Perbandingan Model Prediktif *Customer Churn* dengan Algoritma *Decision Tree*, *Logistic Regression*, dan *Forest* pada Bisnis H&M

Faustine Ilone Hadinata
Sistem Informasi
Universitas Multimedia Nusantara
Banten, Indonesia
faustine.ilone@student.umn.ac.id

Abstract*—Penelitian ini bertujuan mencari algoritma dengan hasil paling optimal dalam melakukan prediksi terhadap variabel *customer churn* guna mengatasi ancaman pelanggan meninggalkan perusahaan retail bernama H&M dalam digitalisasi yang berlangsung. *Dataset* yang digunakan pada penelitian ini bersumber dari Kaggle dan berisi berbagai data mengenai perilaku belanja *online* pelanggan. Metode penelitian yang dipilih adalah *framework* DCOVA & I. Peneliti juga melakukan eksplorasi terhadap atribut mana yang memiliki pengaruh besar terhadap variabel *churn* yang juga berperan sebagai *label* atau target dari model klasifikasi. *Exploratory Data Analysis* dilakukan sebelum pemodelan melalui visualisasi dengan *bar plot. Histogram, pie chart,* dan *box plot,* statistika deskriptif, serta analisis korelasi. Tiga algoritma yang digunakan adalah *Decision Tree, Logistic Regression,* serta *Forest* dengan *misclassification rate* dan *confusion matrix* sebagai evaluasi. *Tools* yang digunakan dalam analisis ini adalah SAS Viya dan SAS *Enterprise* Guide. Hasil dari penelitian ini adalah *Logistic Regression* sebagai algoritma paling optimal dengan *misclassification rate* paling rendah bernilai 0.0858 dan jumlah prediksi benar sebesar 3,047.

Keywords—Churn, DCOVA & I, Classification, Decision Tree, Logistic Regression, Forest, SAS Viya, SAS Enterprise Guide, H&M

I. LATAR BELAKANG & PEMAHAMAN BISNIS

Industri bisnis retail telah mengalami pergeseran besar-besaran seiring berjalan waktu. Kehadiran internet dan perkembangan teknologi mendorong bisnis untuk beradaptasi dan melakukan digitalisasi. Transformasi ini penting dilakukan guna menunjang proses bisnis dan meningkatkan *user satisfaction* pelanggan saat melakukan transaksi. Perubahan dari operasi bisnis tradisional dapat membantu bisnis bersaing dan bertahan dalam lingkungan kompetitif. Digitalisasi di sini juga memicu perubahan pada perilaku belanja pelanggan, dimana berbelanja secara *online* melalui *website*, aplikasi, atau *e-commerce* dinilai lebih mudah dan efisien. Hal ini mendorong bisnis retail untuk mengupayakan pengalaman belanja *online* yang memuaskan pada pelanggan.

Namun, digitalisasi ini juga disertai dengan probabilitas bisnis retail ditinggalkan oleh pelanggan saat gagal beradaptasi dengan perkembangan digital. Hal ini dapat dicegah oleh bisnis dengan cara melakukan analisis mendalam mengenai perilaku *online* pelanggan. Dengan memahami hal tersebut, bisnis dapat mengembangkan strategi pemasaran dan membangun hubungan jangka panjang dengan pelanggan [1]. Pelanggan setia adalah salah satu subjek penting dalam manajemen bisnis yang perlu dipertahankan [2]. Bisnis harus senantiasa berusaha memenuhi kebutuhan pelanggan dengan menawarkan ekslusivitas dalam pengalaman berbelanja sehingga dapat menumbuhkan loyalitas.

H&M adalah salah satu perusahaan retail multinasional yang pertama didirikan pada tahun 1947. *Brand* ini adalah salah satu perusahaan *fashion* terkemuka di dunia. H&M juga telah memperkenalkan *online shopping* sebagai salah satu strategi adaptasi. Pelanggan H&M sendiri tersebar di berbagai penjuru dunia sehingga membutuhkan perhatian lebih dalam pemeliharaan pelanggan. Hal ini dapat dilakukan melalui strategi personalisasi berbasis data guna meningkatkan retensi pelanggan. Sehubungan dengan permasalahan tersebut, tujuan dari penelitian ini adalah membantu memahami atribut atau variabel yang memiliki relevansi atau pengaruh besar pada retensi pelanggan. *Dataset* yang digunakan bersumber dari Kaggle dengan berbagai atribut mulai dari *Account Length, Location Code, Push Status, App Sessions,* dan lain-lain. Dengan demikian, perusahaan ritel atau H&M dapat memaksimalkan strategi penjualan dan hubungan dengan pelanggan sehingga dapat mengurangi *churn rate*.

II. TINJAUAN TEORITIS

Tinjauan teoritis berisi hasil dari berbagai studi pustaka yang dilakukan berhubungan dengan penelitian ini seperti tiga algoritma yang digunakan, antara lain *Decision Tree*, *Logistic Regression*, dan *Forest*, serta *misclassification rate* dan *confusion matrix* sebagai evaluasi. Masing-masing penjelasan dapat dilihat di bawah ini.

A. *Decision Tree*

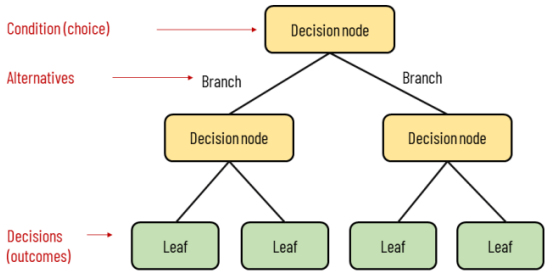

Fig. 1. Ilustrasi *Decision Tree*

Decision Tree adalah salah satu algoritma klasifikasi yang umum digunakan dalam pembelajaran mesin. Metode ini menggunakan struktur pohon yang terdiri atas *node* dan *branch* untuk menggambarkan keputusan atau aturan dalam melakukan prediksi variabel target berdasarkan *input* yang diperoleh [4]. Masing-masing *node* mewakili suatu fitur pada *dataset*. Pembagian data pada setiap *node* dilakukan

berdasarkan perhitungan *entropy* dan *gain*. *Entropy* berfungsi mengukur ketidakpastian atau keacakan data pada suatu *node* melalui rumus di bawah ini [5].

$$Entropy(S) = -\sum P_1 log_2 P_1 \quad (1)$$

Gain berfungsi melakukan pengukuran atas jumlah informasi baru yang diperoleh saat data dibagi berdasarkan atribut tertentu melalui rumus di bawah ini [4].

$$Gain(S, A) = \sum_{V \epsilon V(A)} \frac{|S_V|}{|S|} Entropy(S_V) \quad (2)$$

B. Logistic Regression

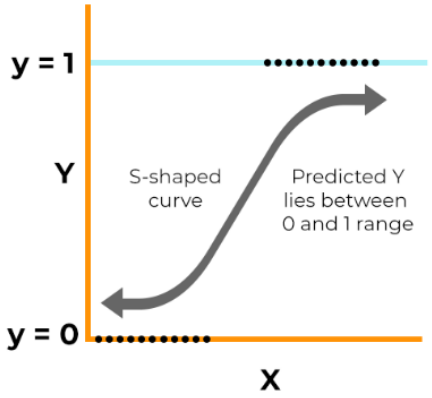

Fig. 2. Ilustrasi *Logistic Regression*

Meskipun mengandung kata 'regresi', *Logistic Regresion* di sini adalah metode klasifikasi. Algortima ini menggunakan fungsi logistik dalam pemodelan probabilitas variabel target dan mengubah nilai kontinu menjadi 0 atau 1 pada klasifikasi biner [7]. Nilai probabilitas tersebut lalu digunakan untuk melakukan klasifikasi data baru pada *label* yang tepat. Terdapat pula rumus dari *Logistic Regression* seperti di bawah ini.

$$Ln\left(\frac{P}{Q}\right) = Ln\left(e^{\beta 0 + \beta 1 X}\right) = \beta_0 \beta_1 X \quad (3)$$

