Regression dapat diinterpretasikan ke dalam nilai odds ratio, yang merupakan perbandingan antara peluang kejadian di satu kelompok dibandingkan dengan kelompok yang direferensikan.

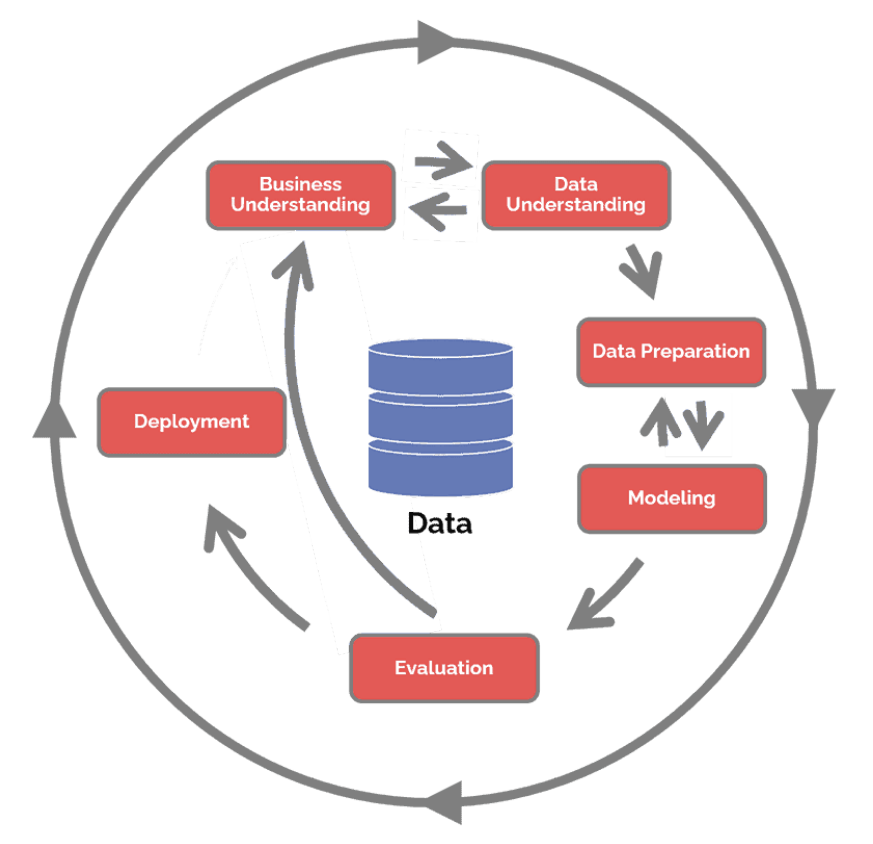
## Support Vector Machine

Support Vector Machine (SVM) merupakan salah satu metode yang ada dalam supervised learning yang digunakan untuk melakukan klasifikasi ataupun regresi. Dalam hal melakukan klasifikasi, SVM dipandang lebih matang dan lebih jelas secara matematis daripada teknik-teknik klasifikasi lainnya. Selain itu, SVM juga dapat melakukan klasifikasi dan regresi secara linear ataupun non-linear.

SVM digunakan untuk menemukan hyperplane terbaik dengan cara memaksimalkan jarak antar kelas. Hyperplane sendiri merupakan sebuah fungsi yang digunakan sebagai pemisah antar kelas. Pada 2-D (dua dimensi), fungsi yang digunakan untuk melakukan klasifikasi disebut line whereas, sedangkan pada 3-D (tiga dimensi) disebut plane similarly. Dan fungsi untuk klasifikasi dalam ruang dimensi yang lebih tinggi disebut hyperplane [8].

Secara garis besar, ketika data tidak dapat dipisahkan secara linear dalam ruang input, maka SVM tidak dapat menemukan pemisah dalam hyperplane, hal ini akan berdampak pada nilai akurasi yang tidak maksimal dan tidak menggeneralisasi dengan baik. Maka dari itu dibutuhkan kernel seperti kernel Linear, Polynomial, dan Radial Basis Function untuk melakukan transformasi data ke ruang dimensi yang lebih tinggi. Kernel sendiri berguna untuk memisahkan data-data secara linear.

## CRISP-DM



*Gambar Siklus Tahapan CRISP-DM*

Metode CRISP-DM merupakan pendekatan yang sistematis dan berulang dalam proses penambangan data yang digunakan untuk memahami bisnis, mempersiapkan data, membangun model, dan mengevaluasi hasilnya. Metode ini populer karena fleksibilitasnya dan dapat diterapkan di berbagai industri [6] termasuk industri tekstil dan sepatu.

Dalam tahap business understanding, fokus utama adalah memahami tujuan bisnis, kebutuhan analisis, dan menetapkan batasan serta sasaran yang ingin dicapai. Tahap pemahaman data melibatkan eksplorasi data untuk memahami karakteristik, kualitas, dan keterkaitan antar variabel dalam dataset. Pada tahap data preprocessing, langkah-langkah seperti identifikasi dan penanganan nilai yang hilang atau outliers dilakukan untuk memastikan integritas data. Ketika data telah dinyatakan siap, tahap pemodelan dilakukan dengan menerapkan algoritma yang sesuai. Pada tahap ini, model statistik atau algoritma pembelajaran mesin dikembangkan dan disesuaikan dengan data yang ada.

Setelah tahap pemahaman bisnis dan pemodelan selesai dilakukan, langkah evaluasi dilakukan. Metode ini membandingkan prediksi label dengan label yang sebenarnya sebagai patokan. Confusion Matrix ini berguna untuk mengevaluasi performa dari model yang telah dibuat. Dengan Confusion Matrix, kita dapat mendapatkan gambaran tentang kemampuan model dalam mengklasifikasikan data dengan benar dan secara keseluruhan. Confusion Matrix terdiri dari empat bagian, yaitu True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). TP menunjukkan jumlah data yang berhasil diklasifikasikan dengan benar sebagai positif, TN menunjukkan jumlah data yang berhasil diklasifikasikan dengan benar sebagai negatif, FP menunjukkan jumlah data yang salah diklasifikasikan sebagai positif, dan FN menunjukkan jumlah data yang salah diklasifikasikan sebagai negatif.

Dalam tahap deployment penyajian hasil, hasil dari proses pemodelan diinterpretasikan dan disajikan dengan cara yang relevan dan mudah dimengerti oleh pihak-pihak yang berkepentingan. Hal ini bertujuan agar hasil analisis machine learning dapat digunakan sebagai dasar untuk pengambilan keputusan yang lebih baik dan strategi bisnis yang efektif [2].

# Metode Penelitian

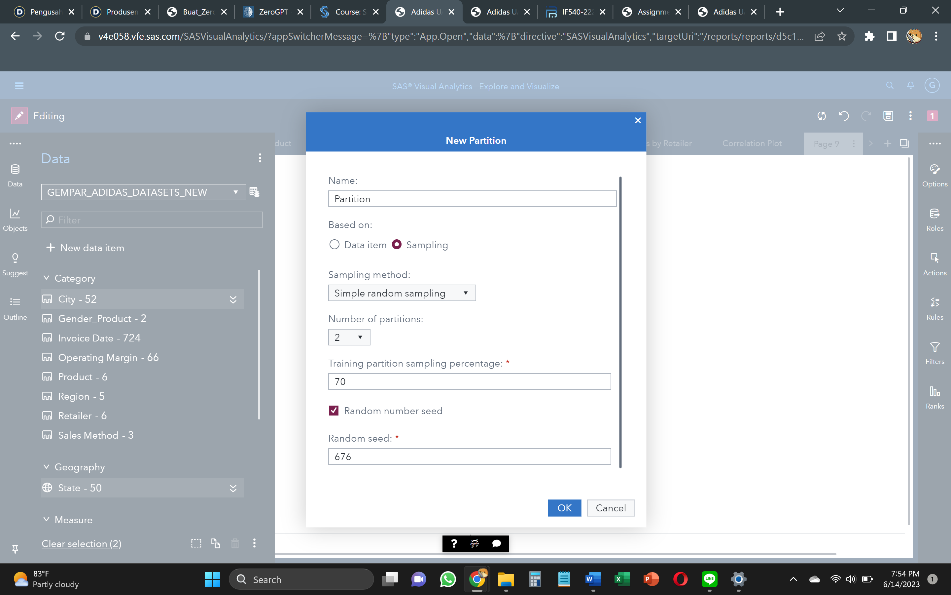
Penelitian data ini menggunakan metode analisa deskriptif dan kuantitatif yang mana hasil pengolahan data akan ditulis secara deduktif untuk memperoleh hasil penelitian yang dibutuhkan. Dengan menggunakan metode ini, fenomena dan permasalahan yang terjadi pada data penjualan Adidas dapat dijelaskan secara komperhensif dan mudah dipahami. Penelitian ini menggunakan perangkat lunak *online* SAS Visual Analytics untuk membuat visualisasi dari seluruh variabel penelitian dan membangun model prediksi dengan beberapa algoritma *machine learning*, yaitu Decision Tree C4.5, Random Forest, Logistic Regression, Bayesian Network, dan SVM.

