PERBANDINGAN ALGORITMA DECISION TREE, RANDOM FOREST DAN NAIVE BAYES PADA PREDIKSI PENILAIAN KEPUASAN PENUMPANG MASKAPAI PESAWAT MENGGUNAKAN DATASET KAGGLE

Wida Azis Rahmat[1], Siti Madinah Ladjamuddin[2], Dipa Teruna Awaludin[3]
Program Studi Teknik Informatika, Fakultas Sains dan
Teknologi Informasi Institut Sains dan Teknologi Nasional
Jl. Moh Kahfi II, Bhumi Srengseng Indah, Jagakarsa, Jakarta
Selatan 12640 Widaazisr26@gmail.com[1],
citymadinah07@istn.ac.id[2], dipateruna@civitas.unas.ac.id[3]

ABSTRAK

Machine learning berfokus pada pembangunan sistem untuk mempelajari dan meningkatkan kinerja berdasarkan data yang dimiliki. Setiap algoritma *machine learning* memiliki performa yang berbeda, dalam penelitian ini berfokus dalam mengukur performa tiga algoritma *machine learning* klasifikasi yaitu algoritma *decision tree*, algoritma *random forest* dan algoritma *naive bayes.* Menggunakan data kepuasan penumpang pesawat dari situs *kaggle,* pada penelitian ini akan dilakukan klasifikasi untuk memprediksi penilaian kepuasan penumpang, Metode *confusion matrix* digunakan dalam mengukur performa akurasi, pengukuran menghasilkan algoritma *random forest* memiliki akurasi paling tinggi sebesar 95%, algoritma *decision tree* sebesar 93% dan algoritma *naive bayes* memiliki akurasi paling rendah sebesar 82%.
Kata Kunci: *Machine Learning, Confusion Matrix, Decision Tree, Random Forest, Naive Bayes*

ABSTRACT

Machine learning focuses on building systems to learn and improve performance based on the data they have. Each machine learning algorithm has a different performance. In this study, the focus is on measuring the performance of three classification machine learning algorithms, namely the decision tree algorithm, the random forest algorithm and the naive Bayes algorithm. Using airplane passenger satisfaction data from the kaggle site, in this study a classification will be carried out to predict passenger satisfaction ratings. The confusion matrix method is used to measure accuracy performance. The measurement results in a random forest algorithm having the highest accuracy of 95%, a decision tree algorithm of 93% and naive bayes algorithm has the lowest accuracy of 82%.
***Keywords**: Machine Learning, Confusion Matrix, Decision Tree, Random Forest, Naive Bayes*

I. PENDAHULUAN

Dalam industri pelayanan, kepuasan konsumen merupakan atribut paling penting dari sebuah perusahaan penyedia layanan jasa, Kepuasan konsumen adalah perasaan senang atau kecewa seseorang yang muncul setelah membandingkan kinerja atau hasil produk yang dipikirkan terhadap kinerja yang diharapkan (Indrasari, 2019). Pada era digital saat ini, sudah banyak perusahaan yang menyadari pentingnya mengukur penilaian kepuasan konsumen terhadap perusahaan mereka. Salah satu tindakan yang dilakukan perusahaan dalam mengukur kepuasan konsumen dengan melakukan analisis menggunakan pemanfaatan teknologi komputasi. *Machine Learning* merupakan bidang studi yang fokus kepada desain dan analisis algoritma sehingga memungkinkan komputer untuk dapat belajar (Ibnu Daqiqil Id, 2021).

Machine learning juga dapat diartikan sebuah komputer yang memiliki kemampuan belajar tanpa diprogram secara eksplisit. Program tersebut memanfaatkan data untuk membangun model dan mengambil keputusan berdasarkan model yang telah dibangun. ML berisi sebuah algoritma yang bersifat generic atau umum dimana algoritma tersebut dapat menghasilkan sesuatu yang menarik atau bermanfaat dari sejumlah data tanpa harus menulis kode yang spesifik (Ibnu Daqiqil Id, 2021). Memilih algoritma menjadi perhatian tersendiri, karena pada setiap algoritma memiliki tingkat performa yang berbeda, memilih suatu algoritma tidak dapat dilakukan secara acak karena hasil yang dikeluarkan mungkin tidak akan relevan.

Penelitian terdahulu oleh (Wijayanto et al., 2021) melakukan penelitian yang berjudul "Analisis Klasifikasi Kepuasan Penumpang Maskapai Penerbangan Menggunakan Algoritma *Naïve*

Bayes". Digunakan dataset sebanyak 129.880 record untuk selanjutnya dari dataset tersebut akan dibagi menjadi data *training* dan data *testing*, dimana pembagian data tersebut akan dibuat dalam 4 kondisi yaitu 90%, 85%, 80% dan 75% sehingga dapat dilihat dari keempat pembagian ini manakah yang memiliki tingkat akurasi yang paling tinggi.

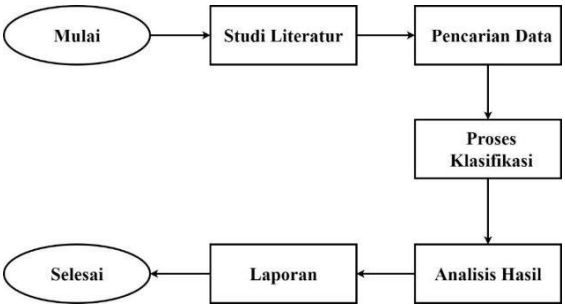
Penelitian terdahulu oleh (Sotarjua & Santoso, 2022) melakukan penelitian yang berjudul "Perbandingan Algoritma *KNN*, *Decision Tree* dan *Random Forest* Pada Data *Imbalanced Class* Untuk Klasifikasi Promosi Karyawan". Pada penelitian ini dilakukan proses klasifikasi dengan algoritma *machine learning* untuk menganalisa performa model pada data klasifikasi karyawan, data yang digunakan pada penelitian ini merupakan data *imbalanced class* sehingga dilakukan teknik *Synthetic Minority Over-Sampling Technique (SMOTE).*

Pada penelitian ini bertujuan untuk mengetahui algoritma yang relevan untuk digunakan dalam melakukan prediksi pada penilaian kepuasan penumpang. Menggunakan bahasa pemrograman *python* karena memiliki keunggungulan seperti fleksibilitas, readability, efisiensi, multifungsi, dan interoperabilitas. Dataset yang di analisis yaitu *Airline Customer Satisfaction* yang didapatkan dari situs *www.kaggle.com*, merupakan hasil data survei tentang kepuasan penumpang pesawat, dataset ini digunakan untuk mengetahui performa algoritma *decision tree*, *random forest*, dan *naive bayes*. Dalam proses implementasi peneliti menambahkan *exploratory data analysis* dan *data cleaning* pada tahapan penelitian, tujuannya untuk menghilangkan noise pada data sehingga saat proses implementasi model algoritma performa yang dihasilkan pada model algoritma dapat lebih baik. *Confusion matrix* merupakan metode yang digunakan untuk mengukur performa model algoritma pada pengujian ini. Parameter pengukuran yang akan digunakan pada penelitian ini menggunakan salah satu matriks dari *confusion matrix* yaitu *accuracy*. (Saputro & Sari, 2020).

II. METODOLOGI PENELITIAN

2.1. Tahapan Penelitian

Agar tercipta nya penelitian yang terstruktur maka peneliti membuat sistem alur kerja, berikut merupakan alur kerja yang akan digunakan pada penelitian ini.



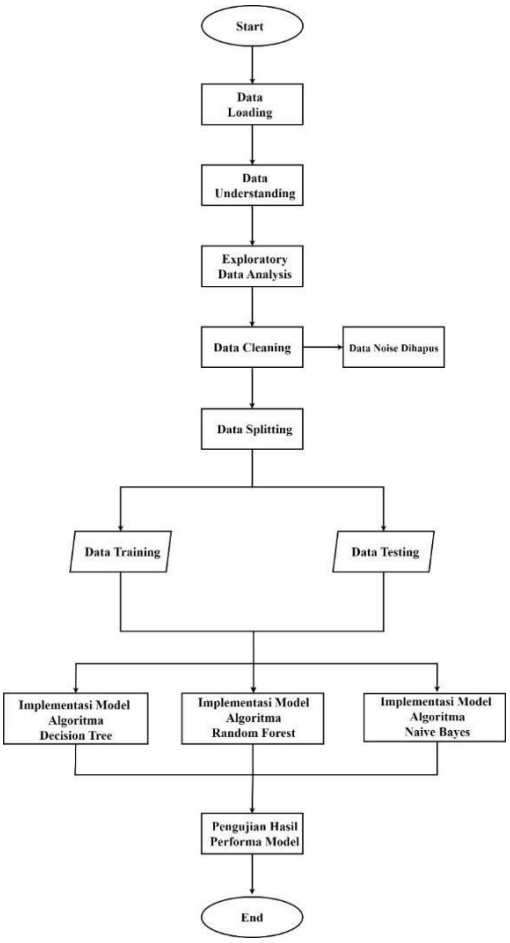
Gambar 2.1 Tahapan Penelitian

a. Studi Literatur

Ini merupakan tahapan awal penelitian yang digunakan dalam memahami topik yang akan dibahas. Pada tahap ini peneliti mengumpulkan referensi atau landasan teori yang sudah ada pada penelitian terdahulu untuk dijadikan landasan dalam menyelesaikan masalah yang ada. Dalam upaya menjawab masalah penelitian atau tujuan penelitian, peneliti menyimpulkan secara eksplisit sesuai dengan pengumpulan teori referensi yang mendasarinya.

b. Pencarian Data

Tahap ini merupakan tahapan dalam mencari dataset yang akan digunakan dalam penelitian ini. Dataset yang digunakan pada penelitian ini merupakan data penilaian kepuasan penumpang maskapai pesawat yang di upload oleh sayantan jana pada tahun 2020 didapatkan dari situs www.kaggle.com. Kaggle merupakan sebuah situs yang menyediakan dataset untuk diolah menggunakan machine learning. Lisensi pada dataset ini pada umum nya merupakan publik domain, yang mana pada lisensi ini memiliki peraturan yang paling bebas bila dibandingkan lisensi lainnya, publik domain merupakan lisensi yang dapat digunakan khalayak umum karena pada lisensi ini sebelumnya memiliki hak cipta namun masa berlakunya sudah habisDalam proses klasifikasi digunakan untuk mengolah data sehingga menghasilkan prediksi terhadap penilaian kepuasan penumpang pesawat sesuai dengan studi kasus yang telah ditetapkan di awal. Pada tahap ini menggunakan tiga algoritma yaitu algoritma *decision tree*, algoritma *random forest*, dan algoritma *naive bayes.*



Gambar 2.2 Tahapan Proses Klasifikasi

Sumber : (Wijayanto et al., 2021)

- Proses *Data Loading*
  Data yang digunakan pada penelitian ini berasal dari situs Bernama kaggle, data ini memiliki tipe file CSV (Comma Separated Value) dengan total baris data 129.880 dan jumlah feature sebanyak 23, dimana terdapat 1 feature yang akan dijadikan label atau target. Sebelum melakukan proses klasifikasi, hal pertama yang dilakukan adalah memasukan data kedalam baris kode jupyter notebook, hal ini dilakukan agar data dapat pahami dan diproses lebih lanjut. Proses ini memanfaatkan menggunakan perintah kode pd.read_csv(). Dalam proses data loading ini penting untuk memastikan bahwa data yang dimasukan kedalam baris kode merupakan data yang ingin di analisis. Selain memasukan file kedalam baris kode proses ini juga digunakan untuk melihat gambaran singkat tentang data kepuasan penumpang maskapai pesawat.
- Proses *Data Understanding*
  Dalam tahapan ini digunakan peneliti intuk mempelajari pola yang terdapat pada. Untuk mengumpulkan informasi yang terdapat pada data, dalam hal ini mengunakan baris kode *describe*() untuk mengetahui nilai yang ada pada *feature*. Peneliti mempelajari terkait nilai pada data tersebut.

Tabel 2.1 Deskripsi Nilai Data Pada Kolom

Nama Variabel	Rata-rata	Min	Max
Age	39.427957	7	85
Flight Distance	1981.409055	50	6951
Seat Comfort	2.838597	0	5
Depature/Arrival Time convenient	2.990645	0	5
Food and Drink	2.851994	0	5
Gate Location	2.990422	0	5
Inflight wifi Service	3.249130	0	5
Inflight Entertainment	3.383477	0	5
Online Support	3.519703	0	5
Ease of online Booking	3.472105	0	5
On board service	3.465075	0	5
Leg room service	3.485902	0	5
Baggage Handling	3.695673	1	5
Checking Service	3.340807	0	5
Cleanliness	3.705759	0	5
Online Boarding	3.352587	0	5
Departure delay In Minutes	14.713713	0	1592
Arrival delay In Minutes	15.091129	0	1584

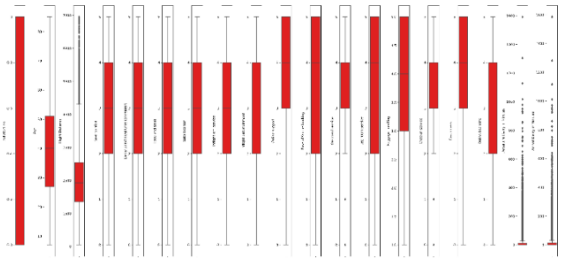
- Proses *Exploratory Data Analysis*
  Exploratory Data Analysis digunakan untuk mengeksplorasi data dalam bentuk grafik agar mempermudah peneliti dalam menganalisis dan menemukan pola-pola tersembunyi dalam data. Kondisi target label juga menjadi fokus dalam proses ini, dimana akan melihat hubungan antara *feature* terhadap target label. Dalam proses ini peneliti menggunakan tiga teknik analisis

untuk mengeksplorasi kondisi data, yaitu teknik statistika deskriptif, analisis univariat dan analisis multivariat.

```
Data columns (total 23 columns):
 #   Column                            Non-Null Count   Dtype
---  ------                            --------------   -----
 0   satisfaction                      129880 non-null  object
 1   Gender                            129880 non-null  object
 2   Customer Type                     129880 non-null  object
 3   Age                               129880 non-null  int64
 4   Type of Travel                    129880 non-null  object
 5   Class                             129880 non-null  object
 6   Flight Distance                   129880 non-null  int64
 7   Seat comfort                      129880 non-null  int64
 8   Departure/Arrival time convenient 129880 non-null  int64
 9   Food and drink                    129880 non-null  int64
 10  Gate location                     129880 non-null  int64
 11  Inflight wifi service             129880 non-null  int64
 12  Inflight entertainment            129880 non-null  int64
 13  Online support                    129880 non-null  int64
 14  Ease of Online booking            129880 non-null  int64
 15  On-board service                  129880 non-null  int64
 16  Leg room service                  129880 non-null  int64
 17  Baggage handling                  129880 non-null  int64
 18  Checkin service                   129880 non-null  int64
 19  Cleanliness                       129880 non-null  int64
 20  Online boarding                   129880 non-null  int64
 21  Departure Delay in Minutes        129880 non-null  int64
 22  Arrival Delay in Minutes          129487 non-null  float64
dtypes: float64(1), int64(17), object(5)
```

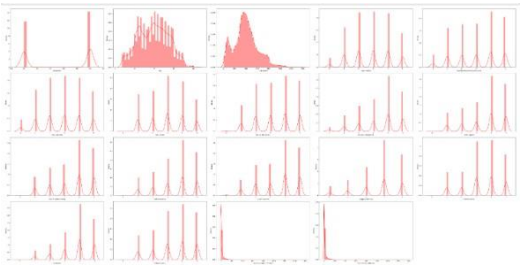
Gambar 2.3 Info Mengenai Dataset Kepuasan

Menggunakan statistika deskriptif peneliti melakukan pengecekan terhadap jumlah kolom, tipe data dan jumlah data. Menggunakan fungsi baris kode *.info()* untuk mendeskripsikan data, menemukan bahwa dataframe lengkap memiliki jumlah 129.880 data dan 22 kolom. Namun dalam pengecekan lebih lanjut ditemukan bahwa terdapat data null pada *feature Arival Delay in Minutes* seperti pada Gambar 2.3.



Gambar 2.4 Hasil Distribusi Data menggunakan *boxplot*

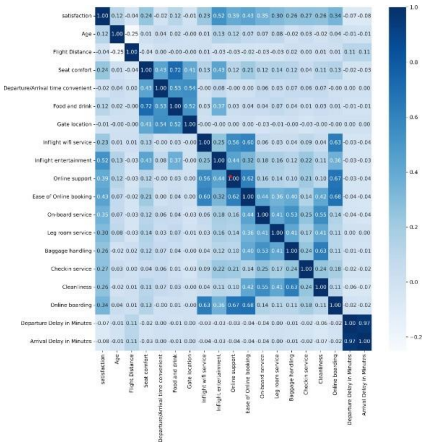
Analisis univariat digunakan untuk melihat distribusi pada kolom data yang berisi nilai numerik, dimana dalam prosesnya peneliti menggunakan baris kode boxplot dan distplot untuk menampilkan hasil dari distribusi data. Dalam pengamatan menggunