UAS SEMESTER GENAP 2023/2024

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS TEKNIK DAN INFORMATIKA

UNIVERSITAS MULTIMEDIA



**Kenny Budiarso Lawson**

**00000081065**

Mata Kuliah:

Big Data Analytics

IS429 - AL

Dosen Pengampu: Iwan Prasetiawan, S. Kom., M. M.

**Ujian Akhir Semester Big Data Analytics (Praktikum)**

Link UAS BDA IS429 Theory: …

Repository Dataset: “Loan\_default.csv”, <https://www.kaggle.com/datasets/nikhil1e9/loan-default>

Bidang Bisnis : Banking

Nama Perusahaan : Lawson Bank

Motto Perusahaan : Exclusive, Trusted, and Friendly

A logo for a bank

Description automatically generated

Lawson Bank adalah sebuah lembaga keuangan yang bergerak di bidang perbankan dengan fokus pada memberikan layanan eksklusif, terpercaya, dan ramah kepada nasabahnya. Sebagai bagian integral dari industri perbankan, Lawson Bank berkomitmen untuk memberikan solusi keuangan yang inovatif dan berkelanjutan kepada para nasabahnya. Dengan motto "Exclusive, Trusted, and Friendly", Lawson Bank menempatkan keunggulan pelayanan dan kepercayaan sebagai prioritas utama dalam setiap interaksi dengan nasabahnya.

Dibangun dengan fondasi integritas dan transparansi, Lawson Bank menawarkan beragam produk dan layanan perbankan yang dirancang untuk memenuhi kebutuhan finansial individu maupun bisnis. Kami menempatkan keamanan dan kepuasan nasabah sebagai fokus utama, dengan berkomitmen untuk memberikan pengalaman perbankan yang aman, nyaman, dan menguntungkan bagi setiap nasabah.

Sebagai bagian dari visi dan misi kami, Lawson Bank tidak hanya bertujuan untuk menjadi mitra keuangan yang handal bagi nasabah, tetapi juga menjadi pilar dalam membangun masa depan finansial yang kokoh dan berkelanjutan bagi masyarakat. Kami percaya bahwa dengan pendekatan yang inklusif, inovatif, dan ramah, kami dapat menjadi lebih dari sekadar sebuah bank, tetapi mitra sejati dalam perjalanan keuangan Anda.

1. **Question 1**

* Manage your **Data Preparation** and **Exploration Process** corresponding to the data curation process.
* The journal writing in section III contains an explanation of EDA (Exploratory Data Analysis)
* Create your **EDA** and **Data preparation process** with the SAS Data Preparation application.

1. **Data Managing & Preparation**

Dataset yang digunakan pada studi kasus ini adalah data eksternal dari kaggle.com berjudul “Loan\_default.csv”, dengan link halaman website tersebut, yaitu: “<https://www.kaggle.com/datasets/nikhil1e9/loan-default>”. Berikut file dan tampilan csv file dataset tersebut:



A screenshot of a computer

Description automatically generated

*(Tampilan file “Loan\_default.csv” pada Microsoft Excel)*

* **Import Dataset ke Casuser – SAS**

File dataset tersebut saya masukkan ke dalam Casuser pada SAS Visual Analytics menggunakan SAS Studio - Develop SAS Code dengan menempatkannya pada library WORK.IMPORT.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

*(Import Dataset ke Casuser)*

A screenshot of a computer

Description automatically generated

*(Dataset Loan\_default sudah tersedia pada data source SAS)*

Terdapat informasi dari dataset yang saya gunakan.Lokasi dataset berada pada [CASUSER/kenny.budiarso@student.umn.ac.id](mailto:CASUSER/kenny.budiarso@student.umn.ac.id). Dataset “Loan\_default” terdiri dari 18 kolom (variabel) dan 255.3 ribu baris (nilai data). Berikut adalah informasi mengenai kolom-kolom (variabel) pada dataset tersebut:

A screenshot of a computer

Description automatically generated

*(Informasi Kolom dan tipe data pada dataset “Loan\_default”)*

Setelah dataset di-import ke dalam program SAS, Langkah selanjutnya adalah mengecek informasi dataset yang saya gunakan. Pengecekkan informasi meliputi informasi data unik dan missing value (null), serta informasi statistik dari tiap-tiap variabel pada dataset (Mean, Median, Modus, Standard Deviation, Standard Error, Min, Max, dan lain-lain).

A screenshot of a computer

Description automatically generated

*(Data Profile – Informasi dataset)*

Pada dataset tersebut, dapat dilihat bahwa tidak ada nilai null/missing value pada tiap variabel. Yang menandakan dataset bersih dari Missing Value.

