# Analisis *Mining Data* pada *Skytrax Airline Reviews Dataset*

Nadya Yuniar Desi Prameswari[1], Muhammad Zamroni Al Fikri[2], Santi Wulan Purnami[3], Irhamah[4] dan Kartika Fithriasari[5]
Departemen Statistika, Fakultas Sains dan Analitika Data, Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia
*e-mail*: yuniarnadya4@gmail.com[1], zamronifikri@gmail.com[2], santi_wp@statistika.its.ac.id[3], irhamah@statistika.its.ac.id[4] dan kartika_f@statistika.its.ac.id[5]

***Abstrak*—Skytrax merupakan perusahaan penasihat mutu dan riset spesialis bagi industri transportasi udara yang berdedikasi untuk meningkatkan pengalaman pelanggan bagi maskapai dan bandara di seluruh dunia dengan cara menyediakan wadah untuk penumpang menuliskan ulasan berdasarkan pengalamannya. Pada penelitian ini akan dilakukan analisis mining data pada Skytrax Airlines Review Dataset untuk mengetahui hal apa saja yang membuat penumpang merekomendasikan jasa maskapai penerbangan kepada orang lain berdasarkan pelayanan yang telah diberikan oleh pihak maskapai penerbangan untuk memberikan pelayanan lanjutan yang lebih baik. Fokus utama data mining yang akan digunakan adalah klasifikasi, dimana algoritma yang akan digunakan untuk mengklasifikasikan dataset adalah Naïve Bayes, Random Forest, dan Logistic Regression. Didapatkan hasil bahwa maskapai penerbangan yang paling banyak digunakan ialah Spirit Airlines. Dari variabel seat comfort, food beverage, ground service, entertainment, value for money penumpang paling banyak memberikan penilaian tidak baik, pada variabel cabin service penumpang paling banyak memberikan penilaian sangat baik dan sebanyak 35.401 ulasan penumpang memutuskan untuk tidak merekomendasikan maskapai penerbangan kepada orang lain. Dari ketiga metode klasifikasi, metode logistic regression merupakan metode yang memiliki hasil klasifikasi terbaik dalam kombinasi data training sebesar 70% dan data testing sebesar 30%. Kurva ROC menunjukkan bahwa nilai AUC yang dimiliki sebesar 0,98 yang berarti bahwa model diklasifikasikan sebagai model yang sangat baik.***

***Kata Kunci*— Data Mining, Logistic Regression ,Maskapai Penerbangan, Naïve Bayes, Random Forest**

## I. PENDAHULUAN

Seiring dengan perkembangan zaman yang makin pesat ini, berbagai kemajuan telah terjadi di berbagai sektor, tidak terkecuali sektor transportasi. Alat transportasi di era ini tidak hanya terbatas pada alat transportasi umum seperti bus dan kapal, tetapi juga berkembang alat transportasi pesawat terbang. Berbagai belahan dunia dapat dijangkau secara mudah dan lebih cepat dengan hadirnya jasa indsutri penerbangan komersial.

Pesawat terbang dan penumpang merupakan dua hal yang tidak dapat dipisahkan. Pihak maskapai berharap para penumpang bisa menjadi loyal untuk terus menggunakan jasa maskapai tersebut dan pihak penumpang menginginkan pelayanan yang terbaik bisa didapat ketika menggunakan maskapai tersebut. Hal ini dapat dicapai ketika terdapat media perantara yang bisa digunakan sebagai acuan pihak maskapai untuk meningkatkan pelayanannya. Salah satu medianya yang digunakan ialah adanya forum maskapai penerbangan yang dapat digunakan untuk memberikan penilaian oleh penumpang terhadap kinerja dan pelayanan maskapai selama menggunakan jasa dari maskapai tersebut.

*Skytrax* merupakan perusahaan penasihat mutu dan riset spesialis bagi industri transportasi udara yang berbasis di London, UK yang bekerja sama dengan maskapai dan bandara di seluruh dunia [1]. *Skytrax* berdedikasi untuk meningkatkan pengalaman pelanggan bagi maskapai dan bandara di seluruh dunia, memberikan wawasan unik, kepiawaian, pengalaman akan persoalan mutu dan pemikiran inovatif untuk membantu mewujudkan terjadinya perubahan di industri transportasi udara. Selain itu, *Skytrax* juga bertindak sebagai pemberi peringkat transportasi udara internasional.

Pada penelitian ini akan dilakukan analisis *mining data* pada *Skytrax Airlines Review dataset.* Hal ini dilakukan untuk mengetahui hal apa saja yang membuat penumpang merekomendasikan jasa maskapai penerbangan kepada orang lain berdasarkan pelayanan yang telah diberikan oleh pihak maskapai penerbangan untuk memberikan pelayanan lanjutan yang lebih baik dimana proses tersebut merupakan salah satu bagian dari data *mining*. Fokus utama data mining yang akan digunakan adalah klasifikasi, dimana algoritma yang akan digunakan untuk mengklasifikasikan dataset adalah *Naïve Bayes*, *Random Forest*, dan *Logistic Regression*. Diharapkan dengan menggunakan tiga metode tersebut mendapatkan nilai akurasi yang baik, sehingga bisa memberi hasil yang tepat dan bisa dijadikan sebagai acuan dalam merekomendasikan maskapai penerbangan.

## II. TINJAUAN PUSTAKA

### A. *Data Mining*

Data *mining* proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* yang digunakan untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar [2]. Tahapan yang dilakukan pada proses *data mining* diawali dari seleksi data dari data sumber ke data target, tahap *pre-processing* yang digunakan untuk memperbaiki kualitas data, transformasi, data mining serta tahap interpretasi dan evaluasi yang menghasilkan output berupa pengetahuan baru yang diharapkan memberikan kontribusi yang lebih baik [3].

