

## Stacking Machine Learning Model for Predict Hotel Booking Cancellations

### Model Machine Learning Stacking untuk Prediksi Pembatalan Pemesanan Hotel

Jus Prasetya<sup>1</sup>, Sefri Imanuel Fallo<sup>2</sup>, Moch Anjas Aprihartha<sup>3</sup>

<sup>1</sup> Alumni Magister Matematika, Universitas Gadjah Mada

<sup>2</sup> Program Studi Matematika, Universitas San Pedro

<sup>3</sup> Program Studi PJJ Informatika, Universitas Dian Nuswantoro

Email: <sup>1</sup>[jusprasetya777@gmail.com](mailto:jusprasetya777@gmail.com), <sup>2</sup>[fallosefrimanuel@gmail.com](mailto:fallosefrimanuel@gmail.com), <sup>3</sup>[moch.anjas.aprihartha@gmail.com](mailto:moch.anjas.aprihartha@gmail.com)

Received: 5 January 2024, revised: 12 February 2024, accepted: 15 February 2024

#### Abstract

The hotel prepares rooms and resources according to the room booking. Advance booking from customers is a relationship between customers and hotels that ensures price stability for customers to enjoy services. Cancellation of hotel bookings and inability to satisfy potential customers is a widespread and alarming problem that can increase hotel operating costs and affect customer satisfaction. Given that the impact on the hospitality industry can be very bad, predicting hotel cancellations can be a solution to help build an appropriate operational strategy. Method used in this research is stacking machine learning model. Stacking consists of two levels, where in this study level 0 (base learner) uses the Naive Bayes, Logistic Regression, and Gradient Boosting Machine algorithms while at level 1 (meta learner) uses the Random Forest algorithm. Accuracy value of the stacking model classification and the gradient boosting machine has the highest accuracy value of 0.87. Sensitivity value of the stacking model is 0.86 and is the highest sensitivity value which means that the stacking model classification is very precise in predicting consumers in canceling hotel reservations. Specificity value of the gradient boosting machine is 0.88 and is the highest specificity value, which means that the gradient boosting machine classification is very precise in predicting consumers who do not cancel hotel reservations. Naive bayes and logistic regression classifications have accuracy, sensitivity, specificity, precision values that are not high.

**Keywords:** Stacking Model, Base Learner, Meta Learner

#### Abstrak

Hotel mempersiapkan kamar dan sumber daya sesuai dengan pemesanan kamar. Pemesanan di awal dari pelanggan merupakan hubungan antara pelanggan dengan hotel yang memastikan kestabilan harga bagi pelanggan untuk menikmati layanan. Pembatalan pemesanan hotel dan



ketidakmampuan untuk memuaskan calon konsumen merupakan masalah yang meluas dan mengkhawatirkan yang dapat meningkatkan biaya operasional hotel dan mempengaruhi kepuasan pelanggan. Mengingat hal itu dampaknya terhadap industri perhotelan bisa sangat buruk, maka dengan memprediksi pembatalan hotel dapat menjadi solusi untuk membantu membangun strategi operasional yang sesuai. Metode yang digunakan pada penelitian ini adalah *stacking machine learning model*. *Stacking* terdiri dari dua level, dimana pada penelitian ini level 0 (*base learner*) menggunakan algoritma *Naive Bayes*, *Logistic Regression*, dan *Gradient Boosting Machine* sedangkan pada level 1 (*meta learner*) menggunakan algoritma *Random Forest*. nilai akurasi klasifikasi *stacking model* dan *gradient boosting machine* memiliki nilai akurasi tertinggi sebesar 0.87. Nilai sensitivitas *stacking model* sebesar 0.86 dan merupakan nilai sensitivitas tertinggi yang berarti klasifikasi *stacking model* sangat tepat memprediksi konsumen dalam pembatalan pemesanan hotel. Nilai Spesifisitas *gradient boosting machine* sebesar 0.88 dan merupakan nilai spesifisitas tertinggi yang berarti klasifikasi *gradient boosting machine* sangat tepat memprediksi konsumen yang tidak melakukan pembatalan pemesanan hotel. Klasifikasi *naive bayes* dan *logistic regression* memiliki nilai akurasi, sensitivitas, spesifisitas, presisi yang tidak tinggi.

**Kata kunci:** *Stacking Model, Base Learner, Meta Learner*

## 1. PENDAHULUAN

Pariwisata adalah salah satu industri yang paling berkembang di dunia dan sangat penting dalam ekonomi global. Dalam hal pembangunan ekonomi, industri pariwisata menghasilkan pendapatan bagi negara melalui konsumsi barang dan jasa oleh wisatawan, pajak, pengembangan usaha, lapangan kerja, dan lain-lain. Namun, di lingkungan pariwisata mengembangkan aktivitasnya selalu dihadapkan dengan ketidakpastian. Dalam hal ini, industri pariwisata sangat sensitif terhadap berbagai faktor eksternal yang dapat berdampak signifikan terhadap pendapatan, seperti ketidakstabilan politik, cuaca, bencana alam, dan lainnya. Selain itu, industri hotel menawarkan produk yang tidak dapat disimpan, sehingga setiap kamar yang tidak dihuni menyebabkan hilangnya pemasukan pendapatan [17].