* **Remove Duplicates**

Pada plan Data Preparation, saya melakukan Remove Duplicates pada kolom “LoanID”, guna memastikan bahwa setiap baris pada dataframe tersebut tidak ada yang ter-duplikat. Sehingga, hasil pemrosesan data berikutnya dapat seakurat mungkin.

A screenshot of a computer

Description automatically generated

*(Proses Remove Duplicates)*

* **Encoding**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

*(Encoding kolom “HasMortgage” menjadi “ENC\_HasMortgage”)*

Tahap selanjutnya adalah melakukan encoding pada kolom-kolom kategorikal dengan nilai “Yes” dan “No”, agar dapat lebih mudah dikelaskan dan dibuat visualisasi kategorikal oleh sistem. Contoh pertama pada gambar di atas adalah proses melakukan One Hot Encoding pada kolom “HasMortgage” yang menunjukkan apakah nasabah memiliki ha katas properti, seperti rumah atau apartemen atau bangunan pribadi lainnya. Nilai pada encoding otomatis diberikan oleh sistem dengan nilai “No” = 1 dan “Yes” = 2. Selain pada kolom “HasMortgage”, proses encoding juga dilakukan pada kolom “HasDependents” dan “HasCoSigner” dengan hasil pada kedua gambar di bawah:

A screenshot of a computer

Description automatically generated

*(Encoding kolom “HasDependents” menjadi “ENC\_HasDependents”)*

A screenshot of a computer

Description automatically generated

*(Encoding kolom “HasCoSigner” menjadi “ENC\_HasCoSigner”)*

1. **Data Exploration**

Data Exploration dilakukan pada dataframe yang sudah diproses dengan Data Curation dan dalam Data Preparation. Berikut adalah tampilan data “LOAN\_DEFAULT\_1” yang sudah dalam keadaaan bersih dan prepared:

A screenshot of a computer

Description automatically generated

*(Tampilan Data yang sudah prepared)*

* + **Distribusi Frekuensi Umur Nasabah**

A screenshot of a computer

Description automatically generated

*(Visualisasi Distribusi Frekuensi Umur Nasabah)*

Eksplorasi data yang pertama yaitu visualisasi distribusi dari frekuensi umur para nasabah Lawson Bank. Pada visualisasi di atas, terdapat 2 chart, yaitu chart total frekuensi tiap kelompok umur dan persentase frekuensi tiap kelompok umur dengan pengelompokkan lebih kecil.

Chart pertama menunjukkan bahwa frekuensi tiap kelompok umur cenderung memiliki distribusi frekuensi yang rata atau setara, namun dengan jeda antar kelompok umur yang cukup banyak. Chart ini menunjukkan bahwa rentang kelompok umur memiliki jeda yang cukup banyak, dengan frekuensi tiap kelompok umur terdistribusi dengan baik atau mendekati normal.

Chart kedua menunjukkan persentase frekuensi kelompok umur dengan dibagi ke dalam 10 pengelompokkan. Pada chart ini, distribusi persentase umur terlihat lebih mendekati setara antara satu kelompok dengan lainnya, kecuali pada tiap ujung bagian, yaitu kelompok umur 18 - 23.1 dan 63.9 – 69yang memiliki persentase sedikit lebih banyak dibanding kelompok lainnya.

Kedua chart ini menunjukkan bahwa distribusi program pinjaman dana oleh Lawson Bank tersebar secara menyeluruh dan merata secara pengelompokkan umur. Dengan begitu, Lawson Bank terbukti berhasil menyediakan langganan peminjaman dana secara merata ke seluruh bagian masyarakat, baik muda maupun berumur lanjut.

* **Frekuensi Tiap Tipe Employment, Education, dan Marital Status**

A screenshot of a computer

Description automatically generated

(Visualisasi Frekuensi Tiap Tipe Employment, Education, dan Marital Status)

Visualisasi kedua menunjukkan hasil eksplorasi beberapa variabel yang diamati, yaitu EmploymentType (Tipe Pekerjaan), Education (Tingkat Edukasi tertinggi/terakhir), dan MaritalStatus (Status Pernikahan).

Dari chart pertama terlihat bahwa jumlah frekuensi tiap tipe pekerjaan cenderung setara di kisaran 63 – 64 ribu orang. Menunjukkan bahwa distribusi frekuensi tiap tipe pekerjaan terdistribusi secara merata.

Chart kedua menunjukkan persentase frekuensi nasabah tiap tingkat edukasi terakhir. Chart tersebut menunjukkan bahwa frekuensi nasabah tiap tipe pekerjaan terdistribusi dengan baik pula dengan perbedaan satu sama lain yang tidak signifikan, yaitu di kisaran persentase 24.9 % - 25.2 %.