### B. *Preprocssing Data*

*Preprocessing* data merupakan tahapan pertama dan merupakan tahapan yang penting dalam *data mining* atau *data analysis* [4]. Pada umunya data pada dunia nyata yang masih mentah atau biasa disebut data primer memiliki kekurangan diantaranta tidak lengkap, banyak *noise*, dan juga tidak konsisten. Oleh karenanya tahapan ini sangat

penting untuk memastikan data sumber diolah sehingga menghasilkan *dataset* yang siap dipakai pada tahapan selanjutnya.

1. Data *Cleaning*

Data *cleaning* merupakan operasi yang melakukan perbaikan dan memfilter terhadap data-data yang rusak, hilang (*missing value)* atau salah (*error*). Selain itu, proses ini juga mengurangi detail data yang tidak perlu. Pembersihan data juga akan mempengaruhi performasi dari teknik data mining karena data yang ditangani akan berkurang jumlah dan kompleksitasnya [5].

2. Data *Transformation*

Data *transformation* atau data transformasi merupakan proses penguahan atau penggabungan data ke dalam format yang sesuai untuk diproses dalam *data* mining, dimana hasil proses tersebut dapat diterapkan lebih efisien. Beberapa metode *data mining* membutuhkan format data yang khusus sebelum bisa diaplikasikan. Sebagai contoh beberapa metode standar seperti *clustering* hanya bisa menerima input data kategorikal. Karenanya data berupa angka numerik yang berlanjut perlu dibagi-bagi menjadi beberapa interval [5].

3. Data *Integration*

Integrasi data merupakan penggabungan data dari berbagai *database* ke dalam satu *database* baru. Proses ini harus dilakukan dengan cermat untuk menghindari ketidak konsistenan data yang sudah ditetapkan. Integrasi data dilakukan pada atribut-aribut yang mengidentifikasikan variabel yang unik seperti atribut nama, jenis produk, nomor pelanggan dan lainnya. Integrasi data perlu dilakukan secara cermat karena kesalahan pada integrasi data bisa menghasilkan hasil yang menyimpang. Sebagai contoh bila integrasi data berdasarkan jenis produk ternyata menggabungkan produk dari kategori yang berbeda maka akan didapatkan korelasi antar produk yang sebenarnya tidak ada [5].

4. Data *Selection*

Pemilihan data dilakukan karena data yang ada pada *database* sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari *database* [5].

### *C. Statistika Deskriptif*

Statistika adalah ilmu mengumpulkan, menata, menyajikan, menganalisa dan menginterpretasikan data angka dengan tujuan membantu pengambilan keputusan yang efektif [6]. Statistika menerjemahkan rumusan matematika menjadi bahasa yang dapat dipahami oleh masyarakat dengan latar pendidikan non-statistika dan juga menerjemahkan berbagai macam persoalan kedalam rumusan matematika.

1. Rata-rata

Rata-rata atau sering disebut *mean* merupakan rasio dari total nilai pengamatan dengan banyaknya pengamatan [6]. Rata-rata merupakan suatu ukuran pemusatan data yang sering digunakan dalam mendeskripsikan suatu data. Berikut adalah rumus untuk menghitung nilai rata-rata.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (1)$$

Keterangan :

$\bar{x}$ = Nilai rata-rata

$x_i$ = Data pengamatan ke-*i* dengan nilai *i* = 1, 2, …, *n*

$n$ = Banyaknya pengamatan

2. Varians (Ragam)

Varians adalah rata-rata hitung deviasi kuadrat setiap data terhadap rata-rata hitungnya. Rumus yang digunakan untuk mengitung varian adalah sebagai berikut.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \quad (2)$$

Keterangan :

$s^2$ = Varians

$x_i$ = Data pengamatan ke-*i* dengan nilai *i* = 1, 2, …, *n*

$\bar{x}$ = Rata-rata

$n$ = Banyaknya data

3. Median

Median salah satu teknik penjelasan kelompok yang didasarkan atas nilai tengah dari kelompok data yang telah disusun urutannya dari yang terkecil sampai yang terbesar, atau sebaliknya dari yang terbesar sampai yang terkecil [6]. Rumus *median* adalah sebagai berikut.

Untuk *n* ganjil :

$$Me = x_{\frac{1}{2}(n+1)} \quad (3)$$

Untuk *n* genap :

$$Me = \frac{x_{n/2} + x_{n/2}+1}{2} \quad (4)$$

Keterangan :

*Me* =Median atau nilai tengah data setelah diurutkan

$\chi$ = Data

$n$ = Banyaknya data

4. Modus

Modus segugus pengamatan adalah nilai yang terjadi paling sering atau yang mempunyai frekuensi paling tinggi [6]. Modus tidak selalu ada. Hal ini terjadi bila semua pengamatan mempunyai frekuensi terjadi yang sama.

Modus untuk data tunggal dapat langsung diketahui dengan menghitung frekuensi masing-masing data yang didapat. Sedangkan modus dalam data kelompok dapat dicari dengan rumus berikut.

$$\text{Modus} = \text{TBB Kelas Modus} + \text{i}\left(\frac{d1}{d1+d2}\right) \quad (5)$$

Keterangan :

TBB = Tepi Batas Bawah

$d_1$ = Beda frekuensi kelas Modus dengan frekuensi kelas sebelumnya

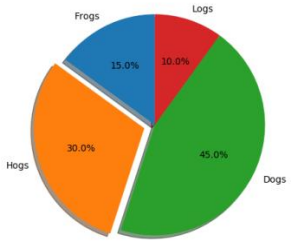
$d_2$ = Beda frekuensi kelas Modus dengan frekuensi kelas sesudahnya

### *D. Visualisasi Data*

Visualisasi data merupakan bentuk eksplorasi data menggunakan gambar-gambar yang bertujuan untuk memberikan informasi yang menarik dan mudah dipahami. Beberapa bentuk visualisasi data yang digunakan adalah sebagai berikut.

1. *Pie Chart*

*Pie chart* merupakan grafik yang merepresentasikan suatu data. Tiap potongnya melambangkan nilai tiap variabel yang sedang diamati [7]. Bentuk *pie chart* disajikan pada Gambar 1 sebagai berikut.