C. Forest

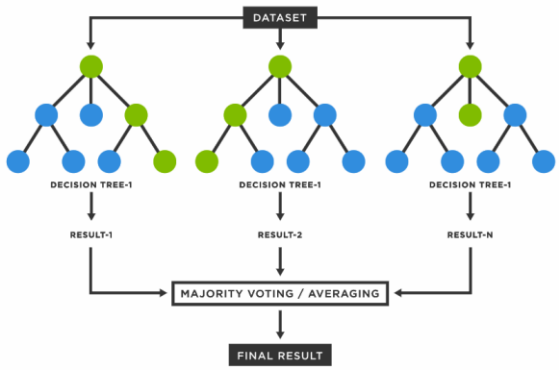

Fig. 3. Ilustrasi *Forest*

Algoritma *Random Forest* adalah penggabungan dari metode *Bagging* dan *Random Sub Spaces* dimana prediksi dari sejumlah besar pohon keputusan disatukan [8]. Keunggulan dari *Random Forest* adalah mampu mengatasi *overfitting* sebab variasi diperoleh melalui *subset* acak fitur dan sampel ulang serta menangani permasalahan ketidakseimbangan data.

D. Confusion Matrix

Confusion Matrix adalah sebuah tabel evaluasi yang berisi perbandingan antara hasil prediksi dengan *label* data asli [6]. *Confusion Matrix* terdiri dari empat bagian utama, antara lain:

1) *True Positive (TP)*
   TP berisi jumlah data yang berhasil diprediksi dengan benar oleh model dimana data bernilai positif.
2) *True Negative (TN)*
   TN berisi jumlah data yang berhasil diprediksi dengan benar oleh model dimana data bernilai negatif.
3) *False Positive (FP)*
   FP berisi jumlah data yang gagal diprediksi dengan benar oleh model dimana data bernilai positif.
4) *False Negative (FN).*
   FN berisi jumlah data yang gagal diprediksi dengan benar oleh model dimana data bernilai negatif.

E. Misclassification Rate

Misclassification Rate adalah salah satu metriks evaluasi yang berfungsi menggambarkan tingkat kesalahan suatu model klasifikasi dalam memprediksi *label* data. Perhitungan tersebut dapat dilakukan melalui rumus berikut, dimana semakin rendah nilai yang diperoleh menunjukkan performa model yang semakin bagus.

$$\frac{FP+FN}{Total\ Data} \quad (4)$$

III. METODOLOGI PENELITIAN

Metode penelitian yang digunakan oleh peneliti adalah DCOVA & I. *Framework* ini terdiri dari proses *Define, Collect, Organize, Visualize, Analyze,* dan *Insights* yang dapat menunjang pengembangan proses bisnis [3].

A. Define

Tahap ini berisi pendefinisian permasalahan dan tujuan penelitian. Permasalahan yang dihadapi perusahaan retail H&M adalah ancaman dari digitalisasi dimana pelanggan dapat dengan mudah meninggalkan suatu bisnis apabila bisnis tersebut gagal beradaptasi dan memenuhi kebutuhan pengguna. Tujuan dari penelitian ini adalah memahami perilaku belanja *online* pelanggan dan mengetahui atribut mana yang memiliki pengaruh besar dalam retensi, sehingga H&M dapat merancang strategi bisnis yang sesuai dengan lapangan yang ada. Analisis mendalam ini juga dapat membantu H&M dalam menjaga dan memelihara pelanggan setia. Tahap *Define* juga meliputi penentuan data yang relevan dan sesuai dengan penelitan, dimana pada penelitian ini, dipilih suatu *dataset* mengenai data *online shopping* pelanggan.

B. Collect

Pengumpulan data dilakukan pada tahap *Collect*, dimana *dataset* yang digunakan berisi atribut seperti tabel di bawah ini. *Dataset* juga diperoleh dari *platform* Kaggle serta terdiri dari 20 *columns* dan 3,333 *rows* atau *record*.

No	Atribut	Penjelasan
1	Account Length	Berapa lama akun pengguna telah terdaftar dalam satuan hari
2	Location Code	Kode lokasi tempat pengguna berada
3	User ID	ID unik yang berfungsi mengidentifikasi pengguna

4	Credit Card Info Save	*Boolean* mengenai apakah pengguna menyimpan informasi kartu kredit pada sistem
5	Push Status	*Boolean* mengenai apakah pengguna mengaktifkan notifikasi
6	Add to Wishlist	Frekuensi pengguna menambahkan produk pada *wishlist*
7	Desktop Sessions	Jumlah sesi *desktop* pengguna
8	App Sessions	Jumlah sesi aplikasi pengguna
9	Desktop Transactions	Jumlah transaksi yang dilakukan oleh pengguna melalui *desktop*
10	Total Product Detail Reviews	Jumlah ulasan dalam rincian produk
11	Session Duration	Rata-rata waktu yang dihabiskan oleh pengguna dalam setiap sesi (satuan menit)
12	Promotion Clicks	Frekuensi pengguna melakukan klik pada penawaran diskon
13	Avg Order Value	Rata-rata jumlah uang yang dikeluarkan oleh pelanggan dalam pembelian
14	Sale Product Views	Jumlah produk diskon yang dilihat oleh pengguna
15	Discount Rate per Visited Products	Rata-rata *discount rate* dari seluruh produk yang dilihat oleh pengguna
16	Product Detail View per App Session	Jumlah rincian produk yang dilihat oleh pengguna dalam setiap sesi aplikasi
17	App Transactions	Jumlah transaksi yang dilakukan oleh pengguna di aplikasi
18	Add to Cart per Session	Frekuensi pengguna menambahkan produk pada keranjang dalam setiap sesi
19	Customer Service Calls	Frekuensi pengguna menghubungi *customer service*
20	Churn	Varibel target dimana 0 berarti pelanggan masih melakukan transaksi pada perusahaan dan 1 berarti pelanggan telah meninggalkan bisnis tersebut

C. *Organize*

Tranformasi pada *dataset* dilakukan pada tahap ini, seperti pengubahan tipe data dari numerik menjadi kategorikal. Proses pemastian ini penting agar eksplorasi dan pemodelan data menjadi lebih mudah dan efisien. Pemeriksaan terhadap nilai *null* juga dilakukan pada tahap *Organize.*

D. *Visualize*

Visualisasi bertujuan memberi gambaran mengenai ditribusi data yang dapat mempermudah pembaca dalam memahami data. Visualisasi pada penelitian ini dilakukan melalui penggunaan *bar plot, pie chart, boxplot,* dan *histogram* terhadap variabel numerik dan kategorikal.

E. *Analyze*

Proses analisa pada penelitian ini dilakukan menggunakan tiga algoritma berbeda sebagai perbandingan, antara lain *Decision Tree, Logistic Regression,* dan *Forest*. Algoritma klasifikasi dipilih sebab *dataset* memiliki *label* atau variabel target (*churn*) sebagai hasil prediksi. Variabel *churn* di sini dapat bernilai 0 atau 1. Terdapat pula beberapa *predictors* dari *response churn*, antara lain *avg order value, location code, push status, account length, session duration, total product detail views,* dan *app sessions*. Evaluasi dilakukan berdasarkan *misclassification rate* dan *confusion matrix*.

F. *Insights*

Visualisasi dan analisis yang dilakukan lalu menghasilkan suatu *insight* atau wawasan yang berguna bagi perusahaan H&M. Wawasan yang diperoleh dapat berupa informasi mengenai perilaku atau kecenderungan pelanggan dalam melakukan pembelian secara *online* ataupun model prediktif mana yang memiliki performa paling bagus.

IV. HASIL DAN PEMBAHASAN

Bagian ini berisi hasil yang diperoleh dari analisis menggunakan SAS Viya dan SAS *Enterprise Guide* dengan dua pembahasan utama, antara lain *Exploratory Data Analysis* dan *Modelling*.

A. *Exploratory Data Analysis* (EDA)

Exloratory Data Analysis atau EDA adalah proses dalam analisis data yang dilakukan sebelum pemodelan. Tujuan EDA adalah meningkatkan pemahaman dan wawasan mengenai data yang digunakan dalam penelitian dengan cara melakukan eksplorasi pada data melalui visualisasi, statisika deskriptif, analisis korelasi, identifikasi *outliers*, dan lain-lain.