Dalam industri perhotelan, hotel mempersiapkan kamar dan sumber daya sesuai dengan pemesanan kamar. Pemesanan di awal dari pelanggan merupakan hubungan antara pelanggan dengan hotel yang memastikan kestabilan harga bagi pelanggan untuk menikmati layanan, tetapi pada saat yang sama dapat meningkatkan biaya pengeluaran dan manajemen sumber daya untuk hotel [7]. Pemesanan kamar hotel secara online, pihak konsumen biasanya memiliki pilihan untuk memesan tanpa pembatalan gratis atau pembatalan gratis dengan harga premium. Kebijakan pembatalan gratis selalu dianggap sebagai asuransi terhadap ketidakpastian perubahan rencana perjalanan. Kebijakan ini terkait dengan perilaku konsumen karena meningkatnya popularitas harga hotel yang dinamis. Dalam konteks ini, pemilihan tarif pembatalan gratis tidak hanya menjamin pengembalian dana jika terjadi perubahan rencana perjalanan tetapi juga memberikan kesempatan kepada konsumen untuk membatalkan dan memesan ulang kamar hotel dengan harga yang lebih rendah jika harga turun sebelum tanggal menginap [11].

Pembatalan pemesanan hotel dan ketidakmampuan untuk memuaskan calon konsumen merupakan masalah yang meluas dan mengkhawatirkan yang dapat meningkatkan biaya operasional hotel dan mempengaruhi kepuasan pelanggan. Efisiensi hotel sangat terdampak, terutama ketika reservasi dibatalkan pada menit-menit terakhir, yang mengarah pada pemanfaatan sumber daya hotel yang tidak optimal. Mengingat hal itu dampaknya terhadap industri perhotelan bisa sangat buruk, maka dengan memprediksi pembatalan hotel dapat menjadi solusi untuk membantu membangun strategi operasional yang sesuai [6]. Pelaku bisnis perhotelan dapat memperoleh manfaat dari prediksi pembatalan, prediksi dilakukan untuk mengidentifikasi pelanggan dengan risiko tertinggi untuk melakukan pembatalan diawal, sehingga menghindari

kerugian besar [10].

Penelitian yang dilakukan [1] membahas prediksi pembatalan hotel untuk mengurangi ketidakpastian dan meningkatkan pendapatan dengan menggunakan lima metode *supervised learning*. [2] juga melakukan penelitian serupa dengan menggunakan metode *XGboost*. Penelitian Sánchez dan Eleazar [17] membahas metode *random forest*, SVM, C.5.0 dan ANN untuk memprediksi pembatalan pemesanan hotel. Penelitian [7] membahas perbandingan metode *logistic regression*, k-NN dan *CatBoost* dan menyimpulkan bahwa metode *CatBoost* merupakan model yang paling baik dalam melakukan prediksi. Penelitian [6] membahas prediksi pembatalan pemesanan hotel dengan menggunakan metode *bayesian networks* dan *lasso regression* dan menunjukkan bahwa penggabungan *bayesian networks* dan model linear/non-linear model *machine learning* menghasilkan kinerja prediksi terbaik. Beberapa penelitian yang telah disebutkan sebelumnya menggunakan beberapa metode *machine learning* dengan membandingkan hasil kinerja setiap metode tersebut dan setiap metode dianggap sebagai *individual learners*. Sejauh ini, belum ada yang melakukan penelitian pembatalan pemesanan hotel dengan algoritma *machine learning* dua level yakni level-0 (*base learner*) dan level-1 (*meta learner*).

Pada penelitian ini akan dilakukan prediksi pembatalan pemesanan hotel dengan menggunakan model *stacking (stacked models)*. *Stacking* adalah metode dengan menggunakan lebih dari satu algoritma *machine learning*. Terdapat dua level yang digunakan dalam metode *stacking*. Pada level-0 data pelatihan akan dilatih dengan model algoritma *naive bayes*, *logistic regression*, dan *gradient boosting machine*, kemudian hasil output tersebut digabungkan dan akan menjadi input untuk pembelajaran pada level-1, algoritma yang digunakan pada level-1 yaitu *random forest*. Dengan integrasi output dari beberapa model, metode ini dapat meningkatkan kinerja prediksi.

## 2. METODOLOGI

### 2.1 Data dan Variabel Penelitian

Data pada penelitian ini bersumber dari *Kaggle Datasets* yang merujuk pada data artikel *Data in Brief-Hotel Booking Demand Datasets* [3]. Data terdiri dari 18 variabel dan 36275 sampel data. Variabel *booking status* merupakan variabel *dependent* (respon) yang bernilai pemesanan dibatalkan (1) atau tidak (0). Adapun variabel yang digunakan disajikan pada Tabel 2.1.

**Tabel 2.1.** Variabel yang digunakan

Variabel	Tipe	Deskripsi
no of adults	Numerik	Jumlah orang dewasa
no of children	Numerik	Jumlah anak
no of weekend nights	Numerik	Jumlah akhir pekan (sabtu atau minggu) tamu yang menginap atau memesan
no of week nights	Numerik	Jumlah hari (senin-jumat) tamu menginap atau memesan
type of meal plan	Kategorik	Jenis makanan yang dipesan
required car parking space	Numerik	Jumlah ruang parkir yang dibutuhkan oleh tamu
room type reserved	Kategorik	Kode tipe kamar yang dipesan.