**Gambar 1.** Bentuk *Pie Chart*

*E. Feature Extraction* menggunakan *Principal Component Analysis (PCA)*

Ekstraksi fitur merupakan proses yang bertujuan untuk menentukan ciri-ciri. Adapun tahapan ekstraksi fitur menggunakan PCA sebagai berikut [8].

1. Hitung rata-rata seluruh sampel data diperoleh dengan menggunakan persamaan:

$$\bar{x}_j = \frac{\sum_{ij=1}^{n} x_{ij}}{n} \quad (6)$$

2. *Adjusted* data (data yang telah disesuaikan) adalah hasil pengurangan dari setiap data dengan rata-rata setiap data yang diperoleh dengan persamaan sebagai berikut:

$$Adjusted\ data = x_{ij} - \bar{x}_j$$
$$X' = Adjusted\ data \quad (7)$$

3. Hitung matrik kovarian (*c*) dihitung dengan menggunakan persamaan berikut:

$$c = \frac{1}{M} X' X'^{T} \quad (8)$$

4. Hitung nilai eigen dan vektor eigen dari matrik kovarian dihitung dengan menggunakan persamaan karakteristik berikut ini:

$$c - \lambda I = 0$$
$$(c - \lambda I)v = 0 \quad (9)$$

Dimana *c* adalah matrik kovarian, *I* adalah matrik identitas, $\lambda$ adalah nilai eigen dan *v* adalah vektor eigen.

5. Hitung nilai eigen yang terbesar yang berkorespondensi terhadap nilai vektor eigen yang tebesar dipilih menjadi *Principal Component*. Vektor eigen yang disusun dari yang terbesar ke yang terkecil dipilih menjadi vektor fitur.

$$v = (eig_1, eig_2, eig_3, ..., eig_n) \quad (10)$$

6. Mencari PC dengan sebagai rata-rata.

$$PC = X' \times v \quad (11)$$

7. Langkah berikutnya untuk melakukan transformasi data untuk menghasilkan PCA.

$$PCA = PC^{T} \times X'^{T} \quad (12)$$

*F. Analisis Klasifikasi*

*Classification* adalah satu bentuk analisis data yang menghasilkan model untuk mendeskripsikan kelas data yang penting. Classification memprediksi kategori (*discrete, unordered*) ke dalam label class. Classification merupakan proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau class data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Beberapa model yang digunakan adalah sebagai berikut.

1. *Naïve Bayess*

*Naïve Bayess* metode yang tidak memiliki aturan dan menggunakan cabang matematika yang dikenal dengan teori probabilitas untuk mencari peluang terbesar dari kemungkinan klasifikasi, dengan cara melihat frekuensi setiap klasifikasi pada data training. *Naïve bayes* merupakan metode klasifikasi populer dan termasuk dalam sepuluh algoritma terbaik dalam data mining, algoritma ini juga dikenal dengan nama *Idiot's Bayes*, *Simple Bayes* dan *Independence Bayes* [9].

Klasifikasi Bayes di dasarkan pada *teorema bayes*, yaitu sebagai berikut.

$$P(x \mid y) = \frac{P(y \mid x) P(x)}{P(y)} \quad (13)$$

Keterangan:

Y : Data dengan kelas yang belum diketahui
X : Hipotesis data y yang merupakan suatu kelas spesifik
P(x|y) : Probabilitas hipotesis *x* berdasarkan kondisi *y* (*posteriori probability*)
P(x) : Probabilitas hipotesis *x* (*prior probability*)
P(y|x) : Probabilitas *y* berdasarkan kondisi hipotesis *x*
P(y) : Probabilitas dari *y*

2. *Random Forest*

*Random Forest* adalah pengklasifikasi yang terdiri dari kumpulan pengklasifikasi pohon terstruktur {h(x,Θk), k=1, ...} dimana {Θk} adalah vektor acak terdistribusi yang identik independen dan masing-masing pohon melemparkan unit suara untuk kelas paling populer diinput. Random forest merupakan pengembangan dari *decision tree*, dimana setiap decision tree telah dilakukan training data menggunakan sampel individu dan setiap atribut dipecah pada tree yang dipilih antara atribut subset yang bersifat acak [10]. Dalam perkembangannya, sejalan dengan bertambahnya *dataset*, maka *tree* pun ikut berkembang. Penempatan *tree* yang saling berjauhan membuat apabila terdapat *tree* disekitar *tree x* berarti pohon tersebut merupakan perkembangan *tree x*.

3. *Logistic Regression*

Regresi logistik merupakan suatu metode analisis data yang mendeskripsikan antara variabel respon dengan satu atau lebih variabel prediktor. Regresi logistik biner variabel responnya yang bersifat dikotomus yang terdiri dari dua kategori yaitu 0 dan 1, sehingga variabel respon akan mengikuti distribusi Bernoulli. Pengujian parameter dalam regresi logistik biner dilakukan baik secara serentak maupun individu. Statistik uji yang digunakan dalam uji serentak adalah statistik uji G atau likelihood ratio test. Sedangkan statistik uji yang digunakan dalam uji parsial adalah statistik uji Wald [11]

*G. Model Evaluation and Selection*

1. *Confusion Matrix*

*Confusion Matrix* adalah tabel yang digunakan untuk menjelaskan performa dari model klasifikasi (*classifier*) pada data *testing* berdasarkan nilai yang didapat. Setiap kolom pada *confusion matrix* menunjukkan kelas prediksi untuk *query* klasfikasi, sedangkan tiap barisnnya menunjukkan kelas aktual dari *query* [12]. Evaluasi model *confussion matrix* menggunakan tabel seperti dibawah ini.

**Tabel 1.** Matrik Klasifikasi untuk *Model 2 Class*

| Classification | | Predicted Class | |
|---|---|---|---|
| | | **Yes** | **No** |
| **Observed Class** | **Yes** | (True Positive) TP | (False Negative) FN |
| | **No** | (False Positive) FP | (True Negative) TN |

Akurasi dapat dihitung dengan menggunakan rumus berikut.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

Keterangan :
TP : Jumlah kasus positif yang diklasifikasikan positif
FP : Jumlah kasus negatif yang diklasifikasikan positif
TN : Jumlah kasus negatif yang diklasifikasikan negatif
FN : Jumlah kasus positif yang diklasifiikasikan negative

2. *Kurva Receiver Operating Characteristics (ROC)*

Kurva *Receiver Operating Characteristics (ROC*) banyak digunakan untuk menilai hasil prediksi, kurva ROC merupakan teknik untuk memvisualisasikan, mengatur, dan memilih pengklasifikasian berdasarkan kinerja mereka [12]. Kurva ROC adalah tool dua dimensi yang dipergunakan untuk menilai hasil kinerja klasifikasi yang menggunakan dua class sebagai keputusannya, objek dipetakan ke salah satu elemen dari himpunan pasangan, positif atau negatif. Pada kurva ROC, TP rate diplot pada sumbu Y dan FP rate diplot pada sumbu X. Untuk klasifikasi data mining, nilai AUC dapat dibagi menjadi beberapa kelompok [13].

**Tabel 2.** Kelompok Nilai AUC

| Nilai AUC | Kelompok |
|---|---|
| 0,90 – 1,00 | *Excellent Classification* |
| 0,80 – 0,90 | *Good Classification* |
| 0,70 – 0,80 | *Fair Classification* |
| 0,60 – 0,70 | *Poor Classification* |
| 0,50 – 0,60 | *Failure* |

The Area Under Curve (AUC) dihitung untuk mengukur perbedaan performasi metode yang digunakan. AUC dihitung menggunakan rumus [14].

$$\theta^{\gamma} = \frac{1}{mn}\sum_{j}^{n}=1\sum_{i}^{m}\psi\left(x_i^{\gamma}, x_j^{\gamma}\right)$$

$$\text{Dimana, } \psi\left(X,Y\right) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X \\ 1 & Y > X \end{cases} \quad (15)$$

Keterangan :
X = Ouput Positif
Y = Output Negatif

3. *Holdout Validation*

Pada kondisi terbatasnya data yang digunakan untuk *training* dan *testing*, diperlukan metode untuk mendapatkan hasil tingkat akurasi dari sebuah metode pada *machine learning*. Salah satu cara untuk validasi adalah dengan menggunakan metode holdout. Metode holdout adalah metode yang akan menyediakan sejumlah data untuk digunakan sebagai data *testing*, dan sisanya sebagai data *training* [15].

Saat proses pengacakan data untuk dibagi sebagai data *training* dan *testing*, sangat mungkin terjadi *overrepresented* pada salah satu atau lebih klasifikasi. Dalam artian bahwa klasifikasi tersebut dominan dibandingkan klasifikasi lainnya, sehingga data *training* dan *testing* yang tercipta menjadi tidak representatif. Maka dari itu diperlukan prosedur *stratification holdout*, dimana dengan prosedur ini dapat dijamin bahwa setiap klasifikasi dapat terwakili pada data training dan testing yang tercipta secara proporsional. Kelas yang terbagi dari hasil proses holdout proporsinya harus sedekat mungkin dengan proporsi aslinya [16].

Dilakukan perulangan terhadap seluruh proses training dan testing beberapa kali dengan data training dan testing yang teracak. Kemudian diambil nilai rata-ratanya. Prosedur ini dikatakan sebagai *repeated holdout*.

4. *K-Fold Cross Validation*

*Cross validation* adalah salah satu metode untuk evaluasi model prediksi. Metode ini bisa mengindikasikan seberapa baik model ini memprediksi data yang belum pernah dilihat sehingga pengujian ini membuat model tidak *overfitting* [17].

Langkah pertama dalam melakukan cross validation adalah melakukan iterasi sebanyak *k*, dimana selama iterasi berlangsung data *testing* dan data *training* tidak pernah sama karena data *testing* pada iterasi sebelumnya menjadi data training dan pada iterasi selanjutnya diambil beberapa data dari data *training* di iterasi sebelumnya untuk dijadikan sebagai *dataset*.

### *H. Maskapai Penerbangan*

Perusahaan penerbangan adalah perusahaan milik swasta atau pemerintah yang khusus menyelenggarakan pelayanan angkutan udara untuk penumpang umum baik baik yang berjadwal maupun yang tidak berjadwal [18].

## III. METODOLOGI PENELITIAN

### *A. Sumber Data*

Pada penelitian ini data yang digunakan adalah data sekunder, diambil dari *website* kaggle yang diakses pada hari Jumat, tanggal 8 Mei 2020 pukul 21.40. Data yang digunakan adalah data penilaian penumpang terhadap maskapai pesawat terbang populer di seluruh dunia pada tahun 2006-2019.

### *B. Variabel Penelitian*

Variabel penelitian yang digunakan tercantum dalam Tabel 1.

**Tabel 3.** Variabel Penelitian

| Variabel | Keterangan |
|---|---|
| *Airlines* | Nama maskapai |
| *Overall* | Poin keseluruhan yang diberikan untuk perjalanan |
| *Author* | Nama penulis penilaian |
| *Review date* | Tanggal memberi penilaian |
| *Customer Review* | Penilaian pelanggan dalam format teks |
| *Aircraft* | Jenis pesawat |
| *Traveller type* | Jenis traveller |
| *Cabin* | Kabin |
| *Date flown* | Tanggal penerbangan |
| *Seat comfort* | Kenyamanan tempat duduk |
| *Cabin service* | Pelayanan kabin |
| *Food beverage* | Makanan dan minuman |
| *Ground Service* | Tata operasi darat |
| *Entertainment* | Hiburan |
| *Value for money* | Transparansi dan akuntabilitas |
| *Recommended* | Biner, variabel target |

### *C. Struktur Data*

Adapun struktur data yang akan digunakan pada penelitian ini sebagai berikut.