Chart ketiga menunjukkan frekuensi nasabah berdasarkan status pernikahan yang terdistribusi dengan baik pula. Dimana frekuensi tiap status pernikahan memiliki nilai dengan selisih sangat sedikit. Seluruh frekuensi tiap status pernikahan berada di kisaran angka 85 ribu nasabah.

Dari visualisasiini, dapat disimpulkan bahwa program peminjaman dana oleh Lawson Bank sudah tersebar secara merata ke seluruh masyarakat walaupun dengan tipe pekerjaan, edukasi terakhir, dan status pernikahan nasabah.

* **Perbandingan Income dengan Loan Amount**

A screenshot of a computer screen

Description automatically generated

*(Visualisasi perbandingan antara nilai rata-rata pendapatan nasabah dengan rata-rata pinjaman per tipe pekerjaan)*

Visualisasi selanjutnya berisikan bar chart yang menunjukkan nilai rata-rata pendapatan dibandingkan dengan nilai rata-rata pinjaman per tipe pekerjaan. Dapat dilihat bahwa pada seluruh tipe pekerjaan, para nasabah cenderung mengambil pinjaman dengan nilai lebih tinggi dibanding pendapatannya. Hal ini dapat terjadi karena besarnya kebutuhan nasabah dan didukung dengan sistem cicilan pinjaman yang dapat sampai 60 bulan, sehingga nasabah merasa tidak terlalu berat dengan kewajiban pinjaman tersebut walau nilainya melebihi pendapatannya. Untuk nilai pada chart di atas, terlihat pula nilai rata-rata pendapatan dan nilai rata-rata besaran pinjaman pada masing-masing tipe pekerjaan terdistribusi merata atau hamper setara, dengan rata-rata nilai pendapatan berada di kisaran angka 82 ribu dan rata-rata pinjaman di kisaran angka 127 ribu.

* **Jumlah Credit Tiap Tipe Employment, serta Frekuensi Kepemilikan Mortgage, Dependents, dan CoSigner**

A screenshot of a computer

Description automatically generated

*(Visualisasi jumlah kredit tiap tipe pekerjaan, frekuensi kepemilikikan hak milik bangunan, tanggungan, dan wakil)*

Visualisasi selankutnya terdapat empat bar chart. Bar chart pertama berisikan informasi rata-rata jumlah kredit yang dimiliki tiap tipe pekerjaan. Dpat dilihat bahwa tiap tipe pekerjaan memiliki rata-rata jumlah kredit yang mirip, yaitu di angka 2.50. Hal ini berarti setiap orang rata-rata memiliki dua atau lebih kredit atau pinjaman yang mereka tanggung.

Bar chart kedua menampilkan informasi frekuensi dari kepemilikan hak miliki bangunan (Rumah/kontrakan/properti pribadi lainnya). Terlihat bahwa status kepemilikan hak milik bangunan nasabah tersebar merata dengan total 127 ribu masing-masing kategori (Yes atau No), atau dengan persentase hampir 50 % masing-masing kategori.

Bar chart ketiga berisikan informasi kepemilikan tanggungan oleh nasabah. Nilai pada chart ini juga tersebar merata dengan nilai kisaran 127 ribu masing-masing kategori. Dengan begitu kedua kategori memiliki persentase hampir 50 % yang menunjukkan data terdistribusi normal.

Bar keempat menunjukkan kepemilikan wakil oleh nasabah saat mengajukan pinjaman. Dapat dilihat bahwa sebagian nasabah memiliki wakil penandatangan dalam mengajukan pinjaman di Lawson Bank.

* **Nilai Rata-Rata Credit Score Tiap Tipe Employment dan Frekuensi Default**

A screenshot of a computer

Description automatically generated

*(Visualisasi Rata-Rata Credit Score dan Frekuensi Nasabah Default/Gagal Bayar)*

Pada visualisai berikutnya terdapat dua chart, yaitu bar chart rata-rata credit score tiap tipe pekerjaan dan pie chart status default (Gagal bayar) keseluruhan nasabah. Dapat dilihat bahwa credi score dari masing-masing tipe pekerjaan berada di kisaran angka yang mirip, yaitu di nilai rata-rata 574, yang mana menunjukkan angka yang cukup baik dari para nasabah dalam menjaga credit score sebagai kredibilitas mereka sebagai peminjam.