1) *Visualisasi*

Visualisasi pertama adalah *bar plot* dengan atribut *location code* dan dipisahkan berdasarkan variabel *churn* dengan nilai 0 atau 1. Dari visualisasi di bawah ini, diperoleh bahwa 415 adalah *location code* dengan jumlah pelanggan paling tinggi, disusul oleh 510, dan jumlah pelanggan paling sedikit ada pada *location code* 408. Wawasan lain adalah jumlah pelanggan setia dengan 0 pada nilai *churn* lebih besar daripada jumlah pelanggan hilang dengan 1 sebagai nilai *churn* pada setiap *location code*. Pelanggan setia dilambangkan oleh *bar plot* berwarna ungu dan pelanggan hilang diwakili oleh *bar plot* berwarna kuning.

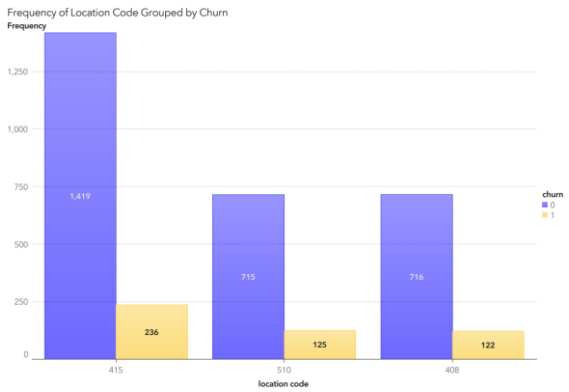

Fig. 4. *Bar Plot* Frekuensi *Location Code* Berdasarkan *Churn*

Histogram di bawah ini menunjukkan distribusi atribut *session duration* dengan 200 sebagai *highest point*. Hal ini berarti pelanggan cenderung menghabiskan waktu selama 200 menit dalam setiap sesi.

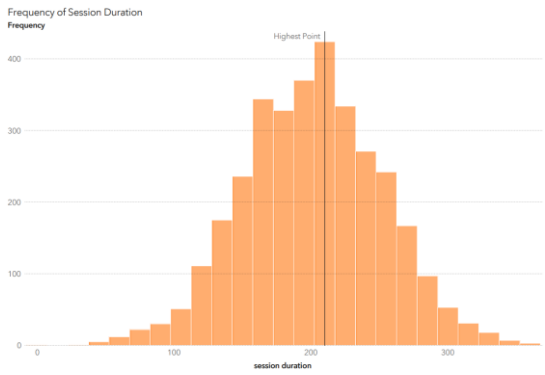

Fig. 5. *Histogram* Distribusi *Session Duration*

Pie chart di bawah ini dibuat dari atribut *push status* dan *churn*. Wawasan yang diperoleh adalah pelanggan cenderung menonaktifkan notifikasi, namun persentase pelanggan setia yang mengaktifkan notifikasi (29.5%) masih lebih besar dibandingkan dengan persentase pelanggan hilang dengan notifikasi aktif (16.6%). Hal ini memicu asumsi bahwa aktivasi notifikasi dapat berpengaruh terhadap retensi pelanggan.

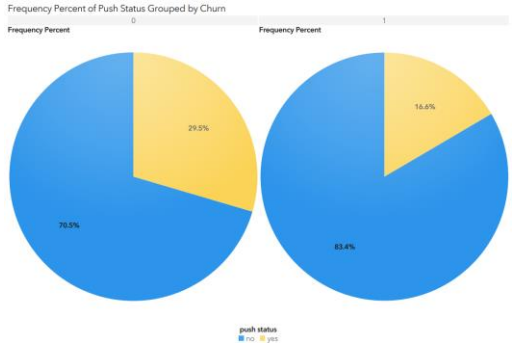

Fig. 6. *Pie Chart* Persentase Frekuensi *Push Status* Berdasarkan *Churn*

Box plot adalah visualisasi yang bertujuan memberi informasi mengenai distribusi, *range*, dan pusat data. *Box plot* di bawah ini dibuat dari atribut *sale product views* berdasarkan *push status*. Wawasan yang diperoleh adalah *range* dan median dari pelanggan dengan aktivasi notifikasi sedikit lebih tinggi daripada pelanggan yang menonaktifkan notifikasi. *Range* di sini dapat dilihat dari *whiskers* dan median dilihat dari garis di tengah *box*. *Range* dari *sale product views* berkisar antara 80-200 pada *push status 'yes'* dan *'no'*.