**Tabel 4.** Struktur Data

| Data ke-*i* | $X_1$ | $X_2$ | ⋯ | $X_{16}$ |
|---|---|---|---|---|
| 1 | $X_{11}$ | $X_{12}$ | ⋯ | $X_{1;16}$ |
| 2 | $X_{21}$ | $X_{22}$ | ⋯ | $X_{2;16}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| n | $X_{n;1}$ | $X_{n;2}$ | ⋯ | $X_{n;16}$ |

Keterangan :
$X_1$ : data variabel ke-1 *(Airlines)*
$X_2$ : data variabel ke-2 *(Overall)*
⋮
$X_{16}$: data variabel ke-16 (*Recommended)*

*D. Langkah Analisis*

Langkah-langkah penelitian yang telah dilakukan berdasarkan dengan tujuan penelitian ini adalah sebagai berikut.

1. Mengidentifikasi masalah dan tujuan penelitian.
2. Mencari dan mengumpulkan data.
3. Melakukan *preprocessing* data.
4. Menghitung statistika deskriptif.
5. Melakukan visualisasi data.
6. Melakukan *feature extraction* menggunakan *principal component analysis (PCA)*.
7. Melakukan analisis klasifikasi pada *dataset*, meliputi *Random Forest, Gaussian Naive Bayess*, dan *Logistic Regression*.
8. Melakukan *model evaluation and selection* berdasarkan metode klasifikasi yang digunakan.
9. Menarik kesimpulan dan saran

## IV. ANALISIS DAN PEMBAHASAN

*A. Preprocessing Data*

1. Data Cleaning

Berdasarkan penelitian ini, untuk menentukan faktor-faktor yang mempengaruhi rekomendasi maskapai penerbangan oleh penumpang dimana didasarkan pada beberapa kriteria penilaian. Variabel dari *dataset* yang memiliki tipe skala yaitu *seat comfort*, *cabin service*, *food beverage*, *ground service*, *entertainment*, dan *value for money*. Variabel *airline* sebagai tambahan informasi untuk mengetahui terdapat berapa jenis maskapai penerbangan yang ada didalam dataset. Untuk beberapa variabel selain yang sudah disebutkan, kemudian dilakukan penghapusan karena dari penelitian ini hanya mempertimbangkan beberapa faktor saja yang mempengaruhi rekomendasi oleh penumpang pesawat.

Salah satu contoh variabel yang tidak digunakan adalah *author*, karena informasi nama penumpang dalam prioritas penelitian ini tidak dipertimbangkan. Adapun variabel *customer review* yang merupakan ulasan dari penumpang pesawat dan pada variabel ini juga terdapat beberapa simbol yang tidak terbaca oleh algoritma, sehingga dilakukan penghapusan.

2. Imputasi *Missing Value*

Data *skytrax airline reviews* yang sudah dilakukan tahap *cleaning data*, kemudian dilakukan cek *missing value*. Data memiliki variabel yang bersifat kategorik, sehingga pada penelitian ini imputasi *missing value* dilakukan dengan menggunakan nilai modus dari setiap variabel.

*B. Statistika Deskriptif*

Statistika deskriptif digunakan untuk mengetahui karakteristik dari data *skytrax airline reviews*, hasilnya ditunjukkan pada Tabel 5.

**Tabel 5.** Statistika Deskriptif

| Variabel | *Unique* | *Top* | *Frequency* |
|---|---|---|---|
| *Airline* | 81 | Spirit Airlines | 2934 |
| *Seat comfort* | 5 | 1 | 20488 |
| *Cabin service* | 5 | 5 | 23658 |
| *Food beverage* | 5 | 1 | 27779 |
| *Ground Service* | 5 | 1 | 35186 |
| *Entertainment* | 5 | 1 | 42329 |
| *Value for money* | 5 | 1 | 21834 |
| *Recommended* | 2 | *no* | 35401 |

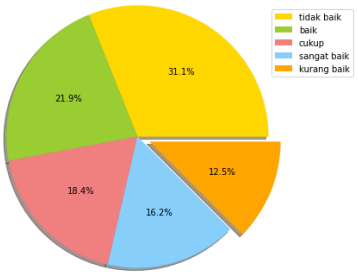
Tabel 5 menunjukkan bahwa maskapai penerbangan yang paling banyak diminati oleh penumpang adalah *Spirit Airlines*, dengan total sebanyak 2934 penumpang yang memilih.

*C. Visualisasi Data*

Visualisasi data yang digunakan pada penelitian ini di antaranya adalah *pie chart* pada variabel-variabel dari data *skytrax airline reviews*.

1. Persentase penilaian penumpang terhadap kenyamanan tempat duduk

Visualisasi untuk menunjukkan persentease penilaian penumpang terhadap kenyamanan tempat duduk pada data *skytrax airline reviews* dapat menggunakan *Pie Chart* yang disajikan pada Gambar 2 berikut.

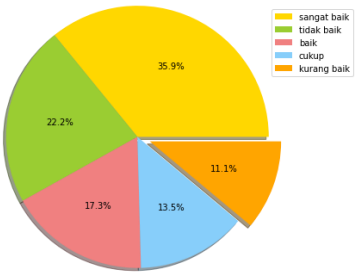



**Gambar 2.** Persentase penilaian kenyamanan tempat duduk

Gambar 2 menunjukkan bahwa dari 65.947 penumpang yang memberikan penilaian, 31,1% menilai bahwa kualitas kenyamanan tempat duduk yang diberikan tidak baik, 16.2% menilai sangat baik dan hanya sedikit penumpang yang memberikan penilaian yang kurang baik.

2. Persentase penilaian penumpang terhadap pelayanan kabin

Visualisasi untuk menunjukkan persentease penilaian penumpang penumpang terhadap pelayanan kabin pada data *skytrax airline reviews* dapat menggunakan *Pie Chart* yang disajikan pada Gambar 3 berikut.