Pada penelitian ini, peneliti menggunakan dataset sekunder penjualan produk Adidas dalam periode 2 tahun (Januari 2020 – Desember 2021) yang disediakan pada platform komunitas data scientist Kaggle <https://www.kaggle.com/datasets/heemalichaudhari/adidas-sales-dataset>. Data terdiri dari 9.469 baris dan 13 variabel sebagai berikut:

1. Tabel Dataset

|  |  |
| --- | --- |
| Atribut | Penjelasan |
| *Retailer* | Nama retail/pengecer dari produk Adidas |
| *Retailer ID* | Nomor unique retailer |
| *Invoice Date* | Tanggal produk terjual yang dicatat |
| *Region* | Pengelompokkan beberapa State |
| *State* | Nama negara bagian Amerika Serikat |
| *Price / Unit* | Harga setiap jenis produk |
| *Unit Sold* | Jumlah unit produk yang dijual per harinya |
| *Operating Profit* | Keuntungan bersih penjualan produk |
| *Total Sales* | Total penghasilan kotor dari penjualan |
| *Seles Method* | Metode penjualan retailer |
| *Product* | Jenis Product yang dijual |
| *Operating Margin* | Persentase selisih dari penjualan - keuntungan |
| *City* | Nama kota asal konsumen |
|  |  |

Kemudian dataset diunggah ke folder di CASUSER (gempar.bambang@student.umn.ac.id). Dataset asli memiliki format .csv, yang saya ubah nama dan formatya menjadi menjadi Gempar Adidas\_Dataset.xlsx. Adapun variabel target yang digunakan adalah RETAILER yang menunjukkan nama-nama retailer yang menjual kembali produk Adidas. Kemudian data dibagi menjadi data training dan data testing dengan rasio sebesar 7:3. Hasil pembagian ini didefiniskan sebagai Partition dengan jumlah random seed sebanyak 676 dan teknik sampling Simple Random Sampling.



*Gambar Pembuatan Partition Menggunakan SAS VA*

Penelitian ini menggunakan teknik data mining CRISP-DM (Vross Industry Standard Process-Data Mining). CRISP-DM sendiri terdiri dari enam tahapan yang menggambarkan siklus ilmu data science [2]. Adapaun enam tahapan yang diimplementasikan dalam penelitian ini adalah:

1. **Understanding Business Objectives**

Tahapan pertama akan melakukan tujuan bisnis dari Adidas tersebut, yakni ingin meningkatkan penjualan dengan cara memfokuskan penempatan dan promosi jenis-jenis produk Adidas yang laris dibeli di retailer sehingga pendapatan meningkat.

2. **Data Understanding**

Tahapan Data Understanding bertujuan untuk memahami data yang tersedia dari pencatatan penjualan internal Adidas. Tren dan anomali yang terjadi terhadap jenis produk yang sama juga dapat ditemukan setelah melakukan fase ini. Pemahaman didapat setelah melakukan visualisasi setiap variabel menggunakan beberapa grafik Exploratory Data Analysis.

3. **Data Preparation**

Tahapan Data Preparation ditujukan untuk memanipulasi data yang tersedia untuk mengadaptasinya ke dalam bentuk yang memungkinkan memperoleh hasil memadai dari model analisis ([5] ; [2]). Pada tahapan ini data akan diidentifikasi dan distandarisasi agar program SAS lebih mampu memahami data.

4. **Data Modelling**

Tujuan dari tahapan Data Modeling membangun hubungan antara variabel penjelas dan variabel yang diteliti, sehingga memungkinkan untuk menyimpulkan nilainya dengan tingkat kepercayaan tertentu [2]. Teknik yang dapat digunakan oleh Adidas dapat berupa metode statistik tradisional seperti neural networks based, decision trees, dan analisa calculated field.

5. **Evaluation**

Pada tahapan Evaluation, Big Data Analyst dapat menghitung kemampuan dari model yang telah dibangun. Model dapat digunakan kembali dengan sampel populasi yang baru (misalnya region Uni Eropa) untuk menguji accuracy, dan validasi tingkat kesalahan yang dihasilkan. Hasil klasifikasi variabel penjualan pun memberikan soslusi bisnis bagi Adidas berdasarkan penemuan tren dari data.

6. **Deployment**

Tahapan Deployment model analisis yang sudah dievaluasi dan dinyatakan layak untuk digunakan akan dijalankan secara konsisten dalam kegiatan operasional Adidas.

## Data Exploration

*Exploratory Data Analysis* (EDA) adalah metode untuk merangkum informasi yang bisa didapatkan dari data dengan menampilkan fitur penting secara logis. EDA perlu dilakukan sebelum data diolah pada data preparation dan data cleansing.

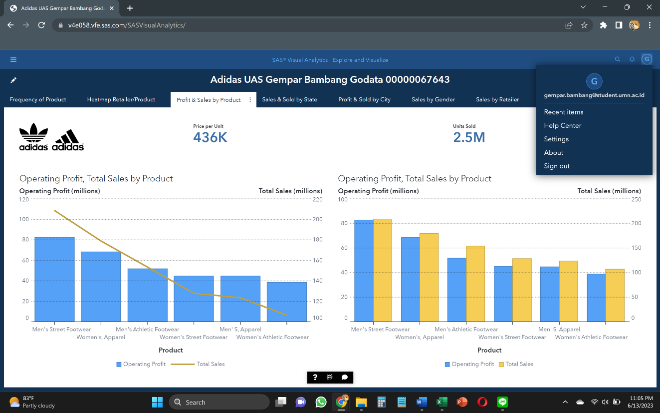


*Gambar Profile Dataset*

Dataset tidak memiliki data bernilai null sehingga tidak perlu melewati tahap data cleansing. Kemudian data akan menempuh tahap data preparation agar menjadikannya siap diolah ke dalam model machine learning. Prosedur pertama ialah standarisasi pada variabel Invoice Date, Sales Method, City, Region, State, Product, dan RetailerID. Prosedur preparasi berikutnya ialah memisahkan isi variabel Product menjadi Gender dan Type lalu Type akan dihapus.

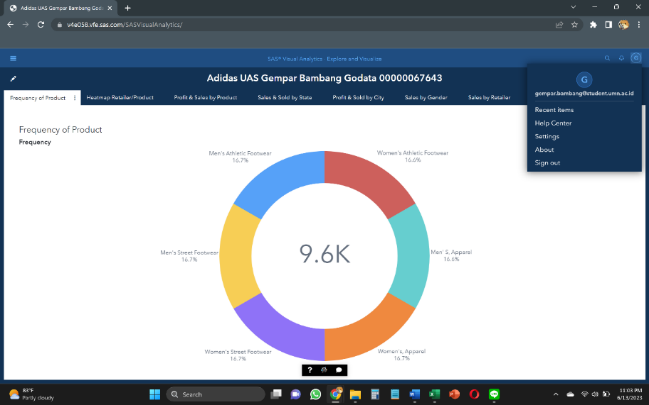
## Analisis Visualisasi Variabel

Beberapa variabel yang digunakan dalam analisis akan divisualisasikan melalui jenis-jenis grafik yang sesuai dengan tipe data dan kebutuhan penelitian.



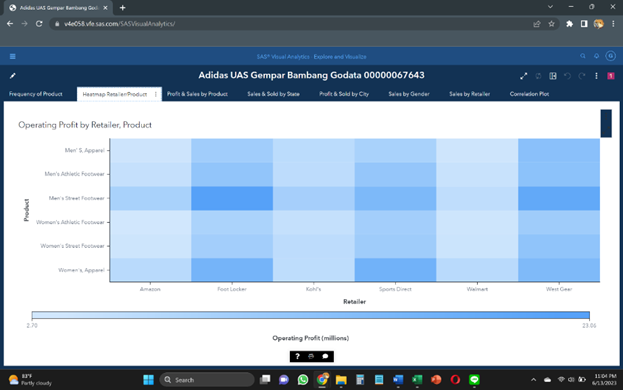
*Gambar Bar Chart Total Profit & Penjualan Berdasarkan Jenis Produk yang Dijual*

Gambar diatas menunjukkan dua judul grafik yang mana grafik di sebelah kiri memberikan informasi bahwa Men’s Street Footwear dan Women’s Apparel menjadi produk yang paling menghasilkan profit tertingi untuk Adidas. Grafik kiri juga memiliki diagram garis kuning yang menunjukkan total penjualan dari jenis produk. Terlihat bahwa Women’s Athlethic Footwear menjadi produk yang sangat sedikit dibeli oleh konsumen. Dengan informasi ini, Adidas dapat memprioritaskan sumber daya dan material untuk produk dengan popularitas lebih tinggi dan menurunkan produksi produk tidak laku.



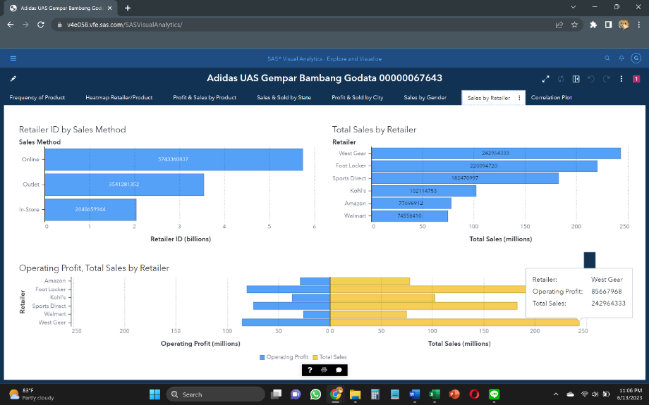
*Gambar Frekuensi Jenis Produk yang Berhasil Dijual*

Gambar di atas menunjukkan sebuah diagram pie dengan jumlah unit produk terjual di tengahnya. Grafik ini memberikan informasi bahwa data yang dikumpulkan telah membagi seluruh jenis produk dengan jumlah yang setara yakni 16,67%.



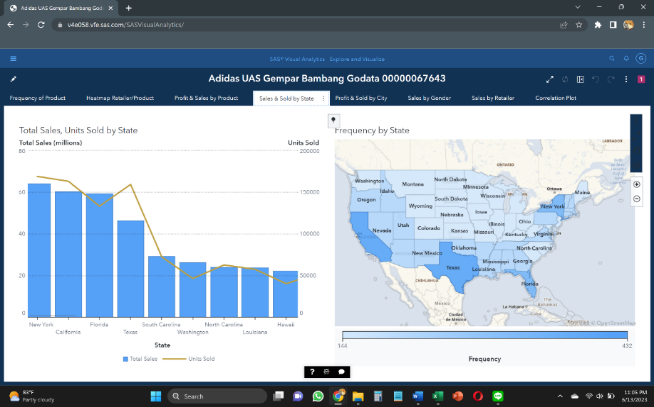
*Gambar Heatmap Tren Penjualan Jenis Produk di Retailer*

Gambar di atas menunjukkan sebuah grafik heatmap yang memberikan informasi mengenai tren jenis produk yang terjual di setiap retailer. Foot Locker, Sports Direct, dan West Gear menjadi retailer yang paling banyak menjual produk Adidas. Informasi ini memberikan gambaran mengenai produk Adidas akan lebih laku dijual di toko retail yang sangat mendukung tema dari produk Adidas, yakni pakaian olahraga.



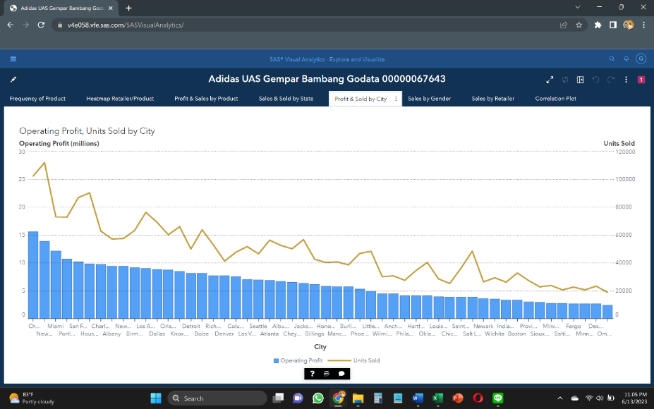
*Gambar Grafik Penjualan & Profit Berdasarkan Retailer*

Gambar di atas memiliki tiga buah grafik di dalamnya yang menunjukkan penjualan yang lebih lengkap berdasarkan retailer dan metode penjualan mereka. Terlihat bahwa profit yang dihasilkan memiliki nilai sepertiga dari total penjualan produk. Selain itu, informasi yang didapatkan dari grafik ialah penjualan melalui online lebih banyak dilakukan dibandingkan melalui metode lain.



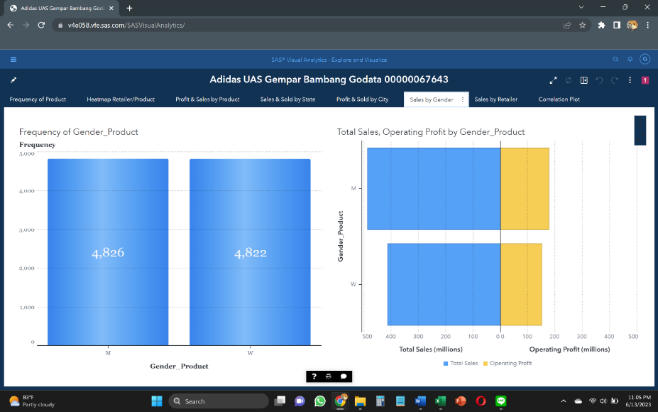
*Gambar Jumlah Penjualan Produk Berdasarkan State*

Gambar di atas memberikan informasi mengenai penjualan Adidas berdasarkan negara bagian AS. Terlihat bahwa negara bagian Texas, California, New York dan Florida menghasilkan pendapatan terbesar bagi Adidas. Sehingga didapatkan wawasan bahwa Adidas perlu memperluas target pasarnya di daerah berwarna biru muda melalui promosi rutin.