lead time	Numerik	Jumlah hari yang telah berlalu antara tanggal masuknya pemesanan ke dalam sistem database dan tanggal kedatangan
arrival year	Numerik	Tahun kedatangan
arrival month	Kategorik	Bulan kedatangan
arrival date	Numerik	Tanggal kedatangan
market segment type	Kategorik	Penetapan segmen pasar
repeated guest	Kategorik	Nilai yang menunjukkan apakah nama pemesanan berasal dari tamu berulang (1), tidak (0)
no of previous cancellations	Numerik	Jumlah pemesanan sebelumnya yang dibatalkan oleh tamu sebelum pemesanan saat ini
no of previous bookings not canceled	Numerik	Jumlah pemesanan sebelumnya yang tidak dibatalkan oleh tamu sebelum pemesanan saat ini
avg price per room	Numerik	Rata-rata Tarif Harian
no of special requests	Numerik	Jumlah permintaan khusus yang dibuat oleh tamu
booking status	Kategorik	Nilai yang menunjukkan apakah pemesanan dibatalkan (1), tidak (0)

## 2.2 Metode Penelitian

Metode yang digunakan pada penelitian ini adalah *stacking machine learning model*. *Stacking* terdiri dari dua level, dimana pada penelitian ini level 0 (*base learner*) menggunakan algoritma *Naive Bayes*, *Logistic Regression*, dan *Gradient Boosting Machine* sedangkan pada level 1 (*meta learner*) menggunakan algoritma *Random Forest*. Pendekatan pertama dalam melakukan *stacking* adalah melatih model masing-masing *base learner* secara terpisah dan kemudian mengintegrasikan output dari beberapa model untuk menjadi input pada *meta learner*.

Setelah melatih masing-masing model pada *base learner*, pada proses menumpuk (*stacking*) terdapat beberapa hal khusus yaitu :

1. Semua model harus dilatih pada training set yang sama
2. Semua model harus dilatih dengan jumlah lipatan CV yang sama
3. Semua model harus menggunakan pembagian lipatan yang sama

Prediksi yang telah divalidasi silang dari semua model harus dipertahankan. Data ini yang digunakan untuk melatih algoritma *meta learner* [4].

## 2.3 Stacking Model

Stacking adalah metode ensemble learning yang dibuat dengan menggunakan lebih dari satu algoritma machine learning. Terdapat dua level (level 0 dan 1) yang digunakan dalam metode stacking. Single model (*base learner*) digunakan di Level-0 dan *meta learner* digunakan di Level-1. Pada level 1, model prediksi diperoleh dari beberapa algoritma machine learning pada Level-0

yang digabungkan melalui meta-model. Dengan kata lain, output dari *base learner* digunakan sebagai nilai input untuk *meta learner* [5]. Hasil prediksi dari *stacking* bergantung pada kinerja setiap *base learner* yang digunakan untuk *stacking*, jumlah algoritma base-model yang dipilih akan mempengaruhi waktu perhitungan [18].

Algoritma *meta learner* terdiri dari tiga tahap :

1. Siapkan metode ensemble
  - Tentukan daftar algoritma  $L$  *base learner*
  - Tentukan algoritma *meta learner*
2. Latih metode ensemble
  - Latih setiap algoritma  $L$  *base learner* dengan data training
  - Lakukan *k-fold cross validation* pada setiap *base learner* dan kumpulkan prediksi yang telah di validasi silang dari setiap *base learner* (penggunaan *k-fold cross validation* harus sama pada setiap *base learner*). Nilai prediksi mewakili  $p_1, \dots, p_L$  pada persamaan (2.1).
  - Nilai prediksi yang telah divalidasi silang  $N$  dari setiap algoritma  $L$  dapat digabungkan membentuk matriks baru berukuran  $N \times L$  (diwakili oleh  $Z$  pada persamaan (2.1). Matriks  $Z$  bersama vektor respon ( $y$ ) disebut data level-1).

$$N \left\{ \begin{bmatrix} p_1 \\ \vdots \\ p_L \end{bmatrix} \right\} \left[ y \right] \rightarrow N \left\{ \begin{bmatrix} Z \\ \vdots \\ Z \end{bmatrix} \right\} \left[ y \right] \quad (2.1)$$

- Latih algoritma *meta learner* pada data level-1 ( $y = f(Z)$ ). Model ensemble terdiri dari model *base learner*  $L$  dan model *meta learner*, yang kemudian dapat digunakan untuk menghasilkan prediksi pada data baru.
3. Memprediksi data baru
    - Untuk menghasilkan prediksi ensemble, pertama hasilkan prediksi dari *base learner*.

Masukkan prediksi tersebut ke dalam *meta learner* untuk menghasilkan prediksi ensemble [4].

## 2.4 Naive Bayes

Teori keputusan bayesian adalah teknik fundamental statistik dalam bidang klasifikasi. Teori keputusan bayesian bertujuan untuk meminimalkan probabilitas membuat keputusan yang salah, atau risiko yang diharapkan. Rumus bayes dapat ditulis dalam bentuk sederhana dengan dihubungkan dengan rumus probabilitas bersyarat yaitu :

$$P\{B|A\} = \frac{P\{A|B\}P\{B\}}{P\{A\}} \quad (2.2)$$

Pada data dengan variabel kontinu, untuk menentukan probabilitas bersyarat  $P\{A|B\}$  maka diasumsikan variabel kontinu berdistribusi normal dengan parameter *mean* dan *variance*. Maka fungsi kepadatan probabilitas dapat mengevaluasi probabilitas bersyarat  $P\{A_i|B_j\}$  untuk setiap kelas secara terpisah dengan persamaan sebagai berikut [15] :

$$P\{A_i|B_j\} = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left(-\frac{(A_i - \mu_{ij})^2}{2\sigma_{ij}^2}\right) \quad (2.3)$$

## 2.5 Logistic Regression

Metode statistik yang paling sering digunakan untuk klasifikasi biner adalah regresi logistik. Pada regresi logistik biner hanya ada dua kategori pada variabel dependen. Persamaan (2.4) merupakan model regresi logistik secara umum dengan terdapat  $p$  variabel independen ( $X_1, X_2, \dots, X_p$ )

$$\pi(X) = E(Y|X) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}} \quad (2.4)$$