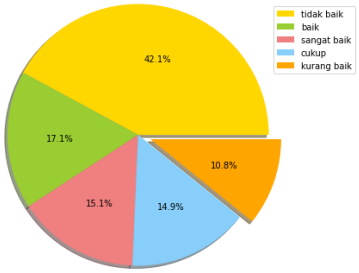

**Gambar 3.** Persentase penilaian pelayanan kabin

Gambar 3 menunjukkan bahwa penilaian yang diberikan penumpang untuk pelayanan kabin yang paling banyak ialah sangat baik dengan persentase 35.9% dan

paling sedikit sebesar 11.1% untuk penilaian yang kurang baik.

3. Persentase penilaian penumpang terhadap makanan dan minuman

Visualisasi untuk menunjukkan persentease penilaian penumpang penumpang terhadap makanan dan minuman pada data *skytrax airline reviews* dapat menggunakan *Pie Chart* yang disajikan pada Gambar 4 berikut.

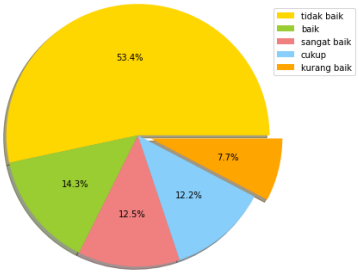



**Gambar 4.** Persentase penilaian makanan dan minuman

Gambar 4 menunjukkan bahwa sebanyak 42.1% dari total penumpang menilai makanan dan minuman yang diberikan oleh pihak maskapai penerbangan dengan tidak baik dan 17.1% dengan baik.

4. Persentase penilaian penumpang terhadap hiburan

Visualisasi untuk menunjukkan persentease penilaian penumpang penumpang terhadap hiburan pada data *skytrax airline reviews* dapat menggunakan *Pie Chart* yang disajikan pada Gambar 5 berikut.

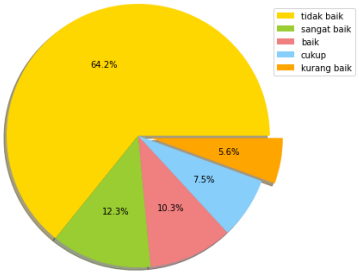



**Gambar 5.** Persentase penilaian hiburan

Gambar 5 menunjukkan bahwa dari 65.947 penumpang yang memberikan penilaian, 53.4% menilai bahwa hiburan yang diberikan tidak baik dan hanya 7.7% penumpang yang memberikan penilaian yang kurang baik.

5. Persentase penilaian penumpang terhadap tata operasi darat

Visualisasi untuk menunjukkan persentease penilaian penumpang penumpang terhadap tata operasi darat pada data *skytrax airline reviews* dapat menggunakan *Pie Chart* yang disajikan pada Gambar 6 berikut..

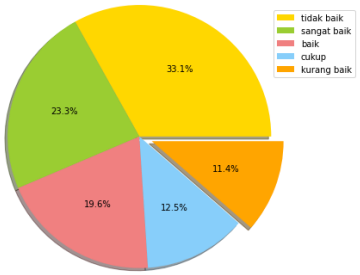



**Gambar 6.** Persentase penilaian tata operasi darat

Gambar 6 menunjukkan bahwa untuk tata operasi darat penumpang paling banyak menilai dengan tidak baik dengan persentase 64.2 dan sebanyak 5.6% penumpang menilai dengan kurang baik.

6. Persentase penilaian penumpang terhadap transparansi dan akuntabilitas

Visualisasi untuk menunjukkan persentease penilaian penumpang penumpang terhadap transparansi dan akuntabilitas pada data *skytrax airline reviews* dapat menggunakan *Pie Chart* yang disajikan pada Gambar 7 berikut.

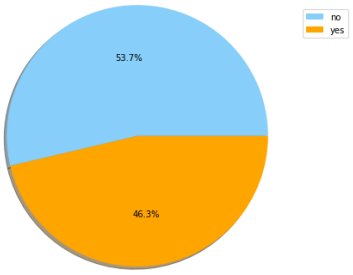



**Gambar 7.** Persentase penilaian transparansi dan akuntabilitas

Gambar 7 menunjukkan bahwa penumpang memberikan penilaian tidak baik sebesar 33.1% untuj transparansi dan akuntabilitas terhadap maskapai penerbangan, disusul dengan 23.3% penumpang menilai dengan sangat baik dan penilaian paling sedikit diberikan untuk kurang baik dengan persentase 11.4%

7. Persentase rekomendasi penumpang terhadap maskapai

Visualisasi untuk menunjukkan persentease rekomendasi penumpang terhadap maskapai pada data *skytrax airline reviews* dapat menggunakan *Pie Chart* yang disajikan pada Gambar 8 berikut.

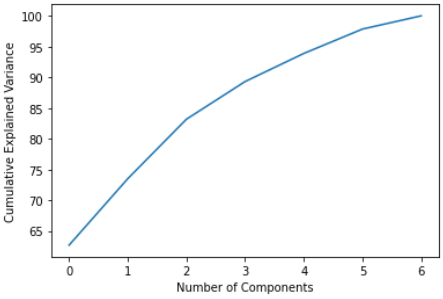



**Gambar 8.** Persentase rekomendasi maskapai

Gambar 8 menunjukkan bahwa dari keseluruhan penilaian yang diberikan penumpang berdasarkan pengalamannya, 53.7% penumpang memutuskan untuk tidak merekomendasikan maskapai penerbangan kepada orang lain dan sebanyak 46.3% memutuskan untuk merekomendasikan maskapai penerbangan kepada orang lain.

*D. Feature Extraction* menggunakan *Principal Component Analysis (PCA)*

Pada penelitian ini untuk mengetahui berapa PC yang harus diambil terlebih dahulu divisualisasikan untuk nilai kumulatif variansi data yang disajikan pada Gambar 9.




**Gambar 9.** *Scree Plot* Kumulatif varians data *Skytrax Airline Reviews*

Nilai kumulaitf *varians* pada *scree plot* dapat disajikan pada Tabel 6 sebagai berikut.

**Tabel 6.** Nilai Kumulatif Varians

| Jumlah PC | Nilai Kumulatif *varians* |
|---|---|
| 1 | 62,68% |
| 2 | 73,50% |
| 3 | 83,18% |
| 4 | 89,31% |
| 5 | 93,88% |
| 6 | 97,85% |
| 7 | 100,00% |

Berdasarkan Gambar 9 dan Tabel 96 dapat dilihat bahwa minimal 80% *varians* data *Skytrax Airline Reviews* bisa dijelaskan oleh *principal component (PC)* dengan mengambil 5 *principal component (PC)*, karena menunjukkan nilai kumulatif *varians* sebesar 83,18%.

Berikut adalah nilai komponen PCA untuk data *Skytrax Airline Reviews* yang disajikan pada Tabel 7 sebagai berikut.

**Tabel 7.** Nilai Komponen PCA

| Variabel | PC1 | PC2 | PC3 | PC4 | PC 5 |
|---|---|---|---|---|---|
| *Seat comfort* | 0,406 | -0,110 | 0,129 | -0,433 | 0,348 |
| *Cabin service* | 0,352 | 0,174 | -0,587 | 0,638 | 0,121 |
| *Food beverage* | 0,403 | -0,269 | 0,076 | 0,054 | 0,644 |
| *Ground Service* | 0,280 | 0,727 | 0,591 | 0,407 | 0,008 |
| *Entertainment* | 0,332 | -0,581 | 0,428 | 0,306 | -0,437 |
| *Value for money* | 0,429 | 0,076 | -0,169 | -0,350 | -0,302 |
| *Recommended* | 0,419 | 0,116 | -0,267 | -0,268 | -0,408 |

Berdasarkan Tabel 7 dapat dilihat terdapat 5 persamaan kombinasi linear. Selain itu, dapat digunakan untuk mengetaui kontribusi terbesar pada setiap PC ada pada variabel yang mana, dengan melihat nilai mutlak dari PC setiap variabel tersebut. Untuk kontribusi terbesar pada PC1 adalah variabel *value for money* sebesar 0,429 dan kontribusi terbesar pada PC2 ialah variabel *ground service* sebesar 0,727. Begitupun juga dengan persamaan kombinasi linear yang lain.

## *E. Analisis Klasifikasi berdasarkan Holdout Non Stratified*

Pengujian data dengan *Holdout Non Stratified* untuk mengetahui nilai akurasi, presisi, dan *recalls* dilakukan dengan menggunakan tiga metode klasifikasi, yang ditunjukkan pada Tabel 6.

**Tabel 6.** Nilai Akurasi dari semua algoritma klasifikasi

| Metode | Akurasi | Presisi | *Recalls* |
|---|---|---|---|
| *Gaussian Naive Bayes* | 0,917 | 0,917 | 0,916 |
| *Random Forest* | 0,930 | 0,923 | 0,925 |
| *Logistic Regression* | 0,937 | 0,932 | 0,927 |

Tabel 6 memperlihatkan perbandingan ketepatan klasifikasi dari ketiga metode berdasarkan *Holdout Non Stratified*. Dapat disimpulkan bahwa metode *Logistic Regression* memiliki nilai akurasi yang jauh lebih tinggi daripada metode klasifikasi yang lainnya, yaitu sebesar 0,937.

## *F. Analisis Klasifikasi berdasarkan Holdout Stratified*

Pengujian data dengan *Holdout Stratified* untuk mengetahui nilai akurasi, presisi, dan *recalls* dilakukan dengan menggunakan tiga metode klasifikasi, yang ditunjukkan pada Tabel 7.

**Tabel 7.** Nilai Akurasi dari semua algoritma klasifikasi

| Metode | Akurasi | Presisi | *Recalls* |
|---|---|---|---|
| *Gaussian Naive Bayes* | 0,917 | 0,917 | 0,915 |
| *Random Forest* | 0,933 | 0,929 | 0,928 |
| *Logistic Regression* | 0,939 | 0,932 | 0,935 |

Tabel 7 memperlihatkan perbandingan ketepatan klasifikasi dari ketiga metode berdasarkan *Holdout Stratified*. Dapat disimpulkan bahwa metode *Logistic Regression* memiliki nilai akurasi yang jauh lebih tinggi daripada metode klasifikasi yang lainnya, yaitu sebesar 0,939.

## *G. Analisis Klasifikasi berdasarkan K-Fold Cross Validation*

Pengujian data dengan *K-Fold Cross Validation* untuk mengetahui nilai akurasi, presisi, dan *recalls* dilakukan dengan menggunakan tiga metode klasifikasi, yang ditunjukkan pada Tabel 8.

**Tabel 8.** Nilai Akurasi dari semua algoritma klasifikasi

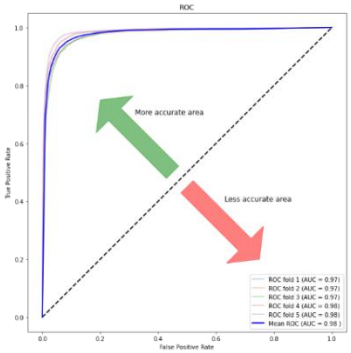
| Metode | Akurasi | Presisi | *Recalls* |
|---|---|---|---|
| *Gaussian Naive Bayes* | 0,914 | 0,917 | 0,915 |
| *Random Forest* | 0,934 | 0,928 | 0,929 |
| *Logistic Regression* | 0,937 | 0,927 | 0,933 |

Tabel 8 memperlihatkan perbandingan ketepatan klasifikasi dari ketiga metode berdasarkan *K-Fold Cross Validation*. Dapat disimpulkan bahwa metode *Logistic Regression* memiliki nilai akurasi yang jauh lebih tinggi daripada metode klasifikasi yang lainnya, yaitu sebesar 0,937.

## *H. Kurva ROC*

Untuk mengukur kinerja suatu sistem atas dasar nilai kesalahan yang terjadi dan tingkat kesuksesan pengenalan suatu sistem (specificity), maka ROC (Receiver Operating Curve) dapat digunakan untuk menghitung nilai kesalahan dan nilai kesuksesan suatu system.