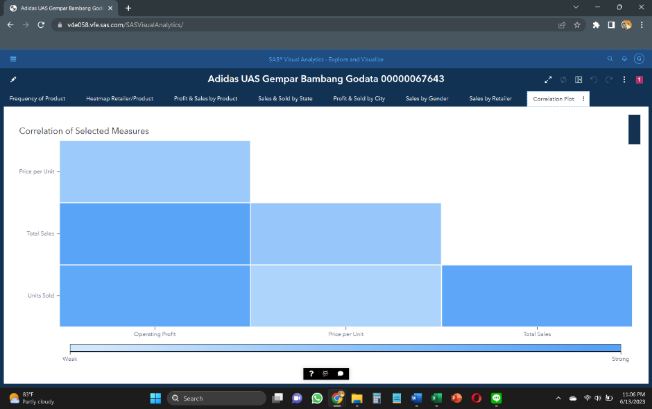
*Gambar Jumlah Keuntungan Produk Berdasarkan State*

Gambar di atas menunjukkan kota di Amerika Serikat yang menjadi konsumen produk Adidas dan jumlah unit produk yang terjual. Informasi yang didapatkan ialah Chicago, New York, dan Miami menempati peringkat tertinggi keuntungan. Hal ini dapat disebabkan oleh ketiga kota tersebut merupakan penduduk bermayoritas kalangan atas.



*Gambar Penjualan Produk & Profit Berdasarkan Gender*

Gambar di atas menunjukkan jumlah produk Adidas yang terjual berdasarkan gender dari produk tersebut. Data yang ada menunjukkan jumlah produk relatif sama yang kemungkinan besar sudah diatur oleh penyedia data. Namun informasi yang didapatkan ialah harga produk pria lebih mahal di bandingkan produk wanita dari grafik kiri. Hal ini menunjukkan pria menjadi target pasar utama dari Adidas yang hobi dengan kegiatan olahraga.



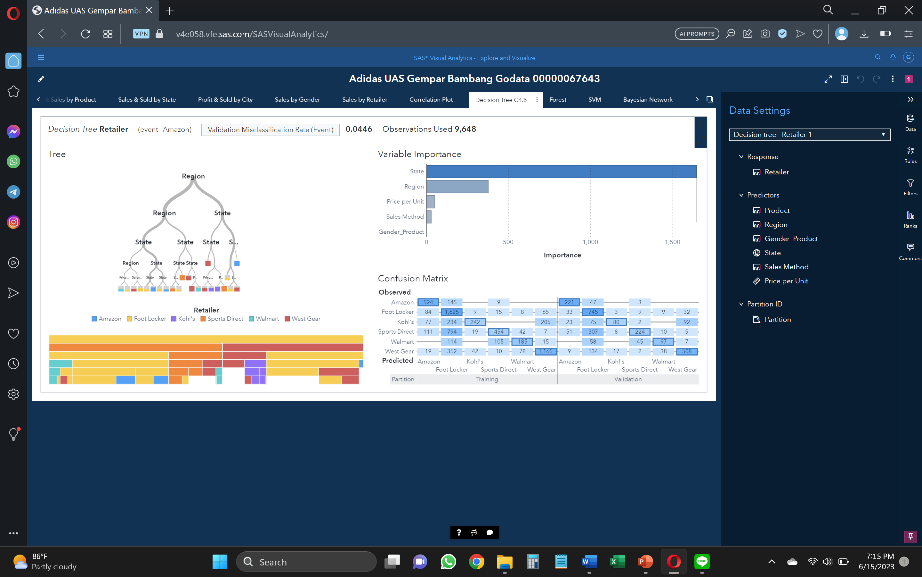
*Gambar Grafik Penjualan & Profit Berdasarkan Retailer*

Gambar di atas menunjukkan visualisasi dari analisis matriks korelasi di antara ketiga variabel numerik penjualan Adidas. Terlihat bahwa harga per unit produk sedikit mempengaruhi jumlah keuntungan dan penjualan dari Adidas.

# Hasil dan Pembahasan

Penelitian ini menjadikan tahapan Data Modeling untuk menghasilkan klasifikasi data berdasarkan variabel Retailer sebagai pusat target penelitian. Sehingga pola-pola tren yang dimiliki data dapat ditemukan, diidentifikasi, dan dimanfaatkan oleh model machine learning yang dibangun.

## Decision Tree C4.5

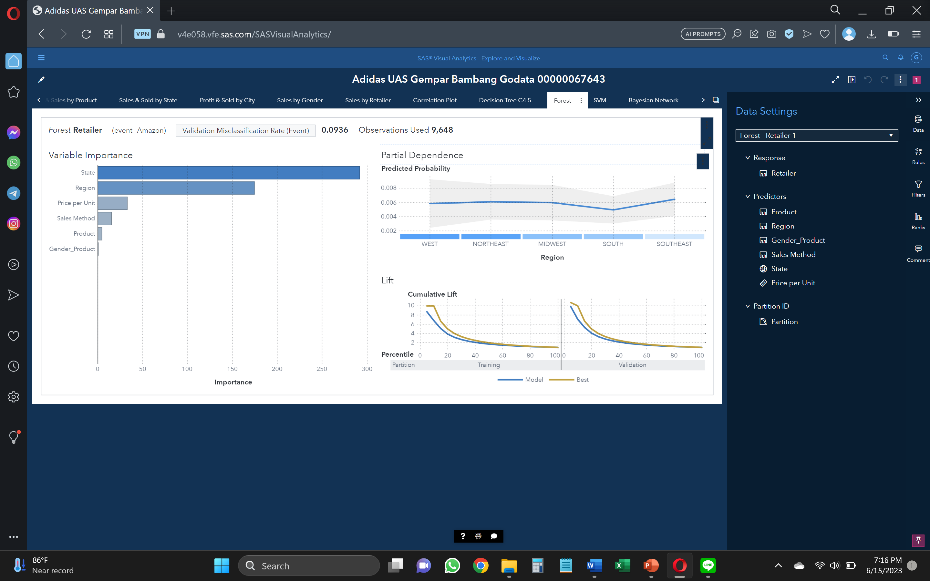


*Gambar Hasil Pemodelan Decision Tree C4.5*

Model Decision Tree menjadi model yang mudah dipahami karena memiliki penalaran seperti cabang pohon “Yes or No”. Variabel Retailer menjadi target dari para predictor yang akan ditentukan untuk semua model. Yaitu Product, Region, Price per Unit, Sales Method dan Gender Product. Confolusion Matrix juga dibagi menjadi Training dan Validation Test yang tidak memiliki perbedaan berarti.

Hasil klasifikasi pun menunjukkan variabel terpenting dalam mengklasifikasikan Retailer yang cocok bagi sebuah produk ialah State dan Region di mana produk Adidas akan dijual. Sebanyak 9.648 baris data berhasil diklasifikasi dengan tingkat kesalahan (misclassification rate) sebesar 4,46%.

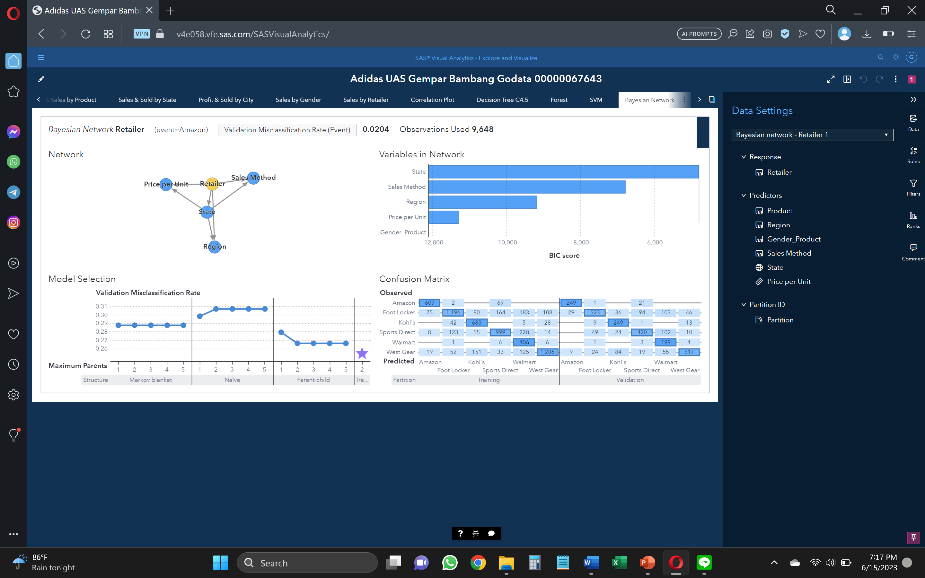
## Forest



*Gambar Hasil Pemodelan Forest*

Model Forest yang dibangun berbasis algoritma Random Forest. Model ini menghasilkan variabel State dan Region yang menjadi fitur terpenting dalam mengklasifikasikan produk yang tepat di retailer. Diagram Cumulative Lift memberikan informasi bahwa seluruh target populasi telah dianalisa menjadi Best Segmentation. Semakin mendekati nol nilai akhir Lift, maka model akan semakin baik dalam membedakan kelompok-kelompok yang telah diklasifikasi. Adapun algoritma Forest menghasilkan tingkat kesalahan cukup tinggo dalam mengklasifikasi sebesar 9,54%.

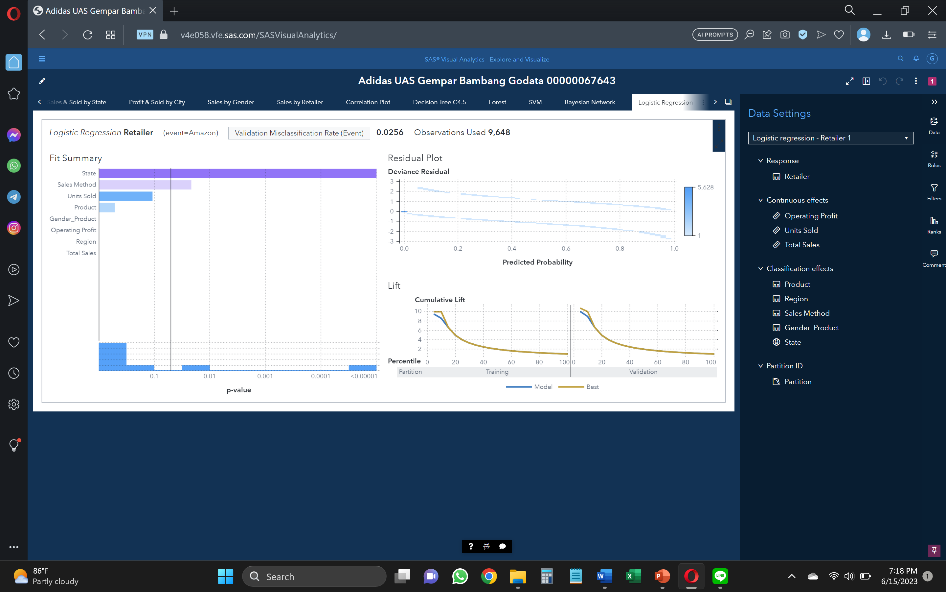
## Bayesian Network



*Gambar Hasil Pemodelan Bayesian Network*

Dalam SAS, pemodelan Bayesian Network menggunakan empat struktur Bayesian, yakni Naïve Tree-augmented naïve, Parent-child, dan markov blanket. Model Bayesian Network yang dibagun dengan predictor yang sama menghasilkan tingkat kesalahan dalam klasifikasi data sebesar 2,04%. State dan Sales Method menjadi variabel yang paling berpengaruh dalam proses pengklasifikasian data terhadap retailer.

## Logistic Regression

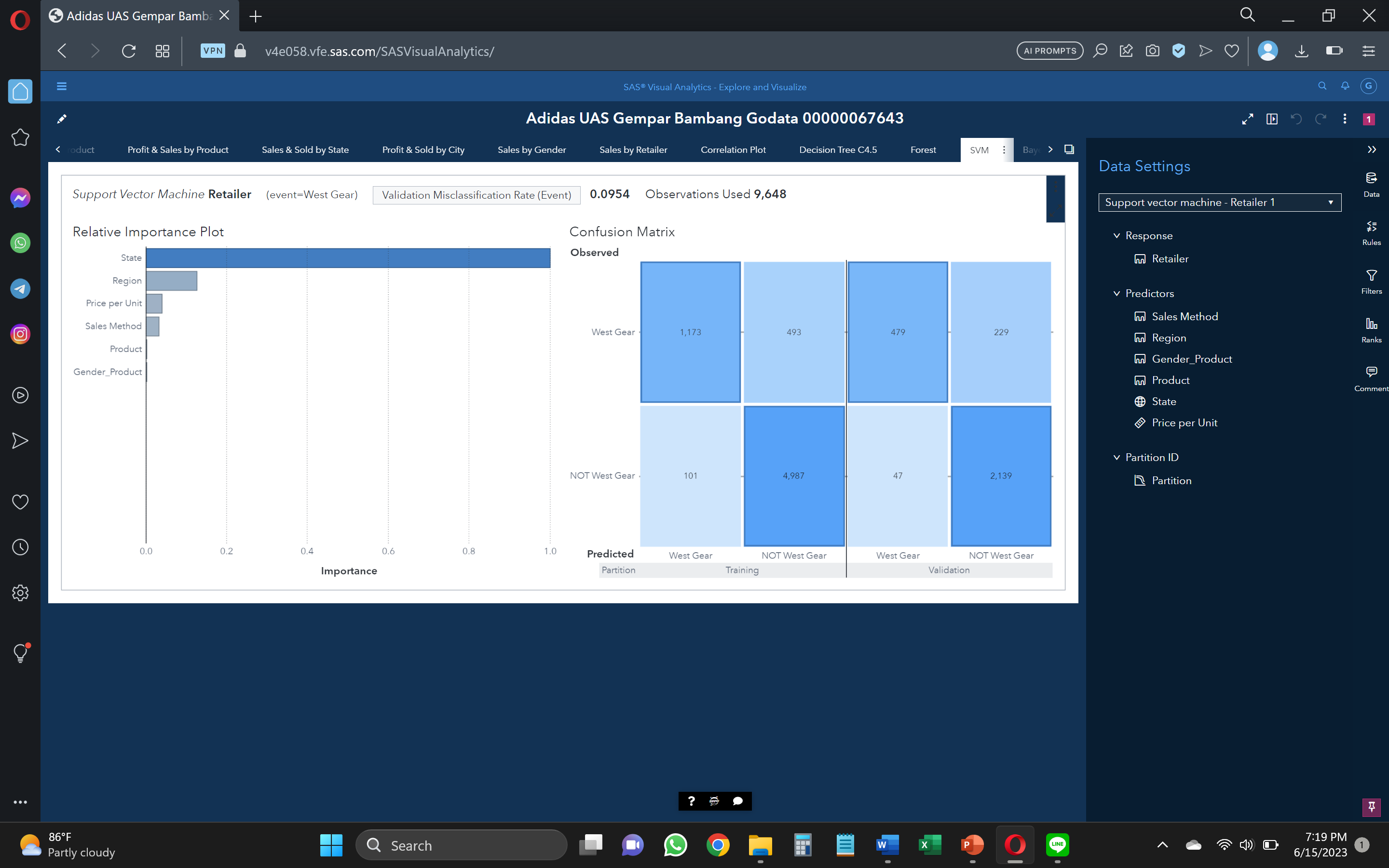


*Gambar Hasil Pemodelan Logistic Regression*

Pemodelan Logistic Regression terhadap variabel target Retailer menggunakan Continous effects berupa Operating Profit, Unit Sold, dan Total Sales untuk menunjukkan hasil klasifikasi akan mempengaruhi nilai dari efek-efek ini. Misclassification Rate yang dihasilkan saat melakukan validation ialah sebesar 2,56%, atau hampir setara dengan penggunaan Bayesian network. Akurasi yang tinggi ini disebabkan data tidak memiliki nilai null yang biasanya akan dibuang oleh algoritma apabila ditemukan.

Informasi berikutnya yang didapat dari grafik di atas ialah variabel State menjadi variabel paling kuat yang jauh mengungguli variabel lain sebagai pengaruh terhadap variabel continous effects. Nilai probabilitas/ signifikansi asimtoik dari model statistik (p-value) ini didominasi oleh nilai sebesar 0.1 yakni sebanyak 62% dari keseluruhan. Nilai sebesar ini menandakan hipotesis yang dibuat algoritma terbukti oleh validation.

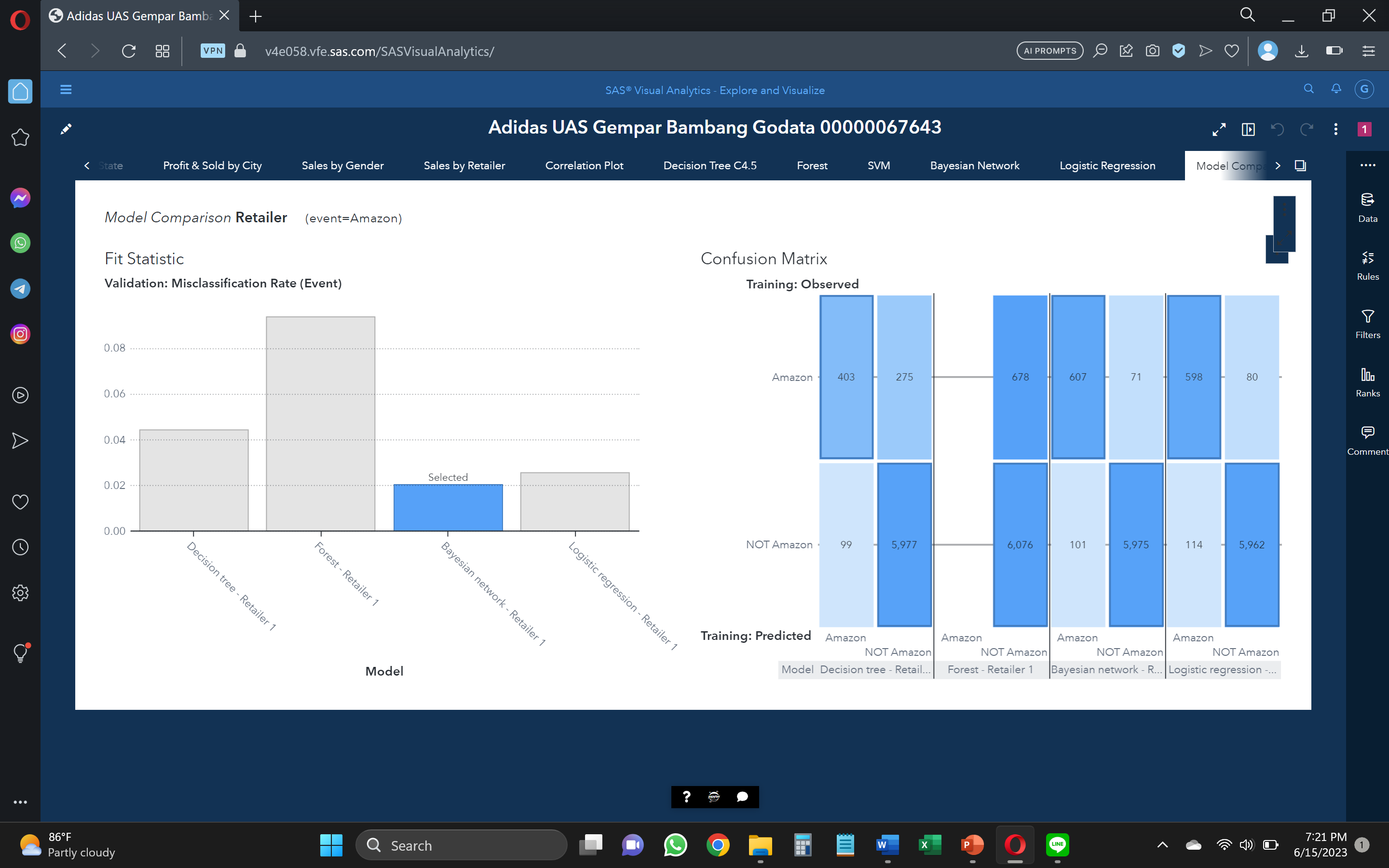
## Support Vector Machine



*Gambar Hasil Pemodelan SVM*

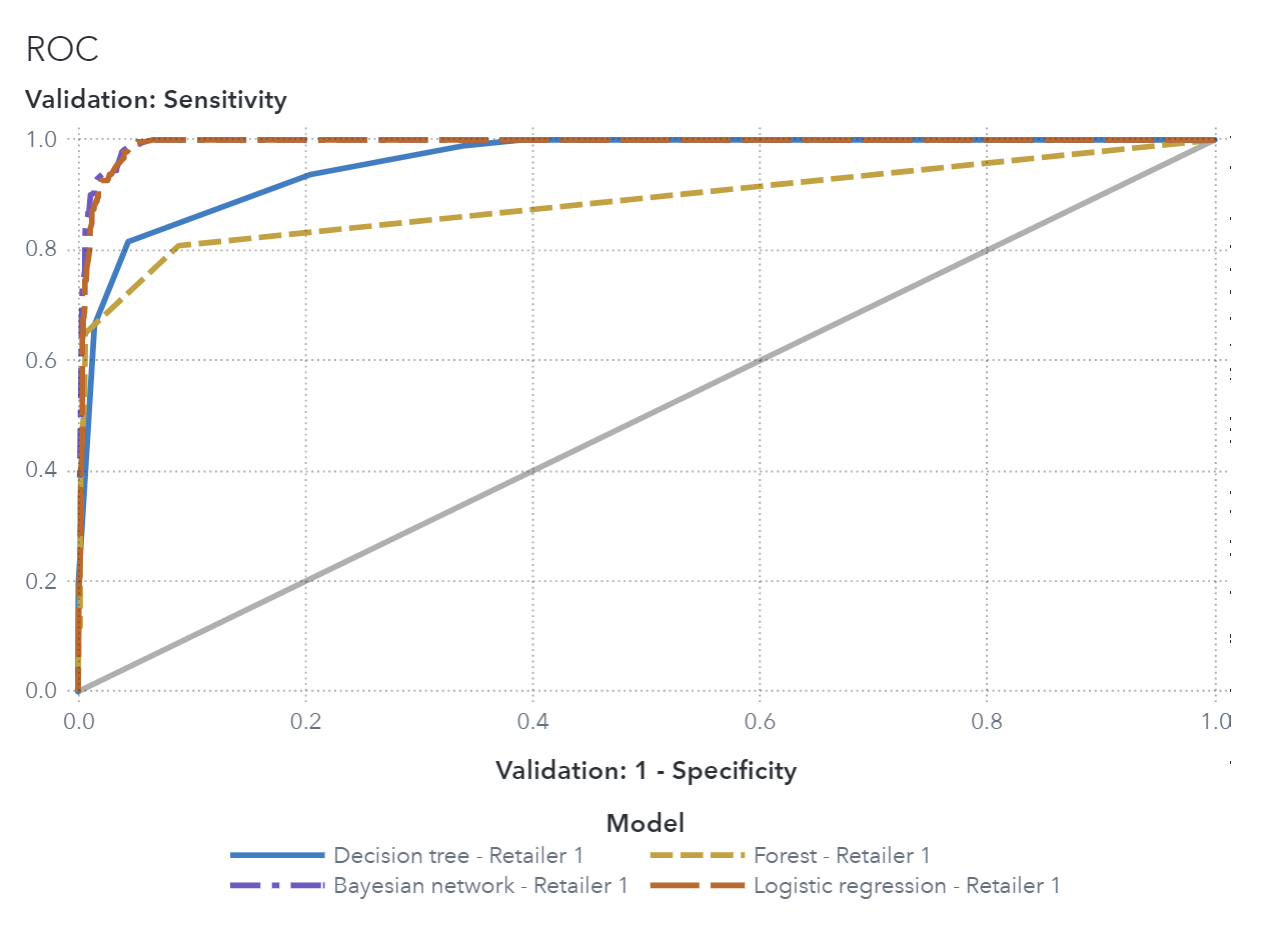
Pemodelan Support Vector Machine (SVM) yang dibagun menghasilkan misclassification rate sebesar 9,54% atau sangat rendah dibandingkan algoritma model lainnya. Contohnya saat mengklasifikasikan produk berdasarkan predictor yang sama, produk yang dijual oleh West Gear diklasifikasi menjadi Not West Gear dengan cukup sering (FP = 493). Adapun variabel yang paling berpengaruh kuat menurut model SVM ialah State retailer sebagai satu-satunya.

## Model Selection and Comparison



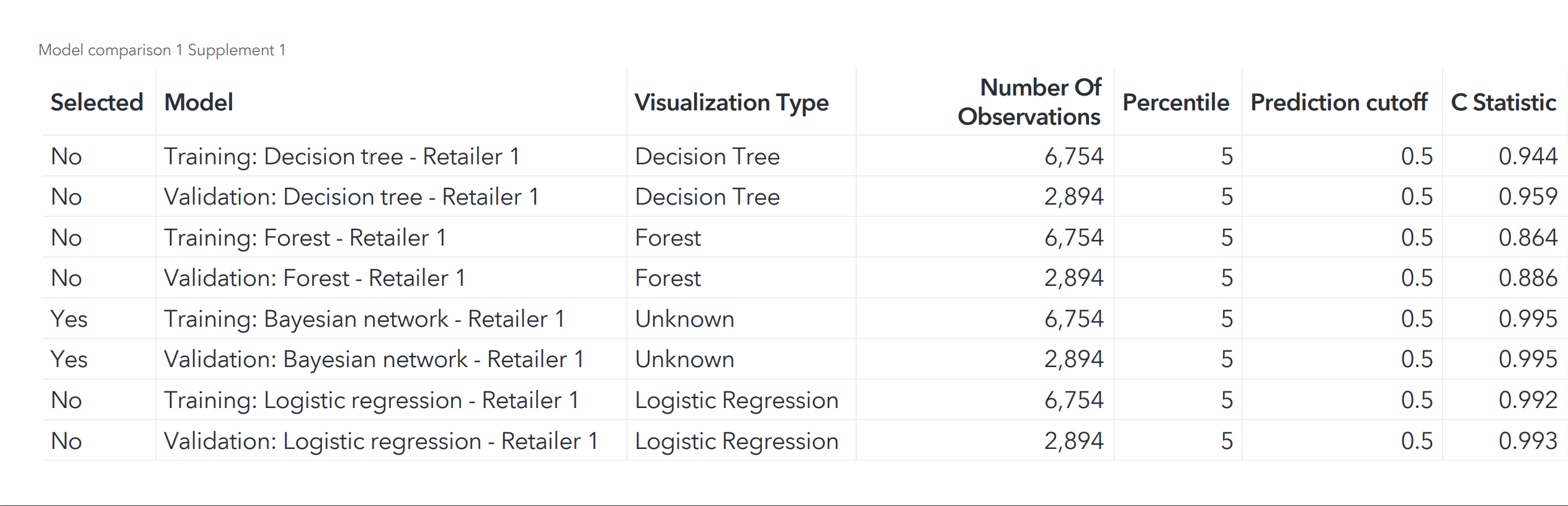
*Gambar Model Selection and Comparison*

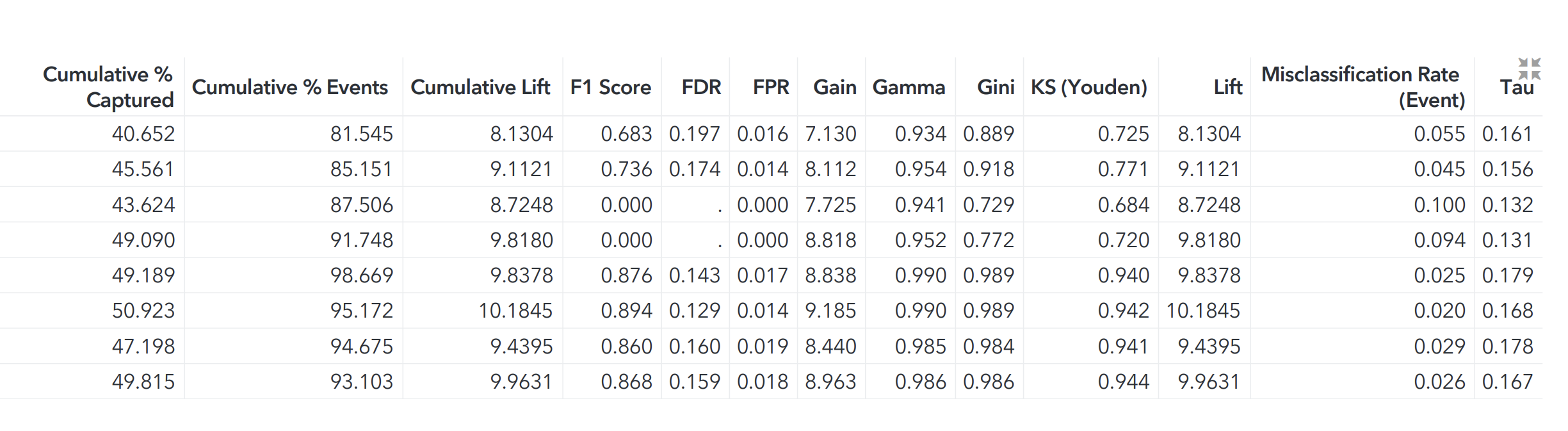
Menggunakan fitur SAS Visual Statistics, empat model yang setara dapat dibandingkan oleh Model Comparison. Keempat model machine learning yang telah dibangun dan diuji, yaitu Decision tree C4.5, Forest, Bayesian Network, dan Logistic Regression telah sesuai dengan kriteria dalam melakukan klasifikasi produk yang tepat dijual oleh suatu retailer. Faktor level misclassification rate (event) yang rendah dipilih sebagai penentu pemilihan model terbaik.

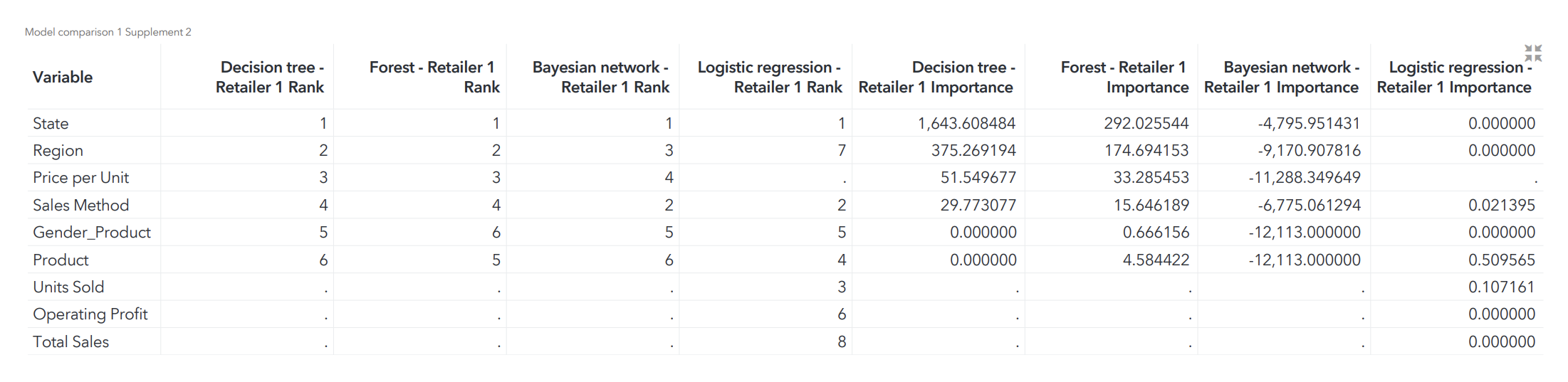
**

*Gambar Grafik ROC Sensitivitas Model*

Oleh karena itu, model Bayesian Network terpilih sebagai algoritma terbaik. Selain disebabkan akuratnya model dalam mengklasifikasikan produk tren Amazon, model juga memiliki sensitivitas terbaik berdasarkan diagram Receiver Operator Characteristic (ROC). Jumlah ketepatan dalam mengklasifikasikan data True Postive dan False Positive hampir senilai 1, yang berarti sangatlah akurat dan dapat digunakan oleh Adidas.