Pada pie chart dapat dilihat persentase nasabah yang berhasil bayar dan gagal bayar. Dimana angka 0 menunjukkan nasabah berhasil bayar dan menjaga kredibilitas pinjamannya, sedangkan angka 1 menunjukkan nasabah gagal bayar pinjamannya. Sebanyak 88.4 % nasabah berhasil bayar dan menjaga kredibilitasnya, sedangkan sebanyak 11.6 % nasabah mengalami gagal bayar dalam melakukan pinjaman di Lawson Bank. Persentase ini menunjukkan hasil yang cukup baik dan dapat menjadi catatan bagi Lawson Bank dalam meningkatkan pengawasan dan ketetatan dalam penyetujuan pengajuan pinjaman dana yang dilakukan oleh nasabah kedepannya.

1. **Question 2**

* Build and manage your **Analytics model** using SAS Software.
* The journal writing in section IV contains an explanation of Modelling your data
* **Create your Analytical Model** by design and concept of SAS VDMML with **Forest or SVM** algorithms **with comparison of other algorithms.**

**Analytics model menggunakan SAS Software**

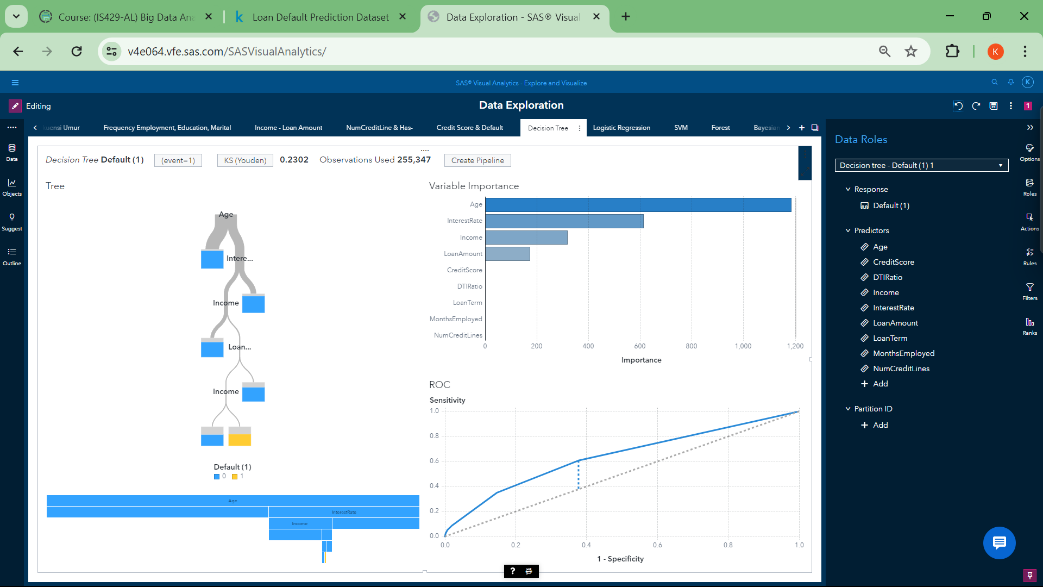
Pada studi kasus ini, dibuat model analitik untuk melihat analisis dan membuat machine learning pada data yang saya miliki. Tahap pertama adalah membuat model analitik yang menggunakan Decision Tree dan Logistic Regression. Sebelum masuk ke tahap tersebut, dilakukan penentuan variabel yang akan digunakan, yaitu “default” sebagai response dan seluruh variabel numerik continuous sebagai predictors.

A screenshot of a computer

Description automatically generated

*(Penentuan variabel Response dan Predictors)*

1. **Decision Tree**



*(Hasil Visualisasi Decision Tree)*

Pada visualisasi Decision Tree tersebut dapat dilihat bahwa akurasi algortima ini pada dataframe yang digunakan sebesar 23.02 % (Berdasarkan nilai KS Youden). Decision Tree ini digunakan dalam menganalisis prediksi kemungkinan default nasabah berdasarkan Predictors data numerik continuous.

Node pada Decision Tree tersebut berakar dari Age sebagai variabel dengan importance tertinggi, yang kemudian dilanjutkan melalui variabel terpenting hingga variabel dengan importance terendah, sampai pada hasil akhir yaitu penentuan apakah nasabah mengalami default atau tidak (1 = default, 0 = tidak default).

A screenshot of a computer screen

Description automatically generatedA screenshot of a computer screen

Description automatically generated

*(Informasi node pada node default & node not default)*

Gambar tersebut menunjukkan informasi pada node bagian akhir, yaitu node default. Informasi tersebut menunjukkan proses pemrosesan informasi dari variabel-variabel yang telah dijadikan predictors hingga data keluar sebagai output yang sudah dianalisis.

A graph with a line

Description automatically generated

*(Line Chart ROC)*

Hasil pada line chart ROC menunjukkan performa kualitas algoritma Decision Tree tersebut, Dimana garis ROC mengalami kenaikkan pada Tengah chart yang kemudian sempat menurun mendekati garis reference. Hal ini menunjukkan algortima tersebut menagalami kenaikkan kualitas di Tengah proses algoritma dan mengalami penurunan di akhir (Namun masih di atas garis reference menunjukkan kualitas masih aman).

1. **Logistic Regression**

A screenshot of a computer

Description automatically generated

*(Visualisasi Logistic Regression)*

Pada visualisasi Logistic Regression, dilakukan pemrosesan variabel numerik continuous dan classifiocation (kategorikal). Hal ini dilakukan untuk melihat keterkaitan antar variabel menjadi chart logistic regression. Dapat dilihat bahwa akurasi Logistic Regression sebesar 36.44 % (Berdasarkan nilai KS - Youden). Nilai akurasi tersebut lebih baik jika dibandingkan dengan akurasi pada Decision Tree sebelumnya. Hal ini menunjukkan kontribusi variabel kategorikal ikut mempengaruhi nilai akurasi menjadi lebih baik.

**Analytical Model menggunakan SAS VDMML**

Pada studi kasus kali ini, saya menggunakan 4 model analisis pada SAS VDMML, yaitu Forest yang dibandingkan dengan 3 model algoritma lainnya, yaitu Bayesian Network, Gradient Boosting, dan Neural Network.

Langkah pertama untuk melakukan analisis model, yaitu menentukan variabel predictors dan response. Response diisi dengan variabel kolom “default” yang menjadi label supervised learning sebagai hasil status gagal baya ratu tidaknya nasabah sebagai hasil prediksi. Variabel predictors berisikan seluruh variabel numerik continuous.

A screenshot of a computer

Description automatically generated

*(Penentuan variabel Response dan Predictors)*

1. **Forest**

A screenshot of a computer

Description automatically generated

*(Visualisasi Random Forest)*

Visualisasi Random Forest memiliki nilai akurasi sebesar 10.53 % (berdasarkan nilai KS - Youden)

* + **Variable Importance**

Grafik batang di bagian kiri atas menampilkan pentingnya variabel dalam model hutan acak. Variabel yang paling berpengaruh terhadap model ditampilkan di bagian atas.

Dalam visualisasi ini, variabel Age (Usia) memiliki pengaruh paling besar terhadap prediksi default pinjaman, diikuti oleh Income (Pendapatan), InterestRate (Suku Bunga), LoanAmount (Jumlah Pinjaman), dan seterusnya.

* + **Error Plot**

Grafik ini menunjukkan tingkat kesalahan klasifikasi (Misclassification Rate) model seiring dengan bertambahnya jumlah pohon dalam hutan acak.

Terdapat dua garis dalam grafik: garis biru menunjukkan tingkat kesalahan klasifikasi pada data pelatihan, sementara garis ungu menunjukkan tingkat kesalahan klasifikasi pada data out-of-bag (OOB).

Secara umum, grafik ini menunjukkan stabilitas tingkat kesalahan klasifikasi seiring bertambahnya jumlah pohon, dengan nilai kesalahan berkisar sekitar 0.115.

* + **Confusion Matrix**

Confusion matrix di bagian bawah menunjukkan performa model dalam bentuk jumlah prediksi benar dan salah untuk masing-masing kelas (default dan non-default).

Angka-angka dalam matriks menunjukkan bahwa model memprediksi dengan benar sebanyak 225,691 contoh sebagai non-default (0) dan 36 contoh sebagai default (1).

Kesalahan prediksi terjadi sebanyak 3 contoh yang seharusnya non-default diprediksi sebagai default, dan 29,617 contoh yang seharusnya default diprediksi sebagai non-default.

Dari visualisasi ini, kita dapat memahami beberapa hal penting tentang performa model Random Forest yang dibangun, yaitu:

* + **Pentingnya variabel**

Usia dan pendapatan adalah faktor utama dalam prediksi default.

* + **Stabilitas model**

Tingkat kesalahan klasifikasi relatif stabil seiring bertambahnya jumlah pohon, menunjukkan model hutan acak yang robust.

* + **Performa model**

Matriks kebingungan menunjukkan bahwa model memiliki performa yang cukup baik dalam mengklasifikasikan non-default, namun perlu perbaikan dalam mengklasifikasikan default, mengingat masih ada jumlah kesalahan yang cukup besar pada prediksi default.

1. **Bayesian Network**

A screenshot of a computer

Description automatically generated

*(Visualisasi Bayesian Network)*

Hasil visualisasi Bayesian Network memiliki nilai akurasi 35.23 % (Berdasarkan nilai KS - Youden).

* + **Network Diagram**

Diagram jaringan di kiri atas menggambarkan hubungan antara variabel prediktor dan variabel target (default). Variabel Default (1) berada di pusat dan dihubungkan dengan variabel prediktor lainnya.

Variabel-variabel yang berhubungan langsung dengan default antara lain: MonthsEmployed (Lama Bekerja), Income (Pendapatan), CreditScore (Skor Kredit), InterestRate (Suku Bunga), Age (Usia), dan DTIRatio (Rasio Debt-to-Income).

* + **Variables in Network**

Grafik batang di kanan atas menampilkan skor BIC (Bayesian Information Criterion) untuk setiap variabel dalam jaringan. Skor BIC digunakan untuk mengukur seberapa baik model fit dengan data; semakin kecil skor BIC, semakin baik fit model.

Variabel dengan skor BIC terkecil, seperti Age dan InterestRate, menunjukkan bahwa mereka memberikan kontribusi signifikan dalam model jaringan Bayesian.

* + **Model Selection**

Grafik ini menunjukkan tingkat kesalahan klasifikasi (Misclassification Rate) berdasarkan jumlah maksimum orang tua (Maximum Parents) yang dipilih untuk model.

Terdapat beberapa metode evaluasi seperti struktur, Markov blanket, dan parent-child. Misclassification Rate menunjukkan performa model pada berbagai konfigurasi. Pada grafik ini, kita melihat bahwa dengan jumlah maksimum orang tua yang sedikit (1 atau 2), model memberikan tingkat kesalahan klasifikasi yang lebih rendah dan stabil.

* + **Confusion Matrix**

Confusion matrix di bagian bawah menunjukkan performa model dalam bentuk jumlah prediksi benar dan salah untuk masing-masing kelas (default dan non-default).

Angka-angka dalam matriks menunjukkan bahwa model memprediksi dengan benar sebanyak 225,362 contoh sebagai non-default (0) dan 603 contoh sebagai default (1).

Kesalahan prediksi terjadi sebanyak 332 contoh yang seharusnya non-default diprediksi sebagai default, dan 29,050 contoh yang seharusnya default diprediksi sebagai non-default.

Dari visualisasi ini, kita dapat melihat beberapa hal penting mengenai performa model jaringan Bayesian yang dibangun:

* + **Struktur jaringan**

Jaringan menunjukkan hubungan langsung antara variabel-variabel prediktor dengan variabel target default.

* + **Pentingnya variabel**

Usia dan suku bunga adalah variabel yang sangat berpengaruh dalam model jaringan Bayesian, ditunjukkan dengan skor BIC yang lebih baik.

* + **Performa model**

Tingkat kesalahan klasifikasi stabil pada berbagai konfigurasi model, namun ada kesalahan prediksi yang signifikan pada kelas default, menunjukkan bahwa model mungkin perlu ditingkatkan untuk memprediksi default lebih akurat.

1. **Gradient Boosting**

A screenshot of a computer

Description automatically generated

*(Hasil visualisasi Gradient Boosting)*

Hasil visualisasi Gradient Boosting menunjukkan performanya dengan nilai akurasi sebesar 36.58 % (Berdasarkan nilai KS - Youden).

* + **Variable Importance**

Grafik batang di bagian kiri atas menampilkan pentingnya variabel dalam model Gradient Boosting. Variabel yang paling berpengaruh terhadap model ditampilkan di bagian atas.

Dalam visualisasi ini, variabel Age (Usia) memiliki pengaruh paling besar terhadap prediksi default pinjaman, diikuti oleh Income (Pendapatan), LoanAmount (Jumlah Pinjaman), InterestRate (Suku Bunga), dan seterusnya.

* + **Iteration Plot**

Grafik ini menunjukkan tingkat kesalahan klasifikasi (Misclassification Rate) model seiring dengan bertambahnya jumlah pohon dalam proses boosting.

Garis pada grafik menunjukkan bahwa tingkat kesalahan klasifikasi terus menurun seiring bertambahnya jumlah pohon, mencapai tingkat yang stabil di sekitar 0.115 setelah sekitar 40 pohon. Ini menunjukkan bahwa model terus belajar dan meningkatkan akurasinya dengan setiap iterasi.

* + **Confusion Matrix**

Confusion matrix di bagian bawah menunjukkan performa model dalam bentuk jumlah prediksi benar dan salah untuk masing-masing kelas (default dan non-default).