Model pada persamaan (2.4) dapat ditransformasi sehingga dapat diinterpretasikan dengan mudah. Transformasi yang digunakan adalah transformasi logit. Maka model menjadi seperti persamaan (2.5) berikut

$$g(X) = \ln \left[ \frac{\pi(X)}{1 - \pi(X)} \right] = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \quad (2.5)$$

Parameter  $\beta$  diestimasi dengan menggunakan *maximum likelihood estimation*. Tahapan-tahapan dalam pemodelan regresi logistik adalah mengestimasi parameter model regresi logistik, melakukan uji signifikansi parameter parameter secara serentak dengan menggunakan uji *likelihood ratio*, melakukan uji signifikansi parameter secara parsial menggunakan uji *Wald*, jika parameter tidak signifikan dalam model, maka variabel tersebut dikeluarkan dalam model regresi logistik, melakukan evaluasi *goodness of fit* model [14].

## 2.6 Gradient Boosting Machine

*Gradient boosting machine* adalah metode ensemble yang melakukan regresi atau klasifikasi dengan menggabungkan output dari masing-masing pohon. *Gradient boosted trees* menggunakan metode yang disebut *boosting* yang menggabungkan *weak learners* secara berurutan sehingga setiap pohon baru mengoreksi kesalahan dari pohon sebelumnya. *Weak learners* biasanya berupa pohon keputusan dengan hanya satu cabang yang disebut *tunggal keputusan*. Langkah pertama yang dilakukan adalah mengisi pohon keputusan tunggal dengan nilai awal kemudian evaluasi dilakukan untuk melihat seberapa baik pohon ini menggunakan fungsi kerugian (*loss function*). Beberapa fungsi kerugian ( $L$ ) misalnya untuk kasus klasifikasi adalah *cross-entropy* dan untuk kasus regresi adalah *root mean square error* (RMSE) [12].

Pada kasus dengan respon kategorik, variabel respon  $y$  biasanya mengambil nilai biner  $y \in \{0,1\}$ , diasumsikan bahwa variabel tersebut berasal dari distribusi bernoulli. Untuk menyederhanakan notasi, diasumsikan label  $\bar{y}$  yang telah ditransformasi, dengan menuliskan  $\bar{y} = 2y - 1$  dan membuat  $\bar{y} \in \{-1,1\}$ . Dalam hal ini, probabilitas respon per kelas dapat diestimasi dengan meminimumkan *negative log-likelihood*, diasosiasikan dengan label kelas baru:

$$\Psi(y, f)_{Bern} = \log(1 + \exp(-2\bar{y}f)) \quad (2.6)$$

Fungsi kerugian ini biasanya disebut sebagai kerugian bernoulli [13].

## 2.7 Random Forest

Definisi umum dari random forest yang diberikan oleh Breiman., (2001), sebagai berikut :

**Definition 2.1** (*Random forest*) diberikan  $(\hat{h}(\cdot, \Theta_1), \dots, \hat{h}(\cdot, \Theta_q))$  merupakan kumpulan pohon prediktor dengan  $\Theta_1, \dots, \Theta_q$  dimana  $q$  merupakan variabel random independen i.i.d dari  $\mathcal{L}_n$ . Prediktor random forest  $\hat{h}_{RF}$  diperoleh dengan menggabungkan kumpulan random trees. Penggabungan dilakukan sebagai berikut [9] :

Rata-rata prediksi pohon individu dalam regresi :

$$\hat{h}_{RF}(x) = \frac{1}{q} \sum_{\ell=1}^q \hat{h}(x, \Theta_{\ell}) \quad (2.7)$$

Majority vote diantara prediksi pohon individu dalam klasifikasi :

$$\hat{h}_{RF}(x) = \underset{1 \leq c \leq C}{\operatorname{argmax}} \sum_{\ell=1}^q 1_{\hat{h}(x, \Theta_{\ell})=c} \quad (2.8)$$

## 2.8 Evaluasi Model Klasifikasi

Matriks klasifikasi adalah metode paling intuitif untuk melihat kinerja klasifikasi. Matriks ini disebut juga dengan *confusion matrix*. Secara visual, matriks menunjukkan distribusi kelas aktual dan kelas yang diprediksi seperti yang disajikan pada Tabel 2.2.

**Tabel 2.2.** Matriks klasifikasi 2-kelas

Klasifikasi 2-kelas		Kelas Prediksi	
		Kelas Positif	Kelas Negatif
Kelas Aktual	Kelas Positif	True Positive (TP)	False Negative (FN)
	Kelas Negatif	False Positive (FP)	True Negative (TN)

Akurasi model diperoleh dari elemen diagonal matriks klasifikasi yang merupakan hasil pengklasifikasian yang benar yaitu kelas aktual positif diprediksi tepat sebagai kelas positif dan kelas aktual negatif diprediksi tepat sebagai kelas negatif [16].

$$\text{Akurasi} = \frac{TP + TN}{TP + FN + FP + TN} \quad (2.9)$$

Matriks klasifikasi juga memiliki empat ukuran dasar yaitu *sensitivity*, *specificity*, *precision*, dan *negative predictive value* [8]:

$$\text{sensitivity, recall, TPR} = \frac{TP}{TP + FN} \quad (2.10)$$

$$\text{specificity, TNR} = \frac{TN}{TN + FP} \quad (2.11)$$

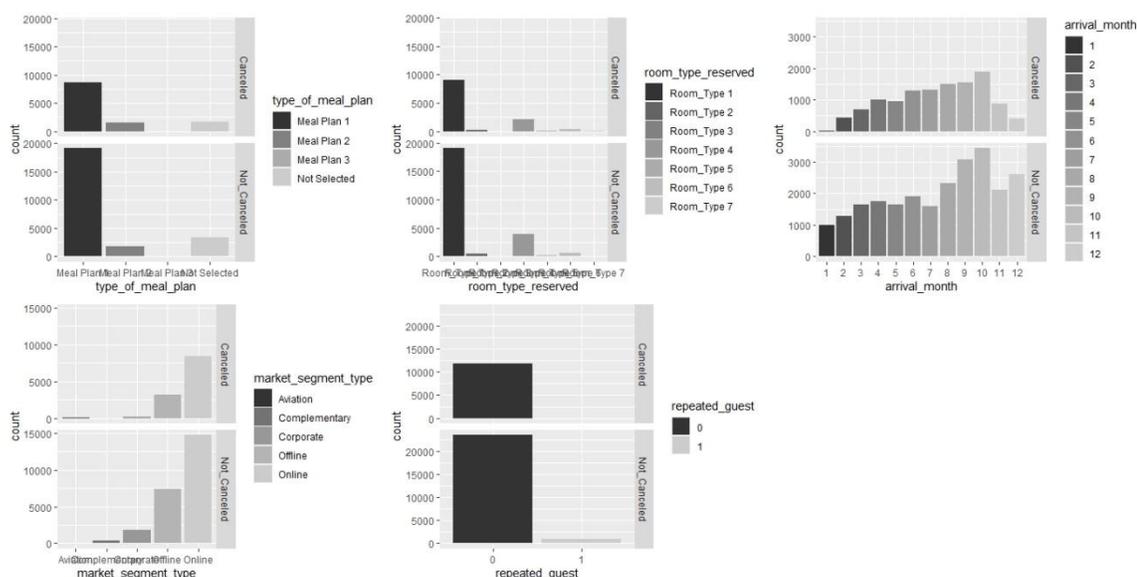
$$\text{precision, PPV} = \frac{TP}{TP + FP} \quad (2.12)$$

$$\text{negative predictive value} = \frac{TN}{TN + FN} \quad (2.13)$$

### 3. HASIL DAN PEMBAHASAN

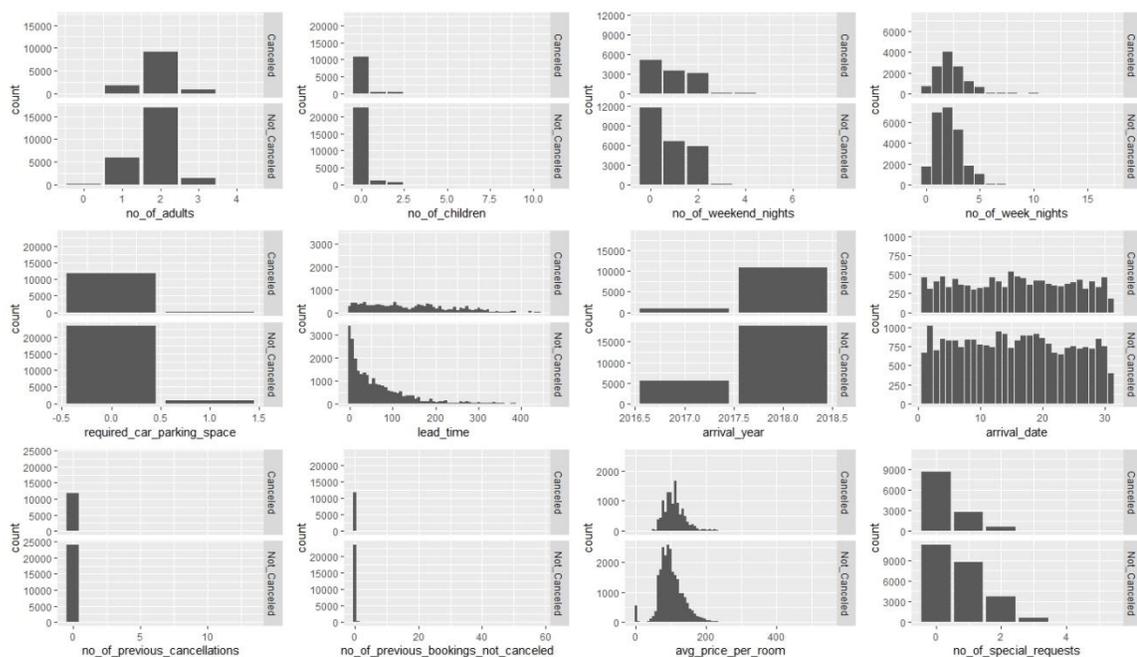
#### 3.1 Eksplorasi Data

Eksplorasi data dilakukan dengan membuat grafik pada variabel independen diasosiasikan dengan variabel dependen (*booking status*). Gambar 3.1 merupakan diagram batang (*bar plot*) variabel independen kategorik yang dihubungkan dengan variabel respon *booking status*, pada gambar 1 menjelaskan bahwa mayoritas konsumen memesan makanan *meal plan 1* pada transaksi booking hotel. Room type 1 merupakan tipe kamar yang paling banyak di pesan oleh konsumen. Pada bulan september dan oktober intensitas konsumen memesan hotel sangat tinggi namun pembatalan pemesanan hotel juga tinggi pada kedua bulan tersebut sedangkan pada bulan desember pemesanan hotel cukup tinggi namun pembatalannya sangat rendah. Segmen pasar online merupakan segmen pasar paling sukses dalam mendatangkan konsumen. Konsumen yang pernah menginap sebelumnya tidak akan membatalkan transaksi pemesanan hotel selanjutnya.



Gambar 3.1. Asosiasi Variabel Independen Kategorik dan Respon

Gambar 3.2 merupakan plot variabel independen numerik yang dihubungkan dengan variabel respon *booking status*, pada gambar 2 menjelaskan bahwa mayoritas konsumen melakukan pemesanan hotel untuk dua orang dewasa dan mayoritas tidak membawa anak. Konsumen menginap didominasi pada hari kerja daripada hari akhir pekan. Mayoritas konsumen melakukan pemesanan hotel kurang dari 100 hari sebelum masa *check in* ke hotel dan juga sedikit konsumen yang memesan ruang parkir. Pada tanggal 15 banyak konsumen melakukan pembatalan pemesanan hotel. Harga rata-rata kamar yang dipesan konsumen dibawah harga \$100 dan cukup banyak konsumen meminta pelayanan tambahan seperti permintaan *twin bed* atau *high floor*.



Gambar 3.2. Asosiasi Variabel Independen Numerik dan Respon

### 3.2 Klasifikasi Pada Algoritma *Base Learner*

Sebelum melakukan teknik klasifikasi pada algoritma *base learner*, data akan dibagi menjadi dua bagian yaitu data training yang berfungsi membuat model klasifikasi pada masing-masing *base learner* dan data testing yang berfungsi menguji kebaikan model yang dibuat oleh data training. Pembagian data menggunakan fungsi `initial_split()` dengan strata pada variabel respon (*booking status*) yang berarti data training dan data testing memiliki proporsi kelas yang sama dengan data aslinya. Seperti pada Tabel 3.1 berikut

Tabel 3.1. Proporsi kelas pada data

Data	Instance (n)	Proporsi Kelas	
		Canceled	Not Canceled
Data Asli	36275	0.3276361	0.6723639
Data Training	29020	0.3276361	0.6723639
Data Testing	7255	0.3276361	0.6723639

Data training akan dilatih pada setiap algoritma *base learner*. Pada penelitian ini algoritma *base learner* terdiri dari *naive bayes*, *logistic regression*, dan *gradient boosting machine*.

Berdasarkan hasil *running* program R-Studio, untuk *base learner naive bayes* dengan *5-fold cross validation* dan *fold assignment* adalah modulo. Maka model algoritma *naive bayes* yang dihasilkan berdasarkan *confusion matrix* seperti pada Tabel 3.2 berikut ini.

Tabel 3.2. Klasifikasi *base learner naive bayes*

Klasifikasi NB	Kelas Prediksi	
	Canceled	Not Canceled
Kelas Aktual Canceled	911	281
Kelas Aktual Not Canceled	1466	4597

Tabel 3.2 menjelaskan bahwa konsumen yang membatalkan pemesanan hotel tepat diprediksi membatalkan pemesanan hotel sebanyak 911 konsumen dan sisanya 281 misklasifikasi. Konsumen yang tidak membatalkan pemesanan hotel tepat diprediksi tidak membatalkan pemesanan hotel sebanyak 4597 konsumen dan sisanya 1466 misklasifikasi. Berdasarkan persamaan (2.9), (2.10), (2.11), (2.12) dan (2.13) diperoleh nilai akurasi sebesar 0.75, nilai sensitivitas sebesar 0.76, nilai spesifisitas sebesar 0.75, nilai presisi sebesar 0.38, dan negative predictive value sebesar 0.94.

Berdasarkan hasil *running* program R-Studio, untuk *base learner logistic regression* dengan *5-fold cross validation* dan *fold assignment* adalah modulo. Maka model algoritma *logistic regression* yang dihasilkan berdasarkan *confusion matrix* seperti pada Tabel 3.3 berikut ini.

**Tabel 3.3.** Klasifikasi *base learner logistic regression*

Klasifikasi LR		Kelas Prediksi	
		Canceled	Not Canceled
Kelas Aktual	Canceled	1315	350
	Not Canceled	1062	4528

Tabel 3.3 menjelaskan bahwa konsumen yang membatalkan pemesanan hotel tepat diprediksi membatalkan pemesanan hotel sebanyak 1315 konsumen dan sisanya 350 misklasifikasi. Konsumen yang tidak membatalkan pemesanan hotel tepat diprediksi tidak membatalkan pemesanan hotel sebanyak 4528 konsumen dan sisanya 1062 misklasifikasi. Berdasarkan persamaan (2.9), (2.10), (2.11), (2.12) dan (2.13) diperoleh nilai akurasi sebesar 0.80, nilai sensitivitas sebesar 0.78, nilai spesifisitas sebesar 0.81, nilai presisi sebesar 0.55, dan negative predictive value sebesar 0.92.

Berdasarkan hasil *running* program R-Studio, untuk *base learner gradient boosting machine* dengan *5-fold cross validation* dan *fold assignment* adalah modulo. Maka model algoritma *gradient boosting machine* yang dihasilkan berdasarkan *confusion matrix* seperti pada Tabel 3.4 berikut ini.

**Tabel 3.4.** Klasifikasi *base learner gradient boosting machine*

Klasifikasi GBM		Kelas Prediksi	
		Canceled	Not Canceled
Kelas Aktual	Canceled	1789	315
	Not Canceled	588	4563

Tabel 3.4 menjelaskan bahwa konsumen yang membatalkan pemesanan hotel tepat diprediksi membatalkan pemesanan hotel sebanyak 1789 konsumen dan sisanya 315 misklasifikasi. Konsumen yang tidak membatalkan pemesanan hotel tepat diprediksi tidak membatalkan pemesanan hotel sebanyak 4563 konsumen dan sisanya 588 misklasifikasi. Berdasarkan persamaan (2.9), (2.10), (2.11), (2.12) dan (2.13) diperoleh nilai akurasi sebesar 0.87, nilai sensitivitas sebesar 0.85, nilai spesifisitas sebesar 0.88, nilai presisi sebesar 0.75, dan negative predictive value sebesar 0.93.

### 3.2 Klasifikasi Pada Algoritma *Stacking Model*

Setelah membuat model pada masing-masing algoritma *base learner* yaitu *naive bayes*, *logistic regression* dan *gradient boosting machine*. Maka model *stacking* dapat dihasilkan dengan menjadikan hasil prediksi (*output*) dari algoritma *base learner* menjadi input untuk algoritma *meta learner* untuk menghasilkan prediksi ensemble. Algoritma *meta learner* yang digunakan pada penelitian ini adalah *random forest*. Berdasarkan hasil *running* program R-Studio, maka

model algoritma *stacking model* yang dihasilkan berdasarkan *confusion matrix* seperti pada Tabel 3.5 berikut ini.

**Tabel 3.5.** Klasifikasi *Stacking model*

Klasifikasi Stacking model		Kelas Prediksi	
		Canceled	Not Canceled
Kelas Aktual	Canceled	1729	278
	Not Canceled	648	4600

Tabel 3.5 menjelaskan bahwa konsumen yang membatalkan pemesanan hotel tepat diprediksi membatalkan pemesanan hotel sebanyak 1729 konsumen dan sisanya 278 misklasifikasi. Konsumen yang tidak membatalkan pemesanan hotel tepat diprediksi tidak membatalkan pemesanan hotel sebanyak 4600 konsumen dan sisanya 648 misklasifikasi. Berdasarkan persamaan (2.9), (2.10), (2.11), (2.12) dan (2.13) diperoleh nilai akurasi sebesar 0.87, nilai sensitivitas sebesar 0.86, nilai spesifisitas sebesar 0.87, nilai presisi sebesar 0.72, dan negative predictive value sebesar 0.94.

**Tabel 3.6.** Perbandingan evaluasi model setiap klasifikasi

	Klasifikasi			
	GBM	NB	LR	Stacking model
Akurasi	0.87	0.75	0.80	0.87
Sensitivitas	0.85	0.76	0.78	0.86
Spesifisitas	0.88	0.75	0.81	0.87
Presisi	0.75	0.38	0.55	0.72
Negative predictive value	0.93	0.94	0.92	0.93

Tabel 3.6 menjelaskan bahwa nilai akurasi klasifikasi *stacking model* dan *gradient boosting machine* memiliki nilai akurasi tertinggi sebesar 0.87. Nilai sensitivitas *stacking model* sebesar 0.86 dan merupakan nilai sensitivitas tertinggi yang berarti klasifikasi *stacking model* sangat tepat memprediksi konsumen dalam pembatalan pemesanan hotel. Nilai Spesifisitas *gradient boosting machine* sebesar 0.88 dan merupakan nilai spesifisitas tertinggi yang berarti klasifikasi *gradient boosting machine* sangat tepat memprediksi konsumen yang tidak melakukan pembatalan pemesanan hotel. Klasifikasi *naive bayes* dan *logistic regression* memiliki nilai akurasi, sensitivitas, spesifisitas, presisi yang tidak tinggi.

Adapun hasil penelitian di atas menunjukkan bahwa metode *ensemble* seperti *stacking model* dan *gradient boosting machine* memiliki nilai akurasi, sensitivitas dan spesifisitas yang tinggi karena metode *ensemble* merupakan metode dimana beberapa model yang dihasilkan akan digabungkan untuk meningkatkan kinerja keseluruhan sehingga dapat meningkatkan nilai akurasi dan mereduksi *overfitting* karena kesalahan prediksi yang dapat terjadi pada satu model dapat diganti dengan prediksi yang benar dari model lainnya, hasil yang didapatkan pada penelitian ini senada dengan penelitian Buyrukoglu dan Savas [5] yang menerapkan metode *stacking model* untuk penentuan posisi pesepakbola dengan algoritma *base learner* yaitu *gradient boosting*, *random forest*, *deep neural network* dan algoritma *meta learner* yaitu *logistic regression* dan hasilnya menjelaskan bahwa metode *stacking model* memiliki nilai akurasi, sensitivitas dan presisi yang tinggi yakni 0.82, 0.83, 0.83 dan *gradient boosting machine* nilai akurasi, sensitivitas dan presisi yang tinggi juga yakni 0.82, 0.82, 0.82 dibandingkan dengan metode *deep neural network* dan *random forest*. Penelitian [7] yang membandingkan beberapa algoritma *machine learning* untuk memprediksi pembatalan pemesanan hotel menjelaskan bahwa metode CatBoost yang merupakan metode *ensemble* memiliki nilai akurasi tertinggi dibandingkan dengan *k-nearest neighbor* dan *logistic regression*. Penelitian Wang, dkk [18] menerapkan metode *stacking model*

untuk prediksi celah pita material hasilnya menjelaskan bahwa metode *stacking model* (pada kasus regresi) memiliki nilai koefisien determinasi ( $R^2$ ) tertinggi sebesar 0.93 dan memiliki nilai RMSE, MAE, MAPE terendah sebesar 0.369, 0.211, 0.109 dibandingkan 10 metode lainnya. Berdasarkan pemaparan di atas menjelaskan bahwa metode ensemble memiliki nilai akurasi yang tinggi seperti yang dijelaskan pada penelitian [7], dan metode *stacking model* pada kasus klasifikasi memiliki nilai akurasi, sensitivitas, spesifisitas, presisi yang tinggi dan pada kasus regresi memiliki nilai koefisien determinasi ( $R^2$ ) yang tinggi seperti yang dijelaskan pada penelitian [5][18].

#### **4. KESIMPULAN**

Berdasarkan hasil pembahasan maka dapat disimpulkan pada proses *stacking* (menumpuk), algoritma pada setiap *base learner* harus dilatih pada data training yang sama, model harus dilatih dengan jumlah lipatan *cross validation* yang sama (pada penelitian ini menggunakan *5-fold cross validation*), model harus menggunakan pembagian lipatan yang sama (menggunakan fungsi *fold assignment* “modulo”). Nilai akurasi klasifikasi *stacking model* dan *gradient boosting machine* memiliki nilai akurasi tertinggi sebesar 0.87. Nilai sensitivitas *stacking model* sebesar 0.86 dan merupakan nilai sensitivitas tertinggi yang berarti klasifikasi *stacking model* sangat tepat memprediksi konsumen dalam pembatalan pemesanan hotel. Nilai Spesifisitas *gradient boosting machine* sebesar 0.88 dan merupakan nilai spesifisitas tertinggi yang berarti klasifikasi *gradient boosting machine* sangat tepat memprediksi konsumen yang tidak melakukan pembatalan pemesanan hotel. Nilai spesifisitas *stacking model* sebesar 0.87 yang berarti sangat tepat juga untuk memprediksi konsumen yang tidak melakukan pembatalan pemesanan hotel. Klasifikasi *naive bayes* dan *logistic regression* memiliki nilai akurasi, sensitivitas, spesifisitas, presisi yang tidak tinggi.

#### **DAFTAR PUSTAKA**

- [1]. Antonio, N., De Almeida, A. and Nunes, L., 2017. Predicting hotel booking cancellations to decrease uncertainty and increase revenue. *Tourism & Management Studies*, 13(2), pp.25-39.
- [2]. Antonio, N., de Almeida, A. and Nunes, L., 2017, December. Predicting hotel bookings cancellation with a machine learning classification model. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1049-1054). IEEE.
- [3]. Antonio, N., de Almeida, A. and Nunes, L., 2019. Hotel booking demand datasets. *Data in brief*, 22, pp.41-49.
- [4]. Boehmke, B. and Greenwell, B.M., 2019. *Hands-on machine learning with R*. CRC press.
- [5]. Buyrukoğlu, S. and Savaş, S., 2023. Stacked-based ensemble machine learning model for positioning footballer. *Arabian Journal for Science and Engineering*, 48(2), pp.1371-1383.
- [6]. Chen, S., Ngai, E.W., Ku, Y., Xu, Z., Gou, X. and Zhang, C., 2023. Prediction of hotel booking cancellations: Integration of machine learning and probability model based on interpretable feature interaction. *Decision Support Systems*, 170, p.113959.
- [7]. Chen, Y., Ding, C., Ye, H. and Zhou, Y., 2022, March. Comparison and analysis of machine learning models to predict hotel booking cancellation. In *2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022)* (pp. 1363-1370). Atlantis Press.
- [8]. Chicco, D. and Jurman, G., 2023. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, 16(1), pp.1-23.

- [9]. Genuer, R., Poggi, J.M., Genuer, R. and Poggi, J.M., 2020. *Random forests* (pp. 33-55). Springer International Publishing.
- [10]. Li, Y., Tang, S.Y., Johnson, J. and Lubarsky, D.A., 2019. Individualized no-show predictions: Effect on clinic overbooking and appointment reminders. *Production and Operations Management*, 28(8), pp.2068-2086.
- [11]. Masiero, L., Viglia, G. and Nieto-Garcia, M., 2020. Strategic consumer behavior in online hotel booking. *Annals of Tourism Research*, 83, p.102947.
- [12]. Mohamed, Z. and Shareef, H., 2022. An Adjustable Machine Learning Gradient Boosting-Based Controller for Pv Applications. *Available at SSRN 4229749*.
- [13]. Natekin, A. and Knoll, A., 2013. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7, p.21.
- [14]. Permai, S.D. and Herdianto, K., 2023. Prediction of Health Insurance Claims Using Logistic Regression and XGBoost Methods. *Procedia Computer Science*, 227, pp.1012-1019.
- [15]. Prasetya, J., 2022. Penerapan Klasifikasi Naive Bayes dengan Algoritma Random Oversampling dan Random Undersampling pada Data Tidak Seimbang Cervical Cancer Risk Factors. *Leibniz: Jurnal Matematika*, 2(2), pp.11-22.
- [16]. Ramasubramanian, K. and Singh, A., 2017. *Machine learning using R* (No. 1). New Delhi, India: Apress.
- [17]. Sánchez-Medina, A.J. and Eleazar, C., 2020. Using machine learning and big data for efficient forecasting of hotel booking cancellations. *International Journal of Hospitality Management*, 89, p.102546.
- [18]. Wang, T., Zhang, K., Thé, J. and Yu, H., 2022. Accurate prediction of band gap of materials using stacking machine learning model. *Computational Materials Science*, 201, p.110899.