*Gambar Tabel Detail Comparison Setiap Model*

# Kesimpulan dan Saran

Dengan pemulihan kembali permintaan produk industri tekstil, maka Adidas juga membutuhkan sebanyak mungikin wawasan terbaru untuk menyiapkan strategi paling efektif untuk menguasai pasar dengan kondisi ekonomi konsumen belum pulih sepenuhnya. Salah satu strategi yang dapat dilakukan oleh Adidas ialah mengalokasikan stok produk ke retailer yang mampu menjualnya dengan cepat dan untung. Wawasan untuk mengembangkan strategi ini dapat melalui sebuah program berbasis machine learning yang dapat menganalisis dan mengklasifikasikan tren produk yang dijual oleh para retailer produk bermerek Adidas.

Penelitian ini menghasilkan lima model yang mampu menganalisis dan mengklasifikasikan tren produk Adidas berdasarkan Retailer. Model tersebut ialah Decision Tree C4.5, Random Forest, Bayesian Network, Logistic Regression, dan Support Vector Machine. Model Bayesian Network terpilih menjadi model terbaik dengan Misclassification Rate yang terendah dari semua model. Bayesian Network memiliki rate sebesar 2,04%, Logistic Regression sebesar 2,56%, Decision Tree C4.5 sebesar 4,46%, Random Forest sebesar 9,36%, dan SVM tidak dapat dibandingkan dengan rate sebesar 9,54%. Oleh karena itu algoritma model Bayesian Network terpilih sebagai model yang paling efektif dalam mengklasifikasikan produk Adidas yang tepat dijual di retailer terlepas dari kriteria spesifik lain yang apabila ingin digunakan oleh Adidas.

Hasil analisis dan klasifikasi model yang telah dibangun memungkinkan untuk direkayasa ulang maupun dikembangkan oleh penelitian selanjutnya. Saran yang dapat diberikan penulis artikel ini ialah dibutuhkannya data penjualan Adidas yang lebih kompleks untuk negara-negara lain yang memiliki perekonomian berbeda seperti di Amerika Serikat. Metode lain untuk dibandingkan dengan hasil penelitian ini juga memerlukan analisis mengenai perkembangan SAS di masa depan.

# Daftar Referensi

1. Nadya Zahira*.* (2023, Mei 15). “Katadata.co.id: Produsen Adidas Dikabarkan Lakukan PHK Sepihak, begini Kata Manajemen”. <https://katadata.co.id/ira/berita/646188f9db5a1/produsen-adidas-dikabarkan-lakukan-phk-sepihak-begini-kata-manajemen>
2. Forradelas, R.R., Alonso, S. L. N., Vázquez, J. J. (2020) ‘Applied Machine Learning in Social Sciences: Neural Networks and Crime Prediction’, Social Sciences 2021, 10. 4. Available <https://www.researchgate.net/publication/347974941_Applied_Machine_Learning_in_Social_Sciences_Neural_Networks_and_Crime_Prediction>
3. Horowitz, B. M. (2019) ‘Chapter 5 - Policy Issues Regarding Implementations of Cyber Attack: Resilience Solutions for Cyber Physical Systems’, Artificial Intelligence for the Internet of Everything, pp 87-100. DOI: <https://doi.org/10.1016/B978-0-12-817636-8.00005-3>
4. I. K. Sahoo, A. Samal, J. Pramanik, & S. Pani, (2019). Exploratory Data Analysis using Python. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8(12), 47-67. <https://doi.org/10.35940/ijitee.L3591.1081219>
5. S. Krishnan, M. Franklin, K. Goldberg, J. Wang, and E. Wu. (2016) ‘Activeclean: An interactive data cleaning framework for modern machine learning’. Paper presented at the 2016 International Conference on Management of Data, San Francisco, June 26–July 1; pp. 2117–20. Available: <https://arxiv.org/pdf/1601.03797.pdf>
6. W. Y. Ayele, “Adapting CRISP-DM for Idea Mining: A Data Mining Process fro Generating Ideas Using a Textual Dataset,” nternational Journal of Advanced Computer Science and Applications. Available <https://arxiv.org/abs/2105.00574v1>
7. S. G. C, V. S. T, R. V. Gowda, P. R. Udupa, and S. Reddy, “A machine learning classification approach for detection of covid 19 using ct images,” EMITTER International Journal of Engineering Technology, vol. 10, pp. 183–194, 2022
8. W. Willy, D. P. Rini, and S. Samsuryadi, “Perbandingan algoritma random forest classifier, support vector machine dan logistic regression clasifier pada masalah high dimension (studi kasus: Klasifikasi fake news),” JURNAL MEDIA INFORMATIKA BUDIDARMA, vol. 5, p. 1720, 10 2021.
9. N. Wuryani, S. Agustiani, I. Komputer, and N. Mandiri, “Random forest classifier untuk deteksi penderita covid-19 berbasis citra ct scan,” Jurnal Teknik Komputer AMIK BSI, vol. 7, 2021. [Online]. <http://ejournal.bsi.ac.id/ejurnal/index.php/jtk>
10. N. Wuryani, S. Agustiani, I. Komputer, and N. Mandiri, “Random forest classifier untuk deteksi penderita covid-19 berbasis citra ct scan,” Jurnal Teknik Komputer AMIK BSI, vol. 7, 2021. Available: <http://ejournal.bsi.ac.id/ejurnal/index.php/jtk>
11. N. Amini and A. Shalbaf, “Automatic classification of severity of covid19 patients using texture feature and random forest based on computed tomography images,” International Journal of Imaging Systems and Technology, vol. 32, pp. 102–110, 1 2022
12. F. Wang and J. Amrhein, “Bayesian Networks for Causal Analysis,” McDougall Scientific Ltd Paper 2776-2018. <https://support.sas.com/resources/papers/proceedings18/2776-2018.pdf>
13. M. Misdram, F. Syarifuddin, A. W. SPIRIT, and undefined 2020, “Klasifikasi data set virus corona menggunakan metode na¨ıve bayes classifier,” jurnal.stmik-yadika.ac.id, vol. 5, pp. 964–972, 2021. [Online]. Available: <https://www.jurnal.stmik-yadika.ac.id/index.php/spirit/article/view/184>
14. N. Nadiah, S. Soim, and S. Sholihin, “Implementation of decision tree algorithm machine learning in detecting covid-19 virus patients using public datasets,” Indonesian Journal of Artificial Intelligence and Data Mining, vol. 5, pp. 37–43, 6 2022
15. I. Budiman, R. Ramadina et al., “Penerapan fungsi data mining klasifikasi untuk prediksi masa studi mahasiswa tepat waktu pada sistem informasi akademik perguruan tinggi,” JUPITER (Jurnal Penelitian Ilmu dan Teknologi Komputer), vol. 7, no. 1, pp. 39–50, 2015
16. Y. So, “A Tutorial on Logistic Regression,” SAS Institute Inc. <https://support.sas.com/rnd/app/stat/papers/logistic.pdf>

# Lampiran

Tautan SAS Visual Analytics Report:

<https://v4e058.vfe.sas.com/links/resources/report?uri=%2Freports%2Freports%2Fd5c12e4c-8ed5-44bf-8d11-630bea7d7ef6>