Angka-angka dalam matriks menunjukkan bahwa model memprediksi dengan benar sebanyak 224,745 contoh sebagai non-default (0) dan 1,231 contoh sebagai default (1).

Kesalahan prediksi terjadi sebanyak 718 contoh yang seharusnya non-default diprediksi sebagai default, dan 28,653 contoh yang seharusnya default diprediksi sebagai non-default.

Dari visualisasi ini, kita dapat melihat beberapa hal penting mengenai performa model Gradient Boosting yang dibangun:

* + **Pentingnya variabel**

Usia dan pendapatan adalah faktor utama dalam prediksi default, diikuti oleh jumlah pinjaman dan suku bunga.

* + **Stabilitas model**

Tingkat kesalahan klasifikasi menurun dan menjadi stabil seiring bertambahnya jumlah pohon, menunjukkan bahwa model Gradient Boosting terus meningkatkan akurasinya.

* + **Performa model**

Matriks kebingungan menunjukkan bahwa model memiliki performa yang cukup baik dalam mengklasifikasikan non-default, namun masih terdapat kesalahan yang cukup besar pada prediksi default, menunjukkan adanya ruang untuk peningkatan lebih lanjut.

1. **Neural Network**

A screenshot of a computer

Description automatically generated

*(Hasil Visualisasi Neural Network)*

Hasil visualisasi menunjukkan performa dengan nilai akurasi sebesar 34.78 % (Berdasarkan nilai KS - Youden).

* + - **Network Diagram (Diagram Jaringan)**

Diagram jaringan neural ini menunjukkan struktur dari model neural network yang telah dibangun. Node pada diagram merepresentasikan neuron di setiap layer, sedangkan garis-garis (link) yang menghubungkan node-node tersebut merepresentasikan koneksi antara neuron di berbagai layer.

Ukuran dan warna dari node dan link merefleksikan kekuatan dan signifikansi dari neuron dan koneksi tersebut. Misalnya, node dengan warna lebih cerah (kuning) dan ukuran lebih besar menunjukkan neuron dengan bobot yang lebih signifikan dalam jaringan.

* + **Neuron Absolute Average**

Diagram ini memberikan informasi tentang rata-rata absolut bobot neuron pada setiap layer. Warna dan ukuran dari node menunjukkan bobot rata-rata absolut tersebut, di mana warna yang lebih terang dan ukuran yang lebih besar menunjukkan bobot yang lebih tinggi.

* + **Iteration Plot (Plot Iterasi)**

Plot ini menunjukkan grafik loss function dan objective function selama proses training. Pada sumbu x ditampilkan jumlah iterasi, sedangkan sumbu y menunjukkan nilai loss dan objective function.

Dari grafik terlihat bahwa nilai loss dan objective function menurun seiring bertambahnya iterasi hingga mencapai stabilitas. Ini menunjukkan bahwa model mengalami proses konvergensi yang baik selama pelatihan.

* + **Confusion Matrix**

Confusion Matrix pada hasil visualisasi ini menunjukkan hasil prediksi model neural network terhadap dataset yang digunakan.

Dalam kasus ini, matriks menunjukkan dua kelas, dengan nilai di dalam matriks mengindikasikan jumlah observasi yang diklasifikasikan dengan benar (diagonal utama) dan yang salah klasifikasi (off-diagonal).

Dari matriks, terlihat bahwa kelas 0 memiliki 225,694 observasi yang diklasifikasikan dengan benar, sedangkan kelas 1 memiliki 29,653 observasi yang diklasifikasikan dengan benar.

Dengan visualisasi Neural Network tersebut, dapat disimpulkan bahwa model neural network yang dibangun menggunakan berbagai variabel prediktor untuk memprediksi kemungkinan kegagalan bayar pinjaman. Proses pelatihan model terlihat berhasil karena nilai loss dan objective function menurun dan stabil. Matriks kebingungan menunjukkan bahwa model ini memiliki performa yang baik dalam mengklasifikasikan data pada kelas yang tepat.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nilai Pembanding** | **Forest** | **Bayesian Network** | **Gradient Boosting** | **Neural Network** |
| Nilai Akurasi:  KS (Youden) | 10.53% | 35.23% | 36.58% | 34.78% |

Kesimpulan Hasil Visualisasi Metode Algoritma:

Model algoritma terbaik adalah Gradient Boosting. Hal ini dapat dinilai dari hasil nilai akurasi (KS - Youden) yang dihasilkan dari visualisasi Gradient Boosting sebesar 35.58% yang paling tinggi dibanding algoritma lainnya. Sedangkan untuk algoritma dengan nilai akurasi terndah adalah algoritma Forest. Berikut adalah penjelasan mengenai alasan perbedaan nilai akurasi antara masing-masing algoritma:

1. **Gradient Boosting (36.58% - Akurasi Tertinggi)**

Gradient Boosting adalah metode ensemble yang menggabungkan prediksi dari beberapa model dasar (biasanya decision trees) untuk meningkatkan akurasi. Alasan Gradient Boosting bisa memiliki nilai akurasi tertinggi adalah:

1. **Sequential Learning**

Gradient Boosting membangun model secara berurutan, dengan setiap model baru mencoba untuk mengoreksi kesalahan dari model sebelumnya. Proses ini memungkinkan model tersebut untuk terus meningkatkan prediksi pada data yang sulit.

1. **Bias-Variance Tradeoff**

Dengan menggabungkan banyak pohon keputusan yang sederhana, Gradient Boosting mengurangi bias dan varians secara bersamaan. Sehingga algoritma ini dapat menghasilkan model yang lebih akurat.

1. **Flexibilitas**

Gradient Boosting dapat menangani berbagai jenis data dan masalah, termasuk yang memiliki hubungan kompleks antara fitur dan target.

1. **Neural Network (34.78%)**

Neural Network adalah model yang sangat fleksibel dan mampu menangkap hubungan non-linear yang kompleks dalam data. Namun, alasan mengapa performanya sedikit di bawah Gradient Boosting bisa termasuk:

1. **Overfitting**

Neural Network, terutama jika terlalu kompleks atau tidak diatur dengan baik (misalnya, tidak menggunakan regularisasi yang cukup), bisa mudah overfit pada data pelatihan, mengurangi performa pada data tes.

1. **Optimisasi**

Proses pelatihan Neural Network memerlukan optimisasi yang baik. Jika hyperparameter (seperti learning rate, jumlah layer, dan neuron) tidak diatur dengan optimal, performa bisa terpengaruh.

1. **Data Requirements**

Neural Network sering memerlukan banyak data untuk dilatih dengan baik. Jika dataset tidak cukup besar, model bisa kesulitan mencapai performa maksimal.

1. **Bayesian Network (35.23%)**

Bayesian Network menggunakan teori probabilitas untuk memodelkan hubungan antara variabel. Kelebihan dan kekurangannya dari Bayesian Network, yaitu:

1. **Probabilistic Modeling**

Bayesian Network efektif dalam menangani ketidakpastian dan dapat memberikan informasi tentang probabilitas kondisi yang terjadi. Ini bisa memberikan hasil yang baik jika relasi probabilistik antara variabel dapat ditangkap dengan baik.

1. **Complexity**

Meskipun fleksibel, Bayesian Network bisa rumit dan membutuhkan asumsi yang kuat mengenai independensi antara variabel. Jika asumsi ini tidak terpenuhi, akurasi bisa terpengaruh.

1. **Forest (10.53% - Akurasi Terendah)**

Forest, atau Random Forest, biasanya merupakan algoritma yang kuat dan akurat. Namun, beberapa faktor yang mungkin menyebabkan akurasi terendah dalam kasus ini adalah:

1. **Feature Importance**

Jika fitur yang digunakan tidak memiliki informasi yang cukup atau tidak relevan, Random Forest bisa memberikan performa yang rendah.

1. **Hyperparameter Tuning**

Performanya sangat bergantung pada parameter seperti jumlah pohon, kedalaman pohon, dan sampel fitur yang digunakan. Jika parameter ini tidak dioptimalkan, hasilnya bisa mengecewakan.

1. **Data Characteristics**

Random Forest bekerja baik pada data yang seimbang dan dengan fitur yang memiliki hubungan linier atau sedikit non-linear. Jika data sangat kompleks atau memiliki banyak noise, performanya bisa menurun.

**Kesimpulan**

Perbandingan ini menunjukkan bahwa **Gradient Boosting** berhasil mencapai **performa tertinggi** karena kemampuannya dalam menangani kesalahan secara iteratif dan menangkap hubungan kompleks dalam data. Neural Network dan Bayesian Network juga performa yang baik, namun sedikit di bawah Gradient Boosting. **Random Forest** yang pada biasanya kuat dalam melakukan prediksi, memiliki performa terendah yang dikarenakan kemungkinan akibat tuning parameter yang tidak optimal atau karakteristik data yang tidak sesuai dengan keunggulannya.

Link Report SAS Explore and Visualize:

<https://v4e064.vfe.sas.com/links/resources/report?uri=%2Freports%2Freports%2F424bbedd-aabe-4cb8-a73b-29351e74ffd4&page=vi14763>