Prediksi Pembatalan Pemesanan Hotel Menggunakan Algoritma Decision Tree, KNN, dan Random Forest

Given Liuwandy[1], Steven[2], William[3]
Program Studi Sistem Informasi Fakultas Teknik dan Informatika, Universitas Multimedia Nusantara, Tangerang, Indonesia
given.liuwandy@student.umn.ac.id
steven14@student.umn.ac.id
william13@student.umn.ac.id

***Abstrak*— Hotel merupakan sebuah bangunan yang berisikan banyak kamar dan fasilitas untuk disewakan kepada masyarakat atau wisatawan untuk bermalam. Dengan jumlah wisatawan yang terus meningkat maka jumlah hotel juga semakin meningkat karena tingginya permintaan. Kemudahan untuk melakukan reservasi hotel secara online menjadi salah satu keuntungan baik bagi calon pengunjung hotel maupun pemilik hotel. Tingginya angka pengunjung hotel juga berbanding lurus dengan angka pembatalan reservasi yang semakin meningkat. Penelitian ini akan melihat faktor-faktor yang menyebabkan pembatalan booking hotel dengan melakukan prediksi menggunakan KNN, Decision Tree, dan Random Forest.**

Kata Kunci* — *Decision Tree, Hotel, KNN , Random Forest

I. LATAR BELAKANG

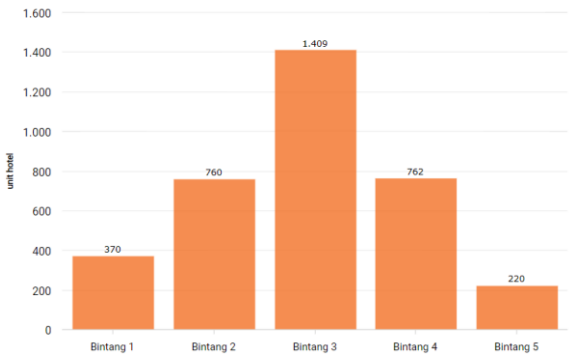
Hotel adalah sebuah bangunan yang dimiliki sekaligus dikelola secara komersial dengan cara memberikan fasilitas baik pelayanan maupun akomodasi. Selain itu, hotel juga dapat diartikan sebagai sebuah bangunan yang memiliki banyak kamar dan disewakan sebagai tempat menginap dan tempat makan bagi orang yang sedang melakukan perjalanan maupun pariwisata. Hotel menjadi salah satu penyedia akomodasi yang dianggap aman dan nyaman bagi orang yang sedang berwisata ke suatu daerah dan menginginkan tempat bermalam yang nyaman [1].

Hotel sendiri dinilai berdasarkan bintang dari hotel tersebut semakin tinggi bintang suatu hotel maka semakin banyak dan lengkap juga fasilitas yang dimiliki oleh hotel tersebut namun kualitas yang didapatkan juga akan berbanding lurus dengan harga yang harus dibayarkan karena semakin tinggi bintang suatu hotel akan semakin mahal juga harga hotel tersebut.

Seiring dengan perkembangan teknologi hampir semua sektor dalam kehidupan manusia mulai diambil alih atau ditambahkan peranan teknologi di dalamnya. Teknologi internet menjadi salah satu teknologi yang kini tidak bisa dilepaskan dari kehidupan manusia dalam menjalankan kehidupan sehari-hari. Teknologi internet tidak hanya memudahkan untuk berkomunikasi dengan orang lain dengan jarak yang jauh tetapi juga menjadi sarana untuk melakukan pemesanan barang, jasa atau layanansecara online mulai dari makanan, pembelian barang secara online dan juga melakukan pemesanan tiket maupun hotel secara online [2].

Berbagai aplikasi kini mendukung masyarakat dalam melakukan pemesanan secara online mulai dari Traveloka, Agoda, RedDoorz, Booking.com, dan masih banyak lagi. Berbaagai aplikasi pemesanan hotel secara online ini tidak hanya memudahkan untuk melakukan reservasi tanpa perlu secara langsung datang ke hotel atau menghubungi pihak hotel tetapi masih banyak keuntungan lain seperti promo yang ditawarkan seperti potongan harga, cashback dan lain sebagainya.

Menurut data Badan Pusat Statistik atau yang biasa disingkat BPS hingga tahun 2021 jumlah hotel yang berklasifikasi bintang di Indonesia mencapai angka 3.521 hotel, dan hotel berbintang 3 menjadi hotel dengan jumlah terbanyak yakni mencapai angka 1.409 hotel dan hotel berbintang 5 menjadi hotel paling sedikit dengan jumlah 220 hotel yang tersebar di Indonesia.



Gambar 1. 1 Jumlah Hotel Di Indonesia

Sangat disayangkan saat adanya pandemik Covid-19 jumlah pengunjung hotel semakin berkurang dimana hal ini terjadi karena adanya lockdown yang ditetapkan oleh pemerintah sehingga jumlah wisatawan semakin berkurang, dimana hal ini sangat berpengaruh terhadap bisnis hotel yang ada di Indonesia [3].

Kemudahan untuk melakukan reservasi hotel di masa ini menjadi hal yang menguntungkan baik bagi

pemesan hotel maupun bagi pemilik hotel. Namun, kemudahan pemesanan ini juga berbanding lurus dengan angka pembatalan booking hotel. Dimana pembatalan reservasi hotel ini bisa sangat merugikan bagi pihak pengelola hotel. Permasalahan pembatalan booking hotel ini menjadi permasalahan pada penelitian ini sehinggapeneliti menggunakan metode prediksi menggunakan tiga algoritma yakni KNN, Random Forest, dan Decision Tree untuk menentukan faktor apa saja yang mempengaruhi orang yang ingin melakukan pembatalan booking hotel.

II. KAJIAN TEORI

A. *Data Mining*

Data Mining telah menjadi seperangkat alat yang populer untuk menangani Big Data (Data dalam jumlah besar). Untuk mengelola, dan menganalisis data untuk pengambilan keputusan terutama dari jenis data dalam jumlah besar, digunakan teknik yang disebut data mining. Data mining akan mengeksplorasi dataset untuk menemukan pengetahuan yang tersembunyi dan tidak diketahui, Data mining juga bisa digunakn untuk melakukan prediksi yang dapat digunakan untuk pengambilan keputusan di masa depan [4]. Data mining adalah ekstraksi pola dan hubungan yang berguna dalam sumber data, seperti: (1) database, (2) teks, (3) web, perhatian dalam data mining adalah data yang bising, data statis, nilai yang hilang, data cadangan, efisiensi algoritma, ukuran dan kompleksitas data.

Data mining juga dikenal sebagai Knowledge Discovery in Databases (KDD). Ada lima elemen utama dalam data mining:

1. Extract, memuat dan juga melakukan transformasi data transaksi ke dalam data warehouse system.
2. Menyimpan dan juga mengatur data dalam multi-dimensional database system.
3. Memberikan akses data kepada business analysts dan IT professionals.
4. Melakukan analisa data menggunakan aplikasi analisis
5. Mempresentasikan data dalam bentuk yang bisa dibaca dan format yang berguna. Seperti table, grafik dan lain-lain.

B. *Decision Tree*

Decision tree merupakan salah satu satu algoritma dalam machine learning yang pertama kali diterapkan pada awal tahun 1966 yang digunakan untuk melakukan klasifikasi didalam data analysis. Decision Tree akan mengubah data menjadi Decision Rules dalam bentuk struktur pohon yang dimana setiap nodenya merepresentasikan atribut sedangkan batang merepresentasikan value dari atribut tersebut. Decision Tree sering juga disebut sebagai CART (Classification and Regression Tree). Dalam pendekatan Decision tree atribut yang diperlukan untuk menganalisis diambil sebagai node awal. Tujuan dari algoritma Decision tree untuk membuat model yang memprediksi nilai dari target variabel, Decision tree menggunakan representasi pohon untuk menyelesaikan masalah di mana node daun adalah label kelas dan atribut direpresentasikan pada node internal dari pohon [5].

Sejak saat itu, CART terus berkembang dan melahirkan beberapa tipe algoritma *Decision Tree* baru, seperrti ID3 (Iterative Dichotomiser 3) yang menggunakan *gain ratio* dan C4.5 yang menduduki peringkat pertama dalam 10 besar algoritma yang digunakan untuk melakukan data mining pada tahun 2008.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Gambar 1. 2 Decision Tree Formula

Keuntungan utama yang didapat jika menggunakan algoritma Decision tree yakni kemampuannya untuk melakukan break down terhadap data dan membuat tampilannya menjadi lebih sederhana. Algoritma ini juga berguna untuk menemukan hubungan yang tercipta antara beberapa variabel dalam dataset, seperti variabel input dan variabel target dalam perbandingan. Decision Tree menjelaskan opsi, resiko, tujuan serta manfaat dari setiap opsi yang ditawarkan.

C. *K-Nearest Neighbor (KNN)*

Algoritma *K-Nearest Neighbor* adalah salah satu metode algoritma analisis data yang mengklasifikasikan sekumpulan data berdasarkan pembelajaran data yang telah diklasifikasikan sebelumnya, seperti pembelajaran terawasi dimana hasil dari query instance diklasifikasikan berdasarkan mayoritas kedekatan dalam kategori KNN. Algoritma K-Nearest Neighbors digunakan untuk mengklasifikasikan objek berdasarkan data training yang paling dekat dengan objek dengan tujuan untuk mengklasifikasikan objek baru berdasarkan sampel data training. "Konsep Tetangga" dalam KNN bekerja berdasarkan asumsi bahwa data akan memiliki kelas atau kategori yang sama dengan data di sekitarnya. Secara sederhana, K-Nearest Neighbors bekerja berdasarkan jarak minimum dari data baru ke data latih dan menentukan K-Nearest Neighbors. Setelah itu, kita mendapatkan nilai mayoritas sebagai hasil prediksi dari data baru tersebut [6].

Algoritma K-Nearest Neighbor sering digunakan dalam aplikasi data mining, pengenalan pola, dan pengolahan citra karena KNN merupakan algoritma supervised learning yang menggunakan data yang sudah ada dan outputnya sudah diketahui [7]. Langkah pertama untuk menghitung metode KNN adalah kita

harus menentukan parameter K, yaitu nilai jumlah tetangga terdekat. Setelah itu, kita harus menghitung jarak Euclidean antara objek dan dataset baru ke semua data training. Dan langkah selanjutnya adalah kita harus mengurutkan hasil langkah sebelumnya secara ascending. Dan kemudian kita harus menentukan tetangga terdekat dari jarak minimum ke K dan menentukan tetangga terdekat berdasarkan objek/data [7].

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Gambar 1. 3 Euclidean Distance Formula

Pada rumus yang ditampilkan dalam gambar 1.3 diatas, kami menggunakan x sebagai atribut dari data training dan y sebagai atribut dari data baru. Kelebihan dari algoritma KNN adalah handal untuk diimplementasikan dan mudah dipelajari terutama dalam bidang geometri. Metode ini dapat berguna dalam dataset yang besar dan tahan terhadap pelatihan data yang *Noist*. Salah satu hal yang paling dirugikan dalam Algoritma K-Nearest Neighbor adalah bias dalam nilai K dan seringkali menjadi tidak relevan

D. *Random Forest*

Random forest merupakan algoritma klasifikasi yang menggunakan pohon flasifikasi dimana setiap dari itu mengaplikasikan bootstrap dari sample data. Setiap variable dalam random forest dipilih secara acak untuk dijadikan sebagai kandidat dari kumpulan variabel dalam pemecahannya. Model ini merupakan model yang memiliki performa paling baik dibandingkan klasifikasilainnya terutama dalam jumlah data yang banyak [8].

Random forest akan mempelajari hubungan antar data pada waktu yang berbeda. Model ini cocok digunakan pada campuran dari data kuantitatif dan kategorikal dan menyediakan variable importance measure. Random forest dicetuskan oleh Leo Breiman dan Adele Cutler. Pada tahun 2000, Leo Breiman yang berasal dari Universitas Berkeley mendefinisikan bahwa pohon keputusan sama dengan kernel dengan margin yang benar. Kemudian pada tahun 2001, ia menerbitkan penelitiannya [9].

$$F - measure = \frac{2 \times recall \times precision}{recall + precision}$$

$$= \frac{2TP}{2TP + FP + FN}$$

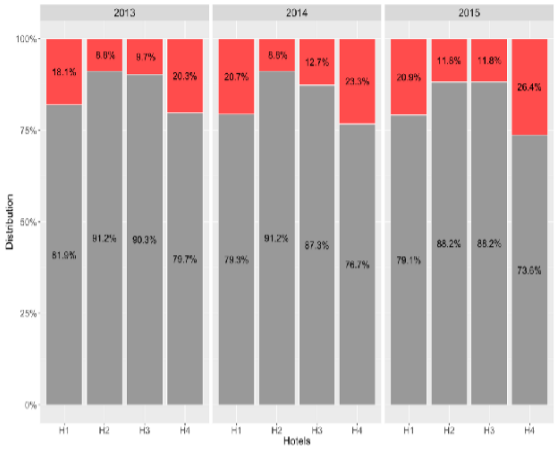
Gambar 1. 4 Random Forest Formula

Pada rumus yang ditampilkan dalam gambar 1.4 diatas menjelaskan *F-Measure* yang didapat dari konstruksi RF Classifier yang menggunakan algoritma *Random Forest* untuk membuat model prediksi. Langkah pertama yang harus dilakukan yakni dengan melakukan konfirmasi dari training set, validation set, dan juga test set, kemudian melakukan ektraksi dengan menggunakan *bootstrap* method dimana Dataset K+1 dimana K digunakan sebagai set training dan sisanya digunakan sebagai K Validasi. Beberapa variabel lainnya yang digunakan dalam rumus tersebut yaitu TP, FP, FN, TN, Precision Rate, serta Recall Values yang didapat dari masing-masing *Classifier*.

E. Penelitian Terdahulu

Penelitian ini menggunakan artikel jurnal yang berjudul *Predicting Hotel Booking Cancellations to Decrease Uncertainty And Increase Revenue (Tourism & Management Studies)* (2017) sebagai penelitian terdahulu yang dijadikan referensi utama dalam menulis jurnal ini. Artikel jurnal tersebut melakukan prediksi terhadap pembatalan reservasi dimana tujuan dari penelitian nya adalah untuk meningkatkan revenue dari hotel dan memangkas ketidak pastian. Penelitian ini melakukan prediksi menggunakan *Two-Class Classification Algorithm*, yakni:

- Boosted Decision Tree
- Decision Forest
- Decision Jungle
- Logically Deep Support Vector Machine (SVM)
- Neural Network



Gambar 1. 5 Rasio Pembatalan Pemesanan Hotel

Hasil dari penelitian yang menggunakan kelima algoritma tersebut mendapatkan nilai yang cukup baik, dimana kelima algoritma tersebut memperoleh hasil nilai AUC diatas 93,5 %, namun model yang paling cocok untuk membuat *Cancellation Prediction Model* adalah algoritma ***Decision Forest*** [10] **.**

III. METODOLOGI

A. *Data Collection*

Penelitian kali ini menggunakan dataset yang didapat dari salah satu website penyedia dataset yang bisa diakses secara gratis, yakni melalui website kaggle.com. Penulis sudah memastikan dataset yang digunakan dalam penelitian kali ini tidak memiliki data NULL dan variabel yang terdapat dalam dataset memiliki hubungan yang relevan dengan topik yang dibahaspersiapan data mencakup proses untuk mengubah dataset menjadi bentuk tabel proporsional yang akan digunakan untuk proses data science selanjutnya, yakni tahap eksplorasi data dan juga data modelling.

B. *Metode Penelitian*

Adapun metode yang digunakan oleh penulis dalam penelitian ini yakni metode kuantitatif yang menjelaskan fenomena apa yang terjadi dari koleksi data numerik dan dianalisa kembali menggunakan metode dasar statistika. Metode penelitian kuantitatif dinilai cocok untuk diterapkan dalam topik penelitian ini karena data yang digunakan merupakan data numerikal dalam bentuk pecahan desimal maupunskala presentase.

C. *Framework (CRISP-DM)*

CRISP – DM (Cross Industry Standar Process – Data Mining) merupakan salah satu metodologi data mining yang dipakai ketika perusahaan ingin mengimplementasikan proses data mining untuk menunjang keperluan industri yang dilakukan oleh data mining expert. CRISP-DM pertama kali di kemukakan oleh 3 perusahaan yang bergerak dalam bidang data-mining pada tahun 1996, yaitu Daimier-Benz, Integral Soluion Ltd (ISL) yang kemudian mengganti nama menjadi SPSS dan perusahaan terakhir yaitu NCR, salah satu perusahaan konsultan spesialis Data Mining.

Gambar 1. 6 CRISP-DM Framework

Berikut merupakan penjelasan mengenai tahapan-tahapan yang dilakukan ketika mengimplementasikan CRISP-DM Framework:

- *Business Understanding* = merupakan langkah pertama dalam proses CRISP-DM yang merujuk pada langkah awal yang dilakukan oleh perusahaan untuk menentukan tujuan serta menentukan permasalahan utama apa yang akan di selesaikan rdasarkan hasil visualisasi yang didapat dari figure 14 yang memvisualisasikan Frekuensi dari CREDIT_SCORE dalam bentuk histogram, dapat disimpulkan.

- *Data Understanding* = Tahap *data understanding* merupakan proses untuk menentukan data apa yang akan digunakan. Tidak hanya mencari data, tetapi *data-mining expert* dalam tahap ini juga memahami kekuatan dan kelemahan dari data yang akan digunakan, apakah data tersebut layak digunakan dan memuat variabel / informasi-informasi penting yang bermanfaat bagi perusahaan atau tidak.Tahap *data understanding* merupakan proses untuk menentukan data apa yang akan digunakan. Tidak hanya mencari data, tetapi *data-mining expert* dalam tahap ini juga memahami kekuatan dan kelemahan dari data yang akan digunakan, apakah data tersebut layak digunakan dan memuat variabel / informasi-informasi penting yang bermanfaat bagi perusahaan atau tidak. Tahapan ini merupakan tahapan yang cukup memakan waktu dan biaya bagi perusahaan dalam proses implementasi CRISP-DM.

- *Data Preparation* = Tahap *data preparation* mencakup proses persiapan data sebelum dilakukan *modelling.* Perbedaan utamaproses *data understanding* dengan *data preparation* yaitu pada proses *Data Preparation* data sudah fix dan siap untuk dilakukan modelling, manipulasi, convert, serta melihat *Missing Value* pada data.

- *Modelling* = Tahap *modelling* tidak hanya memanfaatkan algoritma yang ada, tetapi membuat kombinasi menjadi model prediksi yang lebih baik. Penerapan *modelling* di implementasikan dengan membentuk model-model prediksi dengan bantuan algoritma yang dipilih.

- *Evaluation* = *Evaluation* mencakup evaluasi hasil *modelling* yang dilakukan dalam proses sebelumnya untuk memastikan kembali apakah model tersebut valid sebelum diimplementasikan ke dalam proses

deployment.

- *Deployment* = merupakan proses penerapan algoritma dan model yang sudah ditentukan secara final sebagai machine learning yang bisa digunakan oleh pihak ke 3 untuk kepentingan bisnis.

IV. HASIL DAN ANALISIS

A. *Business Understanding*

Tahap pertama ini merupakan tahapan penting dari data mining. Dikarenakan melalui tahap ini dilakukan pemahaman atas permasalahan yang ingin diselesaikan. Dilakukan pemahaman atas tujuan dan kebutuhan proyek dari perspektif bisnis. Pengetahuan dan pemahaman tersebut diubah dan digunakan dalam mendefinisikan dan menyelesaikan permasalahan proyek tersebut.

Pada industri perhotelan, sering kali ditemukan adanya *cancel booking* atau pembatalan pemesanan hotel. Pembatalan ini berdampak cukup besar bagi hotel, maka dari itu dilakukan prediksi yang memanfaatkan data historis untuk memprediksi pembatalan di masa depan. Prediksi ini dapat membantu pelaku bisnis perhotelan untuk melakukan peningkatan pada sektor bisnisnya. Melalui pemahaman dan pembelajaran faktor yang membuat *customer* melakukan pembatalan.

B. *Data Understanding*

Tahap kedua pada CRISP-DM yang bertujuan dalam pemahaman data yang menjadi acuan solusi dan tujuan bisnis. Pada tahap ini ditentukan data yang tepat untuk dianalisis, serta mengidentifikasi dan memahami kualitas dan limitasi data. Tujuan dari tahap ini adalah mempersiapkan kekurangan yang terjadi, melalukan pendekatan, dan mengidentifikasi sub-kategori.

```
In [2]: df = pd.read_csv('hotel_bookings.csv')

In [3]: df.head()
```

Out[3]:

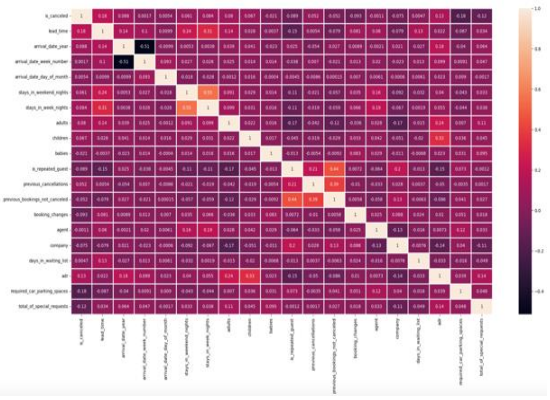
	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_
0	Resort Hotel	0	342	2015	July	27	1	0	
1	Resort Hotel	0	737	2015	July	27	1	0	
2	Resort Hotel	0	7	2015	July	27	1	0	
3	Resort Hotel	0	13	2015	July	27	1	0	
4	Resort Hotel	0	14	2015	July	27	1	0	

Gambar 1. 7 Import dan Membaca Data Head

```
In [4]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  int64
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_week_nights            119390 non-null  int64
 9   adults                          119390 non-null  int64
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  country                         118902 non-null  object
 14  market_segment                  119390 non-null  object
 15  distribution_channel            119390 non-null  object
 16  is_repeated_guest               119390 non-null  int64
 17  previous_cancellations          119390 non-null  int64
 18  previous_bookings_not_canceled  119390 non-null  int64
 19  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64
 26  customer_type                   119390 non-null  object
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  object
dtypes: float64(4), int64(16), object(12)
memory usage: 29.1+ MB
```

Gambar 1. 8 Read Data Info

```
In [11]: plt.figure(figsize = (24, 12))

corr = df.corr()
sns.heatmap(corr, annot = True, linewidths = 1)
plt.show()
```



Gambar 1. 9 Correlation Map

C. *Data Preparation*

Pada tahap ini mencakup seluruh kegiatan preparasi yang dilakukan dalam menyusun *dataset* yang akan diteliti dan dianalisis. Proses persiapan data ini dapat dilakukan berulang kali tanpa urutan tertentu. Dilakukan perancangan data agar sesuai dengan tujuan dan masalah yang terkait dalam *data mining*. Dilakukan pembersihan kekurangan yang sudah ditemukan pada tahap *data understanding* sebelumnya. Seperti data masih memiliki nilai yang hilang ataupun kekurangan lainnya dibetulkan pada tahap ini.

Pada tahap ini memastikan data yang akan digunakan untuk melakukan prediksi benar. Selain membersihkan data, pada tahap ini juga dilakukan metode pembagian data. Data dibagi menjadi 2 tipe

data yaitu *training* (70%) dan *testing* (30%).

```
In [20]: for col in cat_df.columns:
             print(f"{col}: \n{cat_df[col].unique()}\n")

hotel:
['Resort Hotel' 'City Hotel']

meal:
['BB' 'FB' 'HB' 'SC' 'Undefined']

market_segment:
['Direct' 'Corporate' 'Online TA' 'Offline TA/TO' 'Complementary' 'Groups'
 'Undefined' 'Aviation']

distribution_channel:
['Direct' 'Corporate' 'TA/TO' 'Undefined' 'GDS']

reserved_room_type:
['C' 'A' 'D' 'E' 'G' 'F' 'H' 'L' 'B']

deposit_type:
['No Deposit' 'Refundable' 'Non Refund']

customer_type:
['Transient' 'Contract' 'Transient-Party' 'Group']

year:
[2015 2014 2016 2017]

month:
[ 7  5  4  6  3  8  9  1 11 10 12  2]

day:
[ 1  2  3  6 22 23  5  7  8 11 15 16 29 19 18  9 13  4 12 26 17 10 20 14
 30 28 25 21 27 24 31]
```

Gambar 1. 10 Data Preparation

```
In [24]: df.isnull().sum()

Out[24]: hotel                          0
         is_canceled                    0
         lead_time                      0
         arrival_date_year              0
         arrival_date_month             0
         arrival_date_week_number       0
         stays_in_weekend_nights        0
         stays_in_week_nights           0
         adults                         0
         children                       0
         babies                         0
         meal                           0
         market_segment                 0
         distribution_channel           0
         is_repeated_guest              0
         previous_cancellations         0
         previous_bookings_not_canceled 0
         reserved_room_type             0
         deposit_type                   0
         agent                          0
         company                        0
         customer_type                  0
         adr                            0
         required_car_parking_spaces    0
         total_of_special_requests      0
         reservation_status_date        0
         dtype: int64
```

Gambar 1. 11 Checking Null

```
In [31]: y = df['is_canceled']
         X = clean_df
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.30)
         X_train.head()
```

Out[31]:

	hotel	meal	market_segment	distribution_channel	reserved_room_type	deposit_type	customer_type	year	month	day	...	children	babie
111151	1	0	0	0	1	0	0	3	5	7	...	0.0	
9353	0	0	2	2	1	0	0	2	10	24	...	0.0	
73607	1	2	2	2	1	0	0	3	3	11	...	0.0	
49353	1	0	2	2	2	0	2	2	4	14	...	0.0	
12191	0	0	2	2	2	0	0	3	5	31	...	0.0	

5 rows × 26 columns

Gambar 1. 12 Data Splitting (Training dan Testing)

D. *Modelling*

Pada tahap ini akan dilakukan pemilihan dan penerapan model berdasarkan *dataset* yang digunakan. Tahap ini melibatkan penerapan berbagai algoritma machine learning serta metode statistik. Algoritma pada machine learning sendiri terbagi menjadi 3 yaitu *clustering*, *classification*, dan juga regresi. Pada penelitian ini dilakukan *classification* dimana melibatkan 3 algoritma, berikut merupakan hasil akurasi dari ke 3 algoritma yang digunakan:

- K-Nearest Neighbors (KNN)

```
In [32]: knn = KNeighborsClassifier()
         knn.fit(X_train, y_train)

         y_pred_knn = knn.predict(X_test)

         acc_knn = accuracy_score(y_test, y_pred_knn)
         conf = confusion_matrix(y_test, y_pred_knn)
         clf_report = classification_report(y_test, y_pred_knn)

         print(f"Accuracy Score of KNN is : {acc_knn}")
         print(f"Confusion Matrix : \n{conf}")
         sns.heatmap(conf, annot=True)
         print(f"Classification Report : \n{clf_report}")

Accuracy Score of KNN is : 0.8223470518552486
```

Gambar 1. 13 KNN Accuracy

- Decision Tree

```
In [33]: dtc = DecisionTreeClassifier()
         dtc.fit(X_train, y_train)

         y_pred_dtc = dtc.predict(X_test)

         acc_dtc = accuracy_score(y_test, y_pred_dtc)
         conf = confusion_matrix(y_test, y_pred_dtc)
         clf_report = classification_report(y_test, y_pred_dtc)

         print(f"Accuracy Score of Decision Tree is : {acc_dtc}")
         print(f"Confusion Matrix : \n{conf}")
         sns.heatmap(conf, annot=True)
         print(f"Classification Report : \n{clf_report}")

Accuracy Score of Decision Tree is : 0.9133325690702739
```

Gambar 1. 14 Decision Tree Accuracy

- Random Forest

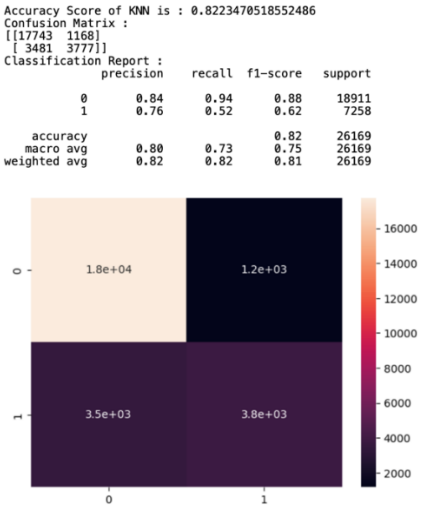
```
In [34]: rd_clf = RandomForestClassifier()
         rd_clf.fit(X_train, y_train)

         y_pred_rd_clf = rd_clf.predict(X_test)

         acc_rd_clf = accuracy_score(y_test, y_pred_rd_clf)
         conf = confusion_matrix(y_test, y_pred_rd_clf)
         clf_report = classification_report(y_test, y_pred_rd_clf)

         print(f"Accuracy Score of Random Forest is : {acc_rd_clf}")
         print(f"Confusion Matrix : \n{conf}")
         sns.heatmap(conf, annot=True)
         print(f"Classification Report : \n{clf_report}")

Accuracy Score of Random Forest is : 0.9204402155221827
```

Gambar 1. 15 Random Forest Accuracy

E. *Evaluation*

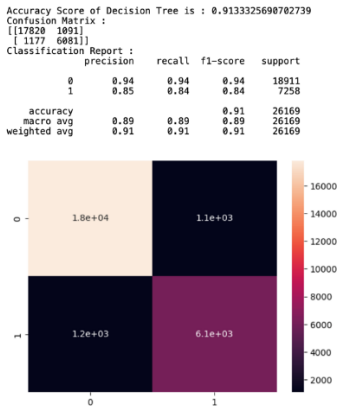
Tahap selanjutnya adalah tahap evaluasi dari model-model yang telah dibuat pada tahap selanjutnya. Dari model-model tersebut dilakukan verifikasi serta pengujian sehingga memastikan model sesuai dengan tujuan dan permasalahan bisnis awal dapat ditangani. Sekaligus pada tahap ini dipilih pemodelan dengan tingkat akurasi, serta penilaian lainnya seperti *misclassification rate* dan lainnya. Pada penelitian ini akan dilihat *confusion matrix* serta nilai *precision*, *recall*, dan akurasi.

- K-Nearest Neighbors (KNN)



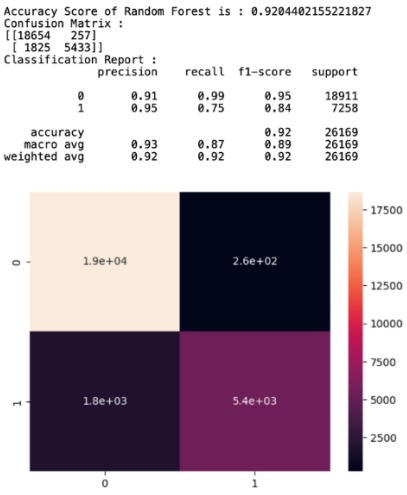
Gambar 1. 16 Evaluating KNN Model

- Decision Tree



Gambar 1. 17 Evaluating Decision Tree Model

- Random Forest



Gambar 1. 8 Evaluating Random Forest Model

F. *Deployment*

Setelah model dievaluasi dan diuji, maka tahap terakhir adalah tahap *deployment* atau penyebaran. Tahap ini diterapkan dengan pembuatan laporan serta mengvisualisasikan hasil pemodelan serta *data mining* lainnya. Tahap ini menerapkan algoritma dan model yang sudah ditentukan secara *final* sebagai machine learning yang bisa digunakan oleh pihak ke 3 untuk kepentingan bisnis perusahaa.

V. KESIMPULAN

Bisnis Hotel menjadi salah satu bisnis yang menguntungkan wisatawan karena bisnis ini menyediakan layanan berupa tempat untuk bermalam atau menikmati makanan dan layanan lainnya. Jumlah cancel booking hotel sangatlah berpengaruh terhadap keuntungan dari pihak hotel, oleh karena itu peneliti memprediksi faktor apa yang menjadi penyebab pembatalan tersebut dan berdasarkan analisa yang dilakukan menggunakan tiga algoritma yakni KNN, Random Forest dan Decision Tree mendapatkan hasil bahwa prediksi ini paling tepat untuk dilakukan menggunakan algoritma Decision Tree.

DAFTAR PUSTAKA

[1] J. R. Gamlath, “Hotel Management Graduates ’ Perception Towards Career in Hotel Industry Post-Pandemic,” no. December, pp. 0–10, 2022, doi: 10.55041/IJSREM17078.

[2] Y. Wu and R. Proyrungroj, “Exploration of the Problems in the Management of the Internet Celebrity High- star Hotels in Sanya under the Self-Media Marketing Exploration of the Problems in the Management of the Internet Celebrity High-star Hotels in Sanya under the Self-Media Marketing,” no. June, 2022.

[3] V. L. Pangkey and N. Rahayu, “Jurnal Public Policy The Impact of Hotel Tax Revenue in the COVID-19 Era as an Increase in Regional Original Income in Indonesia,” vol. 4, pp. 0–5, 2022.

[4] S. Bagga and A. Sharma, “Big Data and Its Challenges: A Review,” *Proc. - 4th Int. Conf. Comput. Sci. ICCS 2018*, no. November, pp. 183–187, 2019, doi: 10.1109/ICCS.2018.00037.

[5] Y. Lee and D. Y. Kim, “The decision tree for longer-stay hotel guest: the relationship between hotel booking determinants and geographical distance,” *Int. J. Contemp. Hosp. Manag.*, vol. 33, no. 6, pp. 2264–2282, 2020, doi: 10.1108/IJCHM-06-2020-0594.

[6] Z. Zuriati and N. Qomariyah, “Klasifikasi Penyakit Stroke Menggunakan Algoritma K-Nearest Neighbor (KNN) Classification of

Stroke Using the K-Nearest Neighbor (KNN) Algorithm," vol. 1, no. 1, pp. 1–8, 2023.

[7] Y. Liang, Y. Pan, X. Yuan, W. Jia, and Z. Huang, "Surrogate modeling for long-term and high-resolution prediction of building thermal load with a metric-optimized KNN algorithm," *Energy Built Environ.*, no. June, 2022, doi: 10.1016/j.enbenv.2022.06.008.

[8] X. Gao, J. Wen, and C. Zhang, "An Improved Random Forest Algorithm for Predicting Employee Turnover," *Math. Probl. Eng.*, vol. 2019, 2019, doi: 10.1155/2019/4140707.

[9] A. O. Alqabbany and A. M. Azmi, "Measuring the effectiveness of adaptive random forest for handling concept drift in big data streams," *Entropy*, vol. 23, no. 7, pp. 1–24, 2021, doi: 10.3390/e23070859.

[10] N. Antonio, A. de Almeida, and L. Nunes, "Predicting hotel booking cancellations to decrease uncertainty and increase revenue," *Tour. Manag. Stud.*, vol. 13, no. 2, pp. 25–39, 2017, doi: 10.18089/tms.2017.13203.

Name	Role
Given Liuwandy (00000042996)	Abstrak, Chapter I (Latar Belakang, Rumusan Masalah) Chapter II (Tinjauan Teoritis)
Steven (00000043310)	Chapter IV (Hasil dan Analisis), Chapter V (Kesimpulan)
William (00000042609)	Chapter II (Tinjauan Teoritis), Chapter III (Metodologi yang digunakan, meliputi setiap fase dalam CRISP-DM)