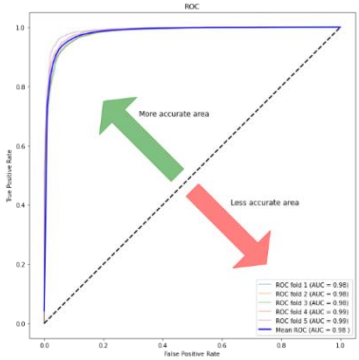
1. *Random Forest*




**Gambar 10.** Kurva ROC dengan *Random Forest*

Gambar 10 menunjukkan bahwa nilai AUC dari model adalah sebesar 0,98 yang menggunakan metode klasifikasi *Random Forest*. Sehingga pada penelitian ini model diklasifikasikan sebagai model yang sangat baik (*excellent*) karena nilai AUC berada di interval nilai 0,9 hingga 1.

2. *Logistic Regression*




**Gambar 11.** Kurva ROC dengan *Logistic Regression*

Gambar 11 menunjukkan bahwa nilai AUC dari model adalah sebesar 0,98 yang menggunakan metode klasifikasi *Logistic Regression*. Sehingga pada penelitian ini model diklasifikasikan sebagai model yang sangat baik (*excellent*) karena nilai AUC berada di interval nilai 0,9 hingga 1.

## V. KESIMPULAN DAN SARAN

### *A. Kesimpulan*

Kesimpulan yang diperoleh dari proses analisis dan pembahasan yang telah dilakukan adalah sebagai berikut:

1. Tahap *preprocessing* yang dilakukan pada data *Skytrax Airline Reviews* mula-mula adalah data *cleaing*. Kemudian, dilanjut dengan imputasi *missing value* pada setiap variabel. Imputasi *missing value* dilakukan dengan menggunakan nilai modus.
2. Dari ulasan yang telah diberikan oleh penumpang, maskapai penerbangan yang paling banyak digunakan ialah *Spirit Airlines*, dari variabel seat comfort, food beverage, ground service, entertainment, value for money penumpang paling banyak memberikan penilaian tidak baik, pada variabel cabin service penumpang paling banyak memberikan penilaian sangat baik dan sebanyak 35.401 ulasan penumpang tidak merekomendasikan maskapai penerbangan kepada orang lain.
3. Metode klasifikasi *Logistic Regression* merupakan metode yang memiliki hasil klasifikasi terbaik dalam kombinasi data *training* sebesar 70% dan data testing sebesar 30%.
4. Berdasarkan kurva ROC menunjukkan bahwa nilai AUC yang dimiliki sebesar 0,98. Hal ini mengandung arti bahwa model diklasifikasikan sebagai model yang sangat baik.

### *B. Saran*

Berdasarkan kesimpulan yang diperoleh, dapat dirumuskan saran sebagai pertimbangan penelitian selanjutnya adalah memilih metode yang memiliki nilai akurasi dan presisi yang lebih baik.

## DAFTAR PUSTAKA


[1] "Tentang Kami : Skytrax," [Online]. Available: https://skytraxresearch.com/id/about-us/. [Accessed 11 Mei 2020].

[2] E. Turban, Decision Support Systems and Intelligent Systems Edisi Bahasa Indonesia Jilid I, Yogyakarta: Andi, 2005.

[3] U. Ayyad, Advances in Knowledge Discovery and Data Mining, MIT Press, 1996.

[4] Q. Y. Fang and L. X. Wei, "A Data Preprocessing Algorith for Classification Model Base On Rough Sets," in *International Conference on Solid State Devices and Material Science*, 2012, pp. 2025-2029.

[5] S. Garcia, J. Luengo and F. Herrera, in *Data Preprocessing in Data Mining*, Cham, Switzerland, Springer International Publishing, pp. 195-243.

[6] R. E. Walpole and R. H. Myers, Ilmu Peluang dan Statistika untuk Insinyur dan Ilmuwan Edisi ke-4, Bandung: Penerbit ITB, 1995.

[7] S. Nugroho, Dasar-Dasar Metode Statistika, Jakarta: Grassindo, 2008.

[8] . M. Z. Nasution, "Face Recognition based Feature Extraction using Principal Component Analysis (PCA)," *Journal of Informatics and Telecommunication Engineering*, vol. 3, pp. 192-201, 2020.

[9] M. Bramer, Principles of Data Mining London, Los A.: Springer Clark, 2007.

[10] R. Lior and M. Oded, Data Mining With Decision Tree Theory and Applications 2nd Edition, World Scientific Publishing Co. Pte. Ltd., 2015.

[11] S. Lemeshow and D. W. Hosmer, Applied Logistic Regression, New York: John Wiley & Sons, 2000.

[12] M. Story and R. G. Congalton, "Accuracy assessment: a user's perspective," *Photogramm. Eng. Remote Sens.*, Vols. 52, no. 3, p. 397–399, 1986.

[13] F. Gorunescu, "Data Mining: Concepts, models and techniques," *Springer Science & Business Media*, vol. 12, 2011.

[14] X. Huo, X. S. Ni, A. K. Smith, T. . W. Liao and T. , "Recent advances in data mining of enterprise data," in *A survey of manifold-based learning methods*, 2007, pp. 691-745.

[15] I. H. Witten, F. Eibe and A. H. Mark, Data Mining: Practical Machine Learning Tools and Techniques 3rd edition, Burlington: Elsevier, 2011.

[16] . A. . A. Freitas, Data Mining and Knowledge Discovery with Evolutionary Algorithms, Verlag Berlin Heidelberg: Springer, 2002.

[17] P. Refaeilzadeh, L. Tang and H. Liu, "Cross-validation," in *Encyclopedia of database systems*, Springer, 2009, p. 532–538.

[18] D. R, Istilah-istilah Dunia Pariwisata, Jakarta: Pradnya Paramita, 1995.