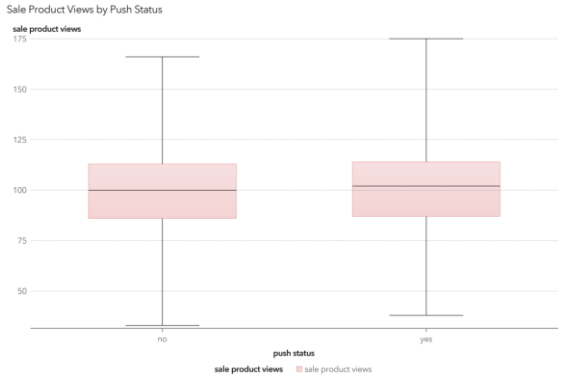

Fig. 7. *Box Plot Sale Product Views* Berdasarkan *Push Status*

2) Statistika Deskriptif dan Analisis Korelasi

Tabel *Simple Statistics* berisi statistika deskriptif dari beberapa variabel numerik seperti *add to wishlist*, *account length*, *app sessions*, *session duration*, *app transactions*, *location code*, dan *churn*. Pada tabel ini, dapat dilihat jumlah data (*N*), rata-rata (*Mean*), standar deviasi (*Std Dev*), penjumlahan (*Sum*), nilai minimum, dan nilai maksimum dari masing-masing atribut.

Tabel paling bawah adalah tabel berisi hasil korelasi *Pearson* dimana baris pertama dalam *churn* adalah nilai korelasi dan baris dua berarti *p-value*. Nilai *p-value* yang < 0.05 menunjukkan hasil yang signifikan. Korelasi tertinggi yang diperoleh ada pada hubungan *app sessions* dengan variabel *churn* sebesar 0.20468 dengan *p-value* < 0.0001. Nilai ini mengindikasikan bahwa terdapat *low positive correlation* antara dua variabel tersebut.

Correlation Analysis

The CORR Procedure

6 With Variables:	add to wishlist account length app sessions session duration app transactions location code
1 Variables:	churn

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
add to wishlist	3333	8.09901	13.68837	26994	0	51.00000
account length	3333	101.06481	39.82211	336849	1.00000	243.00000
app sessions	3333	30.56796	9.26938	101883	0	60.00000
session duration	3333	201.03960	50.71436	670065	0	364.00000
app transactions	3333	4.47945	2.46121	14930	0	20.00000
location code	3333	437.18242	42.37129	1457129	408.00000	510.00000
churn	3333	0.14491	0.35207	483.00000	0	1.00000

Pearson Correlation Coefficients, N = 3333
Prob > |r| under H0: Rho=0

	churn
add to wishlist	-0.08973 <.0001
account length	0.01654 0.3398
app sessions	0.20468 <.0001
session duration	0.09278 <.0001
app transactions	-0.05284 0.0023
location code	0.00617 0.7216

Fig. 8. Tabel Analisis Korelasi

B. *Modelling*

Decision Tree, *Logistic Regression*, dan *Forest* adalah tiga algoritma klasifikasi yang diimplementasikan pada model penelitian ini. Hasil temuan dari masing-masing algoritma dapat dilihat pada penjelasan di bawah ini.

1) Decision Tree

Model ini menghasilkan tiga visualisasi, antara lain *tree*, *variable importance*, dan *confusion matrix*. *Root node* pada *tree* adalah *app sessions* yang juga berperan sebagai variabel dengan *importance* tertinggi. *Session duration* dan *push status* adalah dua variabel penting lain dalam model ini. *Importance* di sini mengacu pada relevansi dari variabel tertentu terhadap hasil *churn* dimana variabel dengan relevansi tinggi berarti memiliki pengaruh yang besar. Terdapat pula nilai *misclassification rate* sebesar 0.1047.

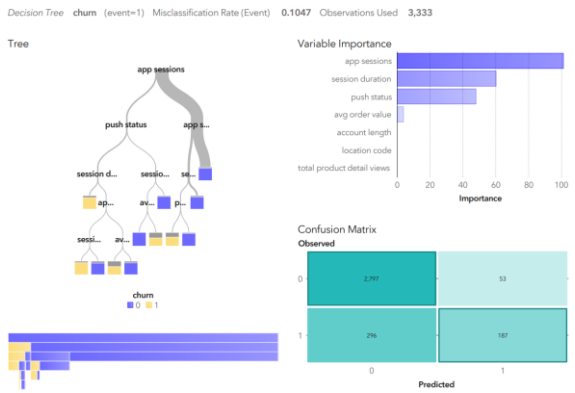

Fig. 9. Model Prediktif *Decision Tree*

Tabel di bawah ini adalah rincian hasil yang diperoleh dari *confusion matrix* model *Decision Tree*. *True Negative* dan *True Positive* menjadi fokus utama yang menunjukkan jumlah prediksi benar.

TABLE I. *CONFUSION MATRIX* MODEL *DECISION TREE*

Observasi	Jumlah Data
True Negative (TN)	2,797
False Negative (FN)	296
True Positive (TP)	187
False Positive (FP)	53

2) Logistic Regression

Model ini menghasilkan tiga visualisasi, antara lain *fit summary*, *residual plot*, dan *confusion matrix*. Pada *fit summary*, diperoleh hasil yang sedikit berbeda dengan model *Decision Tree*, dimana *push status* di sini dinilai lebih penting dibandingkan dengan *session duration*. Namun, *app sessions* tetap memiliki *importance* paling tinggi. Terdapat pula nilai *misclassification rate* sebesar 0.0858.

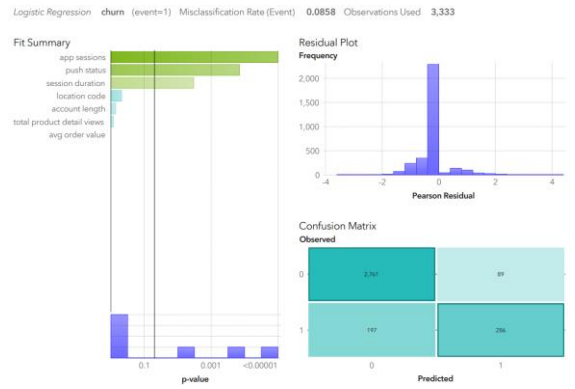

Fig. 10. Model Prediktif *Logistic Regression*

Tabel di bawah ini adalah rincian hasil yang diperoleh dari *confusion matrix* model *Logistic Regression*. *True Negative* dan *True Positive* menjadi fokus utama yang menunjukkan jumlah prediksi benar.

TABLE II. *CONFUSION MATRIX* MODEL *LOGISTIC REGRESSION*

Observasi	Jumlah Data
True Negative (TN)	2,761
False Negative (FN)	197
True Positive (TP)	286
False Positive (FP)	89

3) Forest

Model ini menghasilkan tiga visualisasi, antara lain *variable importance*, *partial dependence*, dan *confusion matrix*. Hasil *variable importance* model ini serupa dengan model *Logistic Regression* dengan urutan *importance* dimulai dari *app sessions*, *push status*, hingga *session duration*. Terdapat pula nilai *misclassification rate* sebesar 0.1146.

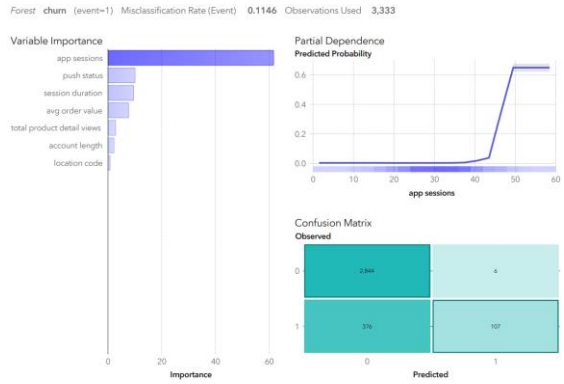

Fig. 11. Model Prediktif *Forest*

Tabel di bawah ini adalah rincian hasil yang diperoleh dari *confusion matrix* model *Forest*. *True Negative* dan *True Positive* menjadi fokus utama yang menunjukkan jumlah prediksi benar.

TABLE III. *CONFUSION MATRIX* MODEL *FOREST*

Observasi	Jumlah Data
True Negative (TN)	2,844
False Negative (FN)	376
True Positive (TP)	107
False Positive (FP)	6

Perbandingan dari tiga model di atas dilakukan guna memperoleh model dengan performa tertinggi. Metriks pembanding yang digunakan adalah nilai *misclassification rate*, *confusion matrix*, jumlah prediksi benar, dan jumlah prediksi salah. Jumlah prediksi benar dihitung dari penjumlahan antara TN dan TP, sementara jumlah prediksi salah adalah hasil penambahan dari FN dan FP. Dari tabel perbandingan di bawah ini, diperoleh hasil bahwa *Logistic Regression* adalah algoritma paling bagus dalam melakukan prediksi *customer churn*. Hal ini didasari oleh nilai *misclassification rate* paling rendah dan jumlah prediksi benar paling tinggi di antara dua model lain.

TABLE IV. *CONFUSION MATRIX* MODEL *FOREST*

Metriks Pembanding		***Decision Tree***	***Logistic Regression***	***Forest***
Misclassification Rate		0.1047	0.0858	0.1146
Confusion Matrix	TN	2,797	2,761	2,844
	FN	296	197	376
	TP	187	286	107
	FP	53	89	6
Jumlah Prediksi Benar		2,984	3,047	2,951
Jumlah Prediksi Salah		349	286	382

V. KESIMPULAN

Kesimpulan yang diperoleh dari penelitian ini adalah *Logistic Regression* adalah algoritma yang mampu menghasilkan performa model paling bagus dalam

memprediksi *customer churn*. Nilai *misclassification rate* model ini adalah 0.0858 dengan jumlah prediksi benar sebesar 3,047. Terdapat pula informasi mengenai atribut dengan signifikansi atau relevansi tertinggi terhadap variabel *churn* dilihat dari visualisasi *variable importance* dan analisis korelasi. Variabel tersebut adalah *app sessions* dengan nilai korelasi sebesar 0.20468 yang menunjukkan hubungan *low positive correlation*.

ACKNOWLEDGEMENTS

Peneliti ingin mengucapkan terima kasih pada Bapak Iwan Prasetiawan selaku pengampu mata kuliah *Big Data Analytics* yang telah memberikan wawasan dan senantiasa membimbing selama proses pembelajaran. Peneliti sangat menghargai segala saran dan ilmu yang telah diberikan oleh Bapak Iwan Prasetiawan.

DAFTAR PUSTAKA

[1] Smith, J. (2022). The Digital Transformation of Retail: Opportunities and Challenges. *Journal of Retailing, 42*(3), 189-205.

[2] Ruhulessin, M. F. (2023). Pentingnya Meraih Loyalitas Pelanggan dalam Kesuksesan Ritel Mewah. *Kompas.com*.

[3] Nga, N. T. Q. & Nhu, G. B. Q. (2020). An empirical study on factors influencing customer impulsive purchase behavior: a case of Ho Chi Minh City in the 4.0 era. *Journal of International Economics and Management*, *30*(3), 17-41.

[4] Jijo, B. T. & Abdulazeez, A. M. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends, 2*(1), 20-28.

[5] Tangirala, S. (2020). Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm. *International Journal of Advanced Computer Science and Applications, 11*(2), 612-619.

[6] Luque, A., Carrasco, A., Martin, A., & Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition, 91*, 216-231.

[7] Gunawan, M. I., Sugiarto, D., & Mardianto, I. (2020). Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Search pada Algoritma Logistic Regression. *Jurnal Edukasi dan Penelitian Informatika, 6*(3), 280-284.

[8] Apriliah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021). Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest. *Jurnal Sistem Informasi, 10*(1), 163-171.