Analisis Klasifikasi pada *Hotel Cancellation Demand* dengan Metode *Naïve Bayes, Logistic Regression, Random Forest,* dan *K-Nearest Neighbor* (KNN)

Zanzabila Rehanisya Firdhani[1*] and Kevina Windy Arlianni[2]
[1]06211940000029: Department Statistika, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[2]06211940000047: Department Statistika, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
*author: kevina.arlianni@gmail.com

ABSTRACT –Industri pariwisata adalah salah satu industri yang paling berkembang di dunia dan memiliki peranan penting dalam ekonomi global [2]. Hotel merupakan suatu bangunan yang menyediakan jasa penginapan, makanan, minuman, serta pelayanan lainnya untuk umum yang dikelola secara komersial teru-tama untuk para wisatawan. Usaha perhotelan yang kian marak di Indonesia menyebabkan tingkat persaingan dalam usaha ini menjadi tinggi sehingga setiap hotel akan berlomba untuk memberikan layanan yang terbaik apalagi ditengah penigkatan pemesanan hotel. Berdasarkan data sebelumnya telah terjadi 40000 pembatalan hotel dengan kerugian mencapai 1 triliun per bulan [6]. Bila mengacu dari data tersebut yakni banyaknya per-mintaan pembatalan pemesanan suatu hotel, maka perlu dilakukan analisis klasifikasi yang mengklasifikasikan faktor-faktor terjadinya pembatalan pemesanan suatu hotel. Untuk menganalisis faktor-faktor tersebut penelitian ini dapat dilakukan beberapa analisis klasifikasi yaitu *Naïve Bayes, regresi logistik, Random Forest dan K-Nearest Neighbor (KNN).* Data di lakukan preprocessing untuk membersihkan dan menyiapkan data lalu dibuat visualisasi data yang mendukung. Berdasarkan analisis klasifikasi menggunakan metode *Repeated Holdout* dan *K-Fold Cross Validation* pada 4 model, model terbaik untuk mengklasifikasikan data pembatalan reservasi hotel yaitu *Random Forest* dengan *K-Fold Cross Validation* yaitu nilai AUC sebesar 0.93 dengan akurasi sebesar 85.9%, sensitifitas sebesar 91.12% dan spesitifitas sebesar 76.99%.

Keywords–Analisis Klasisfikasi, KNN, *Logistic Regression, Naïve Bayes*, Pembatalan Hotel, *Random Forest.*

I. PENDAHULUAN

A. *Latar Belakang*

Dunia kini tengah menghadapi revolusi industri 4.0, dimana sektor industri kini didorong untuk melakukan inovasi dari berbagai sisi mulai dari metode pemasaran produk hingga pengembangan produk. Pariwisata merupakan salah satu industri yang mengalami evolusi dari konvensional menjadi lebih modern [1]. Fenomena ini dipengaruhi oleh permintaan pasar yang didominasi oleh generasi milenial. Industri pariwisata adalah salah satu industri yang paling berkembang di dunia dan memiliki peranan penting dalam ekonomi global [2]. Hal ini dapat dibuktikan pada tahun 1990 jumlah wisatawan mancanegara sedikit melebihi angka 400 juta, sedangkan pada tahun 2017 jumlahnya meningkat menjadi 1300 juta [3]. Industri pariwisata terdiri dari berbagai sektor, salah satunya adalah sektor perhotelan. Hotel merupakan suatu bangunan yang menyediakan jasa penginapan, makanan, minuman, serta pelayanan lainnya untuk umum yang dikelola secara komersial teru-tama untuk para wisatawan. Banyaknya keberadaan hotel memberikan berbagai pilihan fasilitas dan harga kamar yang beraneka ragam. Usaha perhotelan yang kian marak di Indonesia menyebabkan tingkat persaingan dalam usaha ini menjadi tinggi sehingga setiap hotel akan berlomba untuk memberikan layanan yang terbaik apalagi ditengah penigkatan pemesanan hotel.

Penelitian mengenai permintaan pemesanan hotel telah dilakukan oleh Wahyuni dan Wiweka [1]. Namun dari penelitian tersebut, belum mengakomodasi aspek atau faktor yang menyebabkan permintaan pemesanan hotel tersebut tinggi maupun faktor yang menyebabkan orang melakukan pembatalan hotel, padahal dari diketahuinya faktor tersebut semua hotel bisa bersaing untuk meningkatkan kualitas dari dua sisi. Berdasarkan data sebelumnya telah terjadi 40000 pembatalan hotel dengan kerugian mencapai 1 triliun per bulan [4]. Bila mengacu dari data tersebut yakni banyaknya per-mintaan pembatalan pemesanan suatu hotel, maka perlu dilakukan analisis klasifikasi yang mengklasifikasikan faktor-faktor terjadinya pembatalan pemesanan suatu hotel. Sehingga pihak hotel bisa melakukan suatu hal agar terhindar dari kejadian tersebut. Hal tersebut tentunya dilakukan dengan tujuan untuk mengurangi jumlah pembatalan pemesanan hotel. Untuk menganalisis faktor-faktor tersebut penelitian ini dapat dilakukan beberapa analisis klasifikasi dalam ilmu statistika, dimana pada analisis ini yang digunakan adalah *Naïve Bayes, regresi logistik, Random Forest dan K-Nearest Neighbor (KNN)*. Sebelum dilakukan analisis klasifikasi perlu dilakukan preprocessing terhadap data agar memastikan sebuah data layak untuk dilakukan analisis seperti pengecekan miss-ing value dan outlier. Setelah dilakukan analisis juga perlu dilakukan sebuah evaluasi agar model data yang telah dibentuk menghasilkan model yang baik. Evaluasi yang dilakukan pada model ini menggunakan *repeated holdout* dan *k-fold CV* serta dilihat akurasi modelnya.

B. *Rumusan Masalah*

Rumusan masalah berdasarkan latar belakang tersebut adalah

1. Mengklasifikasi dari faktor-faktor yang menyebabkan pembatalan pemesanan hotel
2. Mendeteksi metode klasifikasi paling baik untuk faktor penyebab pembatalan pemesanan hotel

C. *Tujuan Penelitian*

Berdasarkan rumusan masalah di atas, tujuan pada penelitian ini adalah

1. Untuk mengklasifikasi dari faktor-faktor yang menyebabkan pembatalan pemesanan hotel
2. Untuk mendeteksi metode klasifikasi paling baik untuk faktor penyebab pembatalan pemesanan hotel.

II. TINJAUAN PUSTAKA

A. *Statistika Deskriptif*

Statistika deskriptif adalah metode-metode yang berkaitan dengan pengumpulan dan penyajian suatu gugus data sehingga memberikan informasi yang berguna. Statistika deskriptif biasanya digambarkan dalam bentuk tabel, grafik dan diagram. Selain itu, statistika deskriptif menggambarkan perhitungan data kuantitatif seperti nilai rata-rata, nilai median, nilai minimum dan maksimum. Pada penelitian ini, statistika deskriptif yang digunakan adalah rata-rata dan standar deviasi. [5].

B. *Preprocessing Data*

Data mining adalah ekstraksi informasi atau pola yang penting atau menarik dari data yang ada di database yang besar. *Data mining* sangat menarik perhatian industri atau perusahaan masa kini dalam beberapa tahun belakangan ini karena tersedianya data dalam jumlah yang besar dan semakin besarnya kebutuhan untuk mengubah data tersebut menjadi informasi dan pengetahuan yang berguna. Teknik data mining digunakan untuk memeriksa basis data berukuran besar sebagai cara untuk menemukan pola yang baru dan berguna. Data mining merupakan bagian dari proses *Knowledge Discovery in Databases* (KDD). Salah satu tahapan dalam data mining adalah *preprocessing* data. Data *preprocessing* merupakan proses mengubah data mentah menjadi data yang berkualitas sehingga baik untuk menjadi inputan bagi proses data mining. Kegiatan dalam data preprocessing adalah sebagai berikut [6].

1. ***Data Cleaning***

   Data *cleaning* merupakan serangkaian proses untuk mengidentifikasi kesalahan pada data dan kemudian mengambil tindakan lanjut, baik berupa perbaikan ataupun penghapusan data yang tidak sesuai [7]. Ada beberapa hal yang harus ditangani dalam data *cleaning* yakni *missing value* yang merupakan atribut dari beberapa *record* yang nilainya tidak lengkap. Hal ini biasanya disebabkan karena kesalahan ketika proses pengumpulan data. Cara untuk menangani missing value adalah dengan cara mengisi field yang tidak lengkap dengan menggunakan nilai yang plausible berdasarkan keluaran dari algoritma tertentu [8], selanjutkan yakni outlier. Outlier adalah suatu data yang menyimpang dari sekumpulan data yang lain atau tidak mengikuti pola data secara keseluruhan. Dalam suatu kumpulan data biasanya terdapat 10% pengamatan yang outlier yang dapat diatasi dengan menghapus kasus outlier.

2. ***Data Transformasi***

   Data *transformation* dibutuhkan dalam implementasi data mining khususnya pada saat preprocessing data. Pada tahapan ini data diubah atau dikonsolidasikan sehingga proses penambangan yang dihasilkan dapat lebih efisien dan pola yang ditemukan dapat lebih mudah dipahami [9].

3. ***Feature Selection***

   Feature selection adalah suatu proses yang mencoba untuk menemukan subhimpunan dari himpunan fitur yang tersedia untuk meningkatkan aplikasi dari suatu algoritma pembelajaran [10]. *Feature selection* digunakan dibanyak area aplikasi sebagai alat un-tuk menghilangkan fitur yang tidak relevan dan atau fitur berlebihan. Sebuah fitur dikatakan tidak relevan jika memberikan sedikit informasi.

C. *Visualisasi Data*

Visualisasi data merupakan proses penyajian data dalam bentuk grafik yang membuat informasi mudah dimengerti [12]. Visualisasi data memungkinkan pengguna untuk memperoleh pengetahuan yang lebih banyak mengenai data mentah yang didapatkan dari berbagai sumber terpisah. Visualisasi data tidak hanya mengubah data men-jadi grafik visual, akan tetapi visualisasi data juga memerlukan perencanaan. Berbagai macam cara visualization data adalah menggunakan bar chart,histogram, pie chart, boxplot maupun scatterplot [13].

D. *Analisis Klasifikasi*

Proses klasifikasi dalam teknik data mining adalah sekumpulan data yang dapat menghasilkan suatu klasifikasi model (fungsi sasaran). Sehingga diperlukan sebuah dataset pada himpunan tersebut untuk proses klasifikasinya [.

1. ***Naïve Bayes***

   Naive Bayes merupakan sebuah metode pengklasifi-kasian probabilistic sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan . Naïve bayes digunakan untuk memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya. Persemaan dari teorema bayes adalah klasifikasi. Keuntungan menggunakan algoritma Naïve Bayes adalah bahwa jumlah data training yang kecil dapat menentukan estimasi parameter yang dibutuhkan [14]. Naïve Bayes memiliki tingkat akurasi dan kecepatan yang sangat kuat ketika diterapkan pada database dengan ukuran besar. Dalam persamaan teorema Bayes, probabilitas bersyarat adalah diekspresikan sebagai pada persamaan berikut.

$$P(H \mid X) = \frac{P(H \mid X).P(H)}{P(X)}$$

2. *Analisis Regresi Logistik*

Regresi logistik adalah sebuah metode statistik untuk menganalisis variabel respon dengan karakteristik sebagai berikut. Ukuran data ordinal, terdiri dari tiga kategori atau lebih [16]. Model regresi logistik termasuk dalam Model Linear Umum. Model regresi logistik juga bisa disebut model logit. Model logit digunakan untuk memodelkan hub-ungan antar variabel respon. Variabel kategorikal atau kontinu dan variabel prediktor. Jika variabel respon terdiri dari dua kategori disebut dikotomi atau model regresi l*ogistic* biner, tapi bila variabel respon dibagi menjadi lebih dari dua kategori maka dise-but model regresi logistik berganda, dan jika terdapat suatu tingkatan pada kategori tersebut maka disebut model regresi logistik ordinal [17].

3. *Random Forest*

Random forests adalah suatu metode klasifikasi yang terdiri dari gabungan pohon klasifikasi (CART) yang saling independen. Prediksi klasifikasi diperoleh melalui proses voting (jumlah terbanyak) dari pohon-pohon klasifikasi yang terbentuk.*Random forests* merupakan pengembangan dari metode *ensemble* yang pertama kali dikembangkan oleh Leo Breiman (2001) dan digunakan untuk meningkatkan ketepatan klasifikasi. Analisis dengan menggunakan metode random forests dimulai dari pengambilan data dengan teknik resampling bootstrap. Bootstrap adalah suatu metode yang dapat bekerja tanpa membutuhkan asumsi distribusi karena sampel asli digunakan sebagai populasi. Dalam *random forests* proses pengacakan untuk membentuk pohon klasifikasi tidak hanya dilakukan untuk data sampel saja melainkan juga pada pengambilan variabel prediktor. Sehingga, proses ini akan menghasilkan kumpulan pohon klasifikasi dengan ukuran dan bentuk yang berbeda-beda. Hasil yang diharapkan adalah suatu kumpulan pohon klasifikasi yang memiliki korelasi kecil antar pohon. Korelasi yang kecil akan menurunkan hasil kesalahan prediksi *Random Forests* [18].

4. *K-Nearest Neighbor*

K-*Nearest Neighbor* (KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap suatu objek berdasarkan data pembelajaran yang memiliki jarak paling dekat dengan objek tersebut. Data pembelajaran akan diproyeksikan ke dalam K ruang berdimensi banyak, yang masing-masing dimensi merepresentasikan fitur dari data. Ruang ini akan dibagi-bagi menjadi suatu bagian berdasarkan klasifikasi yang telah dilakukan terhadap data pembelajaran. [19] Langkah-langkah yang dilakukan untuk melakukan metode K-Nearest Neighbor ada sebagai berikut:

1. *Generate* data sampel yang akan digunakan sebagai data *training*
2. Inisialisasi K titik sebagai titik-titik pusat (*centroids*) awal
3. Hitung jarak setiap objek data set dengan data training menggunakan perhitungan *Euclidean distance*

Rumus untuk menghitung jarak antar dua titik (x_1, y_1) sebagai titik data set dengan (x_2, y_2) sebagai titik data *training* ditunjukkan pada persamaan berikut.

$$dis(x,y) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

E. Evaluasi Model dan Seleksi

Sebelum membangun model klasifikasi, ada banyak proses untuk mendapatkan hasil terbaik dalam model klasifikasi. Bila kita ingin memperkirakan seberapa akurat m model, terdapat dua jenis metode yang merupakan ukuran untuk evaluasi kinerja clas-sifier, metode *repeated holdout* dan *k-fold cross* validasi. Pertama, matriks konfusi adalah salah satu yang umum pendekatan untuk men-gukur kinerja untuk klasifikasi model. Hal ini dapat didefinisikan metrik untuk pengklasifikasi evaluasi kinerja. Dalam matriks konfusi, kedua kelas tersebut adalah diidentifikasi sebagai kelas positif (+1) dan kelas negatif (−1). Terdapat 4 jenis yakni *True Positive, False Positive, True Negative, False Negative*. Sebagai ditunjukkan pada Tabel 1, setiap kelas prediksi dibandingkan dengan kelasnya kelas aktual untuk setiap contoh untuk menghitung empat ukuran.

Tabel 1. Matriks Konfusi

Kelas Aktual	Kelas Prediksi Positif	Kelas Prediksi Negatif
Positif	*True Positive* (TP)	*False Negative* (FN)
Negatif	*False Positive* (FP)	*True Negative* (TN)

III. METODE PENELITIAN

A. Sumber Data

Data dari artikel "*Hotel Booking Demand Datasets*" ditulis oleh Nuno Antonio, Ana Almeida, dan Luis Nunes

untuk Data in Brief, Volume 22, February 2019. Berisi informasi pemesanan untuk hotel kota dan hotel resor, dan mencakup informasi yang mendukung dimana dapat dilihat pada subbab selanjutnya.

B. *Variabel Penelitian*

Variabel yang digunakan pada penelitian ini disajikan pada Tabel 2 sebagai berikut .

Tabel 2 Variabel Penelitian

Variabel	Tipe Data	Keterangan	Variabel	Tipe Data	Keterangan
Cancel	Kategorik	Pembatalan Reservasi Hotel 0 = no cancel 1 = cancel	**Previous Cancellations**	Numerik	-
Hari Kerja	Numerik		**previous_bookings not_canceled**	Numerik	-
Tahun Kedatangan	Kategori	{2015,2016,2017}	**Tipe Kamar Pesanan**	Kategori	{A,B,C,D,E,F,G,H,L,P}
Bulan Kedatangan	Kategori	{1,2,3,4,5,6,7,8,9,10,11,12}	**Tipe Kamar Ditempati**	Kategori	{A,B,C,D,E,F,G,H,I,K,L,P}
Dewasa	Numerik	-	**Perubahan pemesanan**	Numerik	-
Anak-anak	Numerik	-	**Tipe Deposit**	Kategori	{No Deposit, Non Refund, Refundable}
Makanan	Kategori	{BB, FB,No Meal}	**Tipe Customer**	Kategori	{Transient, Transient-party, Group, Contract}
Negara	Kategori	Kode Negara	**Adr**	Numerik	-
Hotel	Kategori	0 = Resort Hotel 1 = City Hotel			

C. *Langkah Penelitian*

Langkah - langkah pada penelitian ini adalah sebagai berikut.

1. Mengidentifikasi masalah dan tujuan penelitian
2. Mengumpulkan data dari website k*aggle.com*
3. Melakukan *preprocessing* data
4. *Summary statistics* dan membuat visualisasi data
5. Melakukan *Feature selection* dengan ANOVA dan Chi-Square
6. Menganalisis klasifikasi variabel respon ‘cancel’ menggunakan metode *Naive Bayes, Logistic Regression, Random Forest dan K-Nearest Neighbor (KNN)*
7. Membandingkan model menggunakan kriteria akurasi, sensitifitas, spesifisitas, ROC dan AUC
8. Menarik kesimpulan dan saran

IV. HASIL DAN PEMBAHASAN

D. *Data Preprocessing*

Tahap *preprocessing* diperlukan untuk membersihkan data dari yang tidak diperlukan, sehingga metode Klasifikasi akan lebih optimal dalam perhitungannya. Berikut hasil deteksi *missing values* semua variabel menggunakan Python.

Tabel 3 Jumlah *Missing Values*

No	Variabel	Missing Value
1	Cancel	0
2	Hari Kerja	0
3	Tahun Kedatangan	0
4	Bulan Kedatangan	0
5	Dewasa	0
6	Anak-anak	4
7	Makanan	0
8	Negara	488
9	Pembatalan Sebelumnya	0
10	Booking sebelumnya yang tidak dicancel	0
11	Tipe Kamar Pesanan	0
12	Tipe Kamar Ditempati	0
13	Perubahan pemesanan	0
14	Tipe Deposit	0
15	Tipe Customer	0
16	Adr	0

17	Jenis Hotel	0

Berdasarkan Tabel 3, dapat diketahui bahwa terdapat *missing value* pada variabel Anak-Anak dan variabel Negara. Pada variabel anak-anak terdapat 0,335% missing value, sedangkan pada variabel Negara terdapat 40,87% missing value. Variabel anak-anak merupakan variabel numerik sehingga *missing value* pada variabel tersebut dapat diatasi dengan melakukan imputasi nilai mean variabel Anak-anak. Sedangkan pada variabel negara yang bertipe kategori, *missing value* diimputasi dengan menggunakan nilai modus. Setelah dilakukan imputasi, dilanjutkan dengan deteksi *outlier*. Deteksi *outlier* dilakukan dengan menggunakan nilai *z-score* dimana data dengan nilai z-*score* < 3 akan dilakukan penghapusan, sehingga akan diperoleh data bersih tanpa adanya *outlier.*

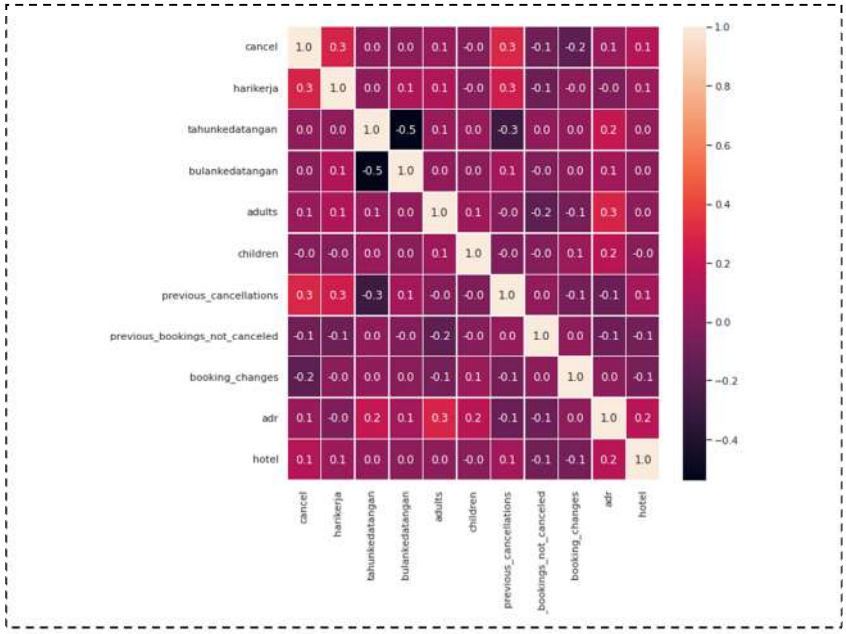
E. Visualisasi Data

Setelah dilakukan *preprocessing* data, selanjutnya dilakukan analisis statistika deskriptif dan visualisasi data dengan menggunakan berbagai plot serta mendapatkan *summary data.* Berikut merupakan hasil pengolahan data dengan menggunakan statistika deskriptif.

Tabel 4 Statistika Deskriptif Variabel Numerik

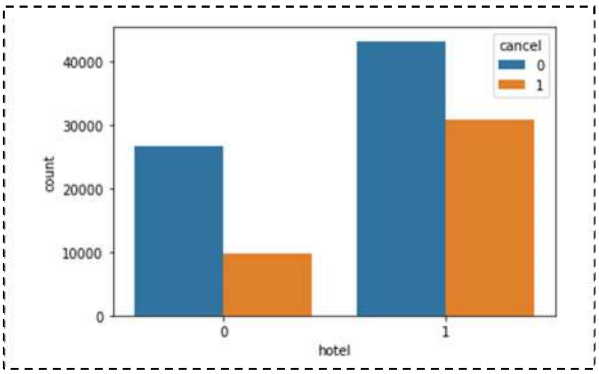
No	Variabel	*Mean*	Standar Deviasi
1	Hari Kerja	100,296	98,779
2	Tahun Kedatangan	2016	0,71
3	Bulan Kedatangan	7	3,1
4	Dewasa	1,858	0,476
5	Anak-anak	0,0405	0,197
6	Pembatalan Sebelumnya	0,0525	0,225
7	Booking sebelumnya yang tidak dicancel	0,0395	0,297
8	Perubahan pemesanan	0,1698	0,452
9	Adr	98,918	42,73

Tabel 4 menunjukkan karakteristik tentang variabel data dengan tipe data numerik pada data, bahwa dari data tersebut menunjukkan rata-rata orang datang pada bulan ke 7, jumlah dewasa sebanyak 2 orang. Sehingga dapat dilanjukan ke analisis berikutnya yakni visualisasi data.



Gambar 1 Plot Korelasi antar Variabel

Berdasarkan Gambar 1 diatas dapat diamati korelasi antar variabel dari data permintaan pemesanan hotel. Semakin mendekati nilai 1 maka hubungan atau korelasi antar variabel semakin kuat. Korelasi terbesar yakni antara variabel cancel dengan tipe deposit yaitu sebesar 0.46 sehingga pembatalan hotel sangat berhubungan dengan tipe deposit yang dipilih oleh pelanggan.

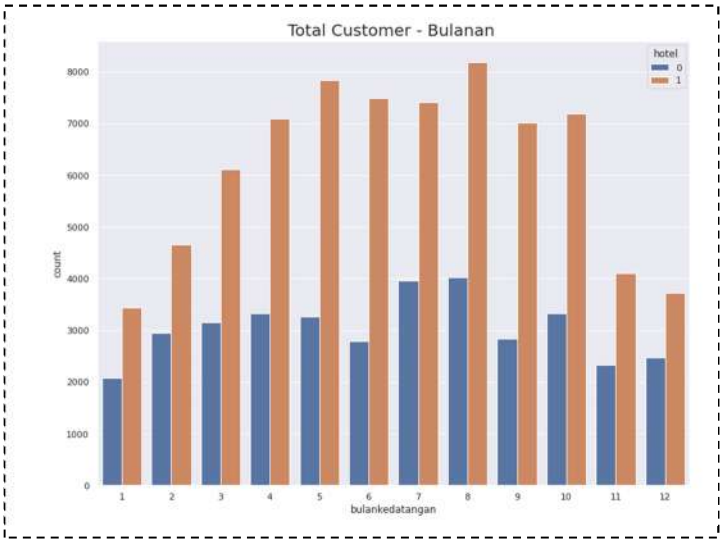


Gambar 2 Jumlah Pembatalan berdasarkan Jenis Hotekl

Gambar 2 menunjukkan bahwa pembatalan hotel lebih banyak terjadi pada city hotel seperti yang terlihat pada Gambar 2. Meskipun demikian, jumlah pelanggan di city hotel lebih banyak dibandingkan dengan *resort hotel.*

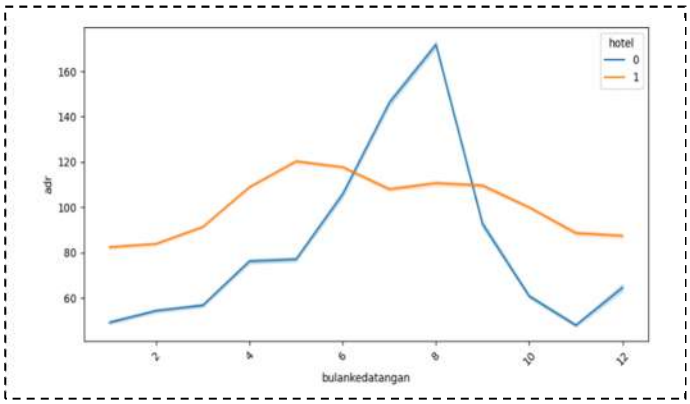
Gambar 3 Jumlah Kedatangan berdasarkan Bulan

Berdasarkan Gambar 3 dapat diketahui bahwa jumlah kedatangan pada bulan Agustus merupakan paling banyak dalam setahun dimana pada bulan itu merupakan awal musim gugur. Dan paling sedikit yaitu pada bulan Januari dengan estimasi pada musim dingin sehingga sedikit pula yang melakukan kedatangan di hotel karena cuaca yang tidak mendukung untuk bepergian ke luar rumah.



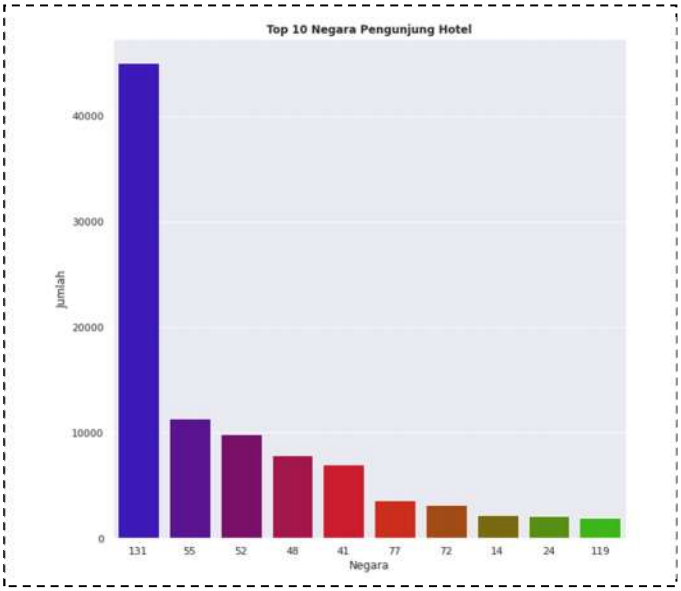
Gambar 4 Total Customer berdasarkan Jenis Hotel setiap Bulan

Berdasarkan Gambar 4, jika diperhatikan, terlihat bahwa jumlah *cancellation* di setiap bulannya mencapai hampir 50% dari total jumlah customer yang datang di setiap bulan.



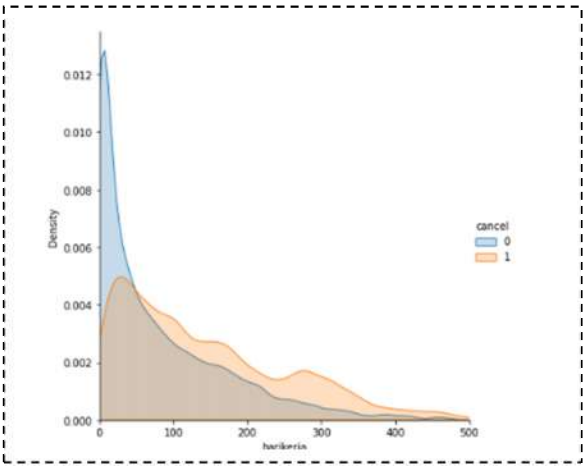
Gambar 5 Jumlah Kedatangan berdasarkan Bulan

Gambar 5 menunjukkan bahwa tipe hotel resort lebih banyak dipilih pada bulan Juli dan Agustus dimana saat itu merupakan hari libur musim panas sehingga banyak keluar yang mengunjungi hotel resort untuk berlibur. City hotel memiliki jumlah pelanggan yang relative stabil dalam semua bulan dalam setahun dilihat berdasarkan Gambar 5 yaitu garis kuning yang tidak terlihat berfluktuasi.



Gambar 6 Top 10 Negara Pengunjung Hotel

Gambar 6 menunjukkan bahwa pengunjung hotel mayoritas berasal dari Benua Eropa seperti Portugal (PRT) sebanyak 45.000 pelanggan lalu pengunjung dari negara Great Britania Raya (GBR), Prancis dan Spanyol.



Gambar 7 Hari Kerja Berdasarkan Status Pembatalan Hotel

Berdasarkan Gambar 7, dapat diketahui bahwa pada waktu tunggu (hari kerja hotel) lebih dari 60 hari, pelanggan cenderung membatalkan reservasi mereka dengan kata lain tingkat pembatalan lebih tinggi jika hari kerja pada pemrosesan reservasi lebih dari 60 hari.

F. *Feature Selection*

Feature Selection digunakan untuk mengidentifikasi faktor yang paling penting atau yang paling penting dan mempengaruhi variabel respon. Dalam penugasan ini, teknik pemilihan fitur digunakan untuk memprediksi faktor paling penting yang mempengaruhi apakah pelanggan memilih untuk membatalkan pemesanan hotel mereka atau tidak. Sehingga variabel yang dipilih sebagai target adalah variabel cancel yakni untuk mengetahui apakah faktor yang utama dipilih pelanggan sehingga mereka memilih untuk membatalkan pemesanan. Pada data ini dilakukan *feature selection* dengan metode *ANOVA with f_classif* untuk variabel numerik dan *Chi-Square* untuk variabel kategorik. Dengan metode *ANOVA f_classif* diperoleh *score ANOVA* tertinggi pada variabel berikut yakni variabel dewasa, hari_kerja, anak-anak dan adr. Lalu pada nilai Chi-Square didapatkan variabel kategorik terbaik adalah negara, tipe deposit, assigned_room_type, previous cancellations, previous_bookings_not_canceled, tipe kamar yang dipesan, hotel, tipe konsumen, makanan, bulan kedatangan dan tahun kedatangan,

G. *Analisis Klasifikasi*

Selanjutnya akan dilakukan klasifikasi terhadap data menggunakan beberapa metode yakni 4 metode *Naïve Bayes*, *Logistic Regression*, *Random Forest* dan *K-Nearest Neighbor* (KNN) dengan *Repeated Holdout* dan *K-Fold* (k=5). Sehingga didapatkan hasil sebagai berikut yang tertera pada berikutnya.

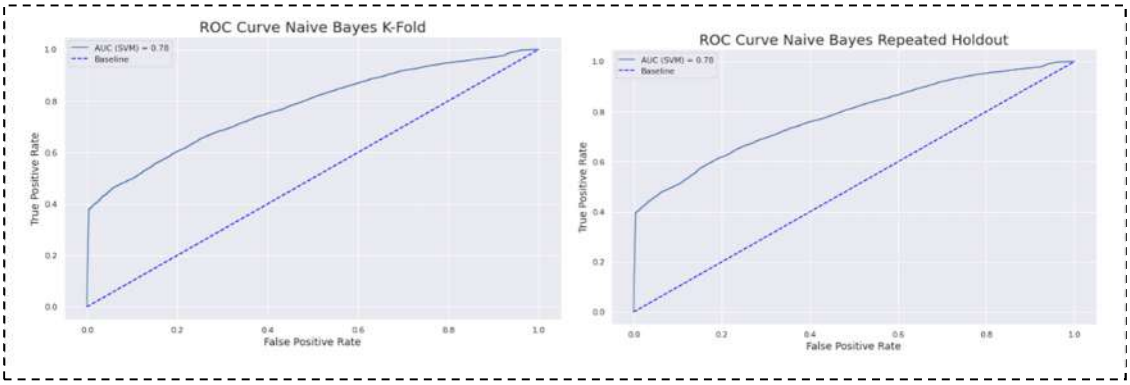
1. *Naïve Bayes*

   Data yang telah bersih kemudian diklasifikasikan menggunakan *Naïve Bayes* dengan *Repeated Holdout* dan *K-Fold Cross Validation* (k=5) didapatkan *confusion matrix* sebagai berikut.

Tabel 5 Perbandingan Nilai Kritera Model *Naïve Bayes*

Metode	Akurasi	Sensitifitas	Spesitifitas	AUC
K-Fold CV (k=5)	0.7698	0.9836	0.4053	0.79
Repeated Holdout	0.7690	0.9843	0.4016	0.78

Perbandingan Kurva ROC antara *Repeated Holdout* dan *K-Fold* sebagai berikut pada Gambar 8.



Gambar 8 Kurva ROC Model *Naïve Bayes*

Nilai AUC dari metode *Naïve Bayes* dengan *Repeated Holdout* dan *K-Fold* bernilai 0.78 dan 0.79 yang berarti metode *Naïve Bayes* cukup baik dalam mengklasifikasikan data.

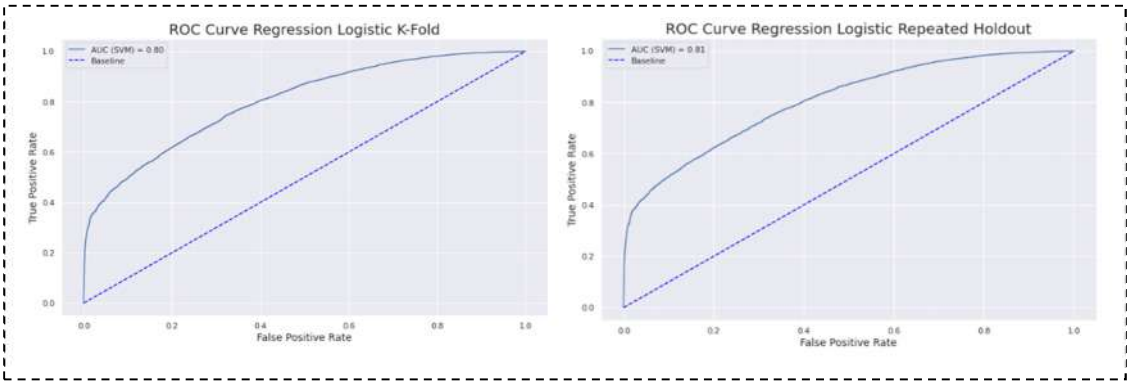
2. *Logistic Regression*

Data yang telah bersih kemudian diklasifikasikan menggunakan *Logistic Regression* dengan *Repeated Holdout* dan *K-Fold Cross Validation* (k=5) didapatkan *confusion matrix* sebagai berikut.

Tabel 6 Perbandingan Nilai Kritera Model *Logistic Regression*

Metode	Akurasi	Sensitifitas	Spesitifitas	AUC
K-Fold CV (k=5)	0.7518	0.8848	0.550	0.80
Repeated Holdout	0.7505	0.8872	0.5172	0.81

Perbandingan Kurva ROC antara *Repeated Holdout* dan *K-Fold* sebagai berikut pada Gambar 9.



Gambar 9 Kurva ROC Model *Logistic Regression*

Nilai AUC dari metode *Logistic Regression* dengan *Repeated Holdout* dan *K-Fold* bernilai 0.80 yang berarti metode *Logistic Regression* baik dalam mengklasifikasikan data.

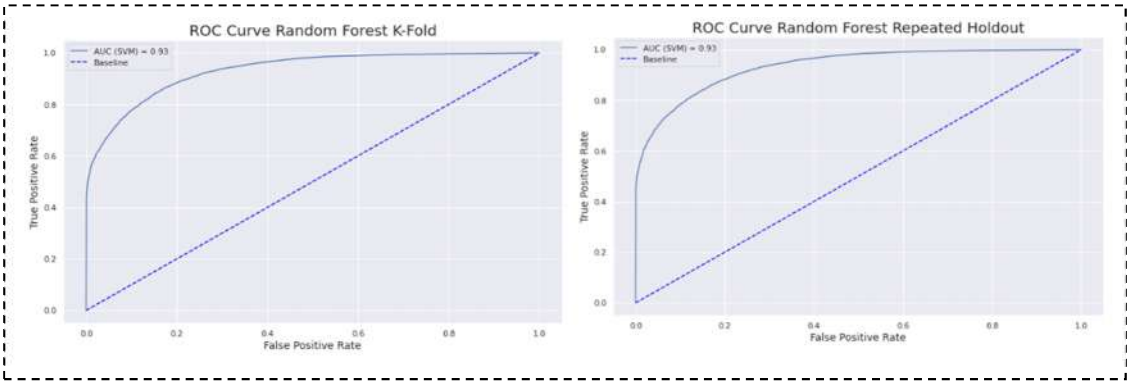
3. *Random Forest*

Data yang telah bersih kemudian diklasifikasikan menggunakan *Random Forest* dengan *Repeated Holdout* dan *K-Fold Cross Validation* (k=5) didapatkan *confusion matrix* sebagai berikut.

Tabel 7 Perbandingan Nilai Kritera Model *Random Forest*

Metode	Akurasi	Sensitifitas	Spesitifitas	AUC
K-Fold CV (k=5)	0.8593	0.9118	0.7699	0.93
Repeated Holdout	0.8554	0.9115	0.7596	0.93

Perbandingan Kurva ROC antara *Repeated Holdout* dan *K-Fold* sebagai berikut pada Gambar 10.



Gambar 10 Kurva ROC Model *Random Forest*

Nilai AUC dari metode *Random Forest* dengan *Repeated Holdout* dan *K-Fold* bernilai 0.93 yang berarti metode *Random Forest* sangat baik dalam mengklasifikasikan data.

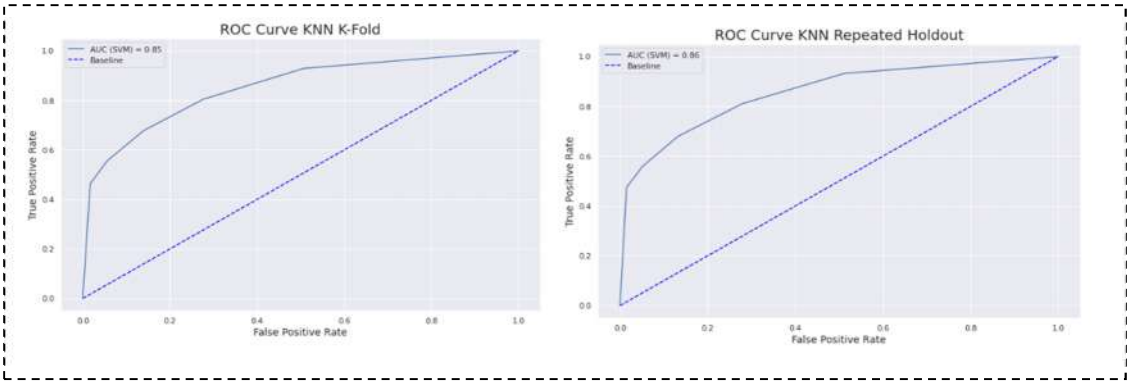
4. *K-Nearest Neighbor* (KNN)

Data yang telah bersih kemudian diklasifikasikan menggunakan *K-Nearest Neighbor* dengan *Repeated Holdout* dan *K-Fold Cross Validation* (k=5) didapatkan *confusion matrix* sebagai berikut.

Tabel 8 Perbandingan Nilai Kritera Model KNN

Metode	Akurasi	Sensitifitas	Spesitifitas	AUC
K-Fold CV (k=5)	0.7964	0.8669	0.6761	0.85
Repeated Holdout	0.7952	0.8600	0.6697	0.86

Perbandingan Kurva ROC antara *Repeated Holdout* dan *K-Fold* sebagai berikut pada Gambar 11.



Gambar 11 Kurva ROC Model KNN

Nilai AUC dari metode KNN dengan *Repeated Holdout* dan *K-Fold* bernilai 0.85 dan 0.86 yang berarti metode KNN baik dalam mengklasifikasikan data.

H. Perbandingan Model

Berikut merupakan tabel hasil perbandingan 4 model menggunakan kriteria akurasi, sensitifitas, spesifisitas, dan AUC.

Table 9 Perbandingan Kriteria 4 Model Klasifikasi

Model	*K-Fold (k=5)*				*Repeated Holdout*			
	Akurasi	Sensitifitas	Spesitifitas	AUC	Akurasi	Sensitifitas	Spesitifitas	AUC

Naïve Bayes	0.7698	0.9836	0.4053	0.79	0.7690	0.9843	0.4016	0.78
Logistic Regression	0.7518	0.8848	0.550	0.80	0.7505	0.8872	0.5172	0.81
Random Forest	**0.8593**	**0.9118**	**0.7699**	**0.93**	0.8554	0.9115	0.7596	0.93
K-Nearest Neighbor (KNN)	0.7964	0.8669	0.6761	0.85	0.7952	0.8600	0.6697	0.86

Hasil analisis klasifikasi dengan berbagai model dan menggunakan metode *Repeated Holdout* dan *K-Fold Cross Validation* di atas menunjukkan bahwa model terbaik yaitu *Random Forest* dengan *K-Fold Cross Validation* yaitu nilai AUC terbesar yaitu 0.93 dengan nilai akurasi sebesar 85.9%, nilai sensitifitas sebesar 91.12% dan nilai spesitifitas sebesar 76.99%.

V. KESIMPULAN

Variabel yang digunakan berdasarkan hasil feature selection dengan ANOVA dan Chi-Square adalah variabel dewasa, hari_kerja, anak-anak, adr, negara, tipe deposit, assigned_room_type, previous cancellations, previous_bookings_not_canceled, tipe kamar yang dipesan, hotel, tipe konsumen, makanan, bulan kedatangan dan tahun kedatangan,

Model terbaik untuk mengklasifikasikan data pembatalan reservasi hotel yaitu *Random Forest* dengan *K-Fold Cross Validation* yaitu nilai AUC sebesar 0.93 dengan akurasi sebesar 85.9%, sensitifitas sebesar 91.12% dan spesitifitas sebesar 76.99%.

REFERENCES

[1] S. S. Wachyuni and K. Wiweka,, "Kepuasan Wisatawan Dalam Penggunaan E-Commerce Agoda", vol. 8, no. No 1, 2020.

[2] A. J. Sanchez-Medina and E. C.-. Sanchez, "Using machine learning and big data for efficient forecasting of hotel", International Journal of Hospitality Management, no. 89, 2020.

[3] World Tourism Organization, NWTO Tourism Highlights, 2018 Edition ed., Madrid: UNWTO , 2018.

[4] R. A. Bahtiar and J. P. Saragih, "Dampak Covid-19 terhadap perlambatan ekonomi sektor umkm," Jurnal Bidang Ekonomi Dan Kebijakan Publik,, vol. 7, no. 6, pp. 19-24, 2020.

[5] R. E. Walpole, Pengantar Statistika edisi ke -3, Jakarta: PT. Gramedia Pustaka Umum, 1995.

[6] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques Third Edition, 225 Wyman Street, Waltham, MA 02451, USA: Elsevier Inc, 2012.

[7] J. Kaiser, "Dealing with Missing Values in Data," *Jurnal Sistem Integrasi,* vol. 5, no. 1, pp. 42-51, 2014.

[8] S. a. Willems, "Multivariate Regression S-Estimators for Robust Estimation and Inference," *Statistica SInica,* pp. 981-1001, 2005.

[9] B. Fry, Visualizing Data: Exploring and Explaining Data with the Processing Environment, Sebastopol: o'Reilly, 2007.

[10] D. C. Montgomery, Introduction to Statistical Quality Control Sixth Edition, New Jersey: John Wiley & Sons Inc, 2013.

[11] Besterfield, Quality Control 8th Edition, Jakarta: PT Gramedia Pustaka Utama, 2006.

[12] G. K. Battacharya and A. R. Johnson, Statistics Principles and Methods 6th Edition, United State of America: John Wiley & Sons Inc., 2010.

[13] A. S. Fitrani, I. R. I. Astutik and M. A. Rosid, "Analysis of classification algorithm in pension types," Journal of Physics, 2019.

[14] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression, New York: Second Edition, John Willey & Sons, 2000.

[15] M. A. S. Cahyanto, "Analysis of Students' Misconception Based on the Use of Learning Objectives in Classification of Materials and Their Properties," Journal of Physics, 2019.

[16] Breiman, L. (2001). Random Forests. Machine Learning. 45(1).5-32.

[17] Breiman, L., Friedman, J. H., Olshen, R. A.,dan Stone, C. J.. (1993). Classification and Regression Trees. New York : Chapman Hall.

[18] "k-nearest neighbors algorithm - Wikipedia." [Daring]. Tersedia pada:https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm. [Diakses:06-Juni-2022].

[19] "Research of Spatial Data Query Optimization Methods Based on KNearest Neighbor Algorithm," ResearchGate. [Daring]. Tersedia pada:https://www.researchgate.net/publication/285618786_Research_of_Spatial_Data_Query_Optimization_Methods_Based_on_KNearest_Neighbor_Algorithm. [Diakses: 06-Jun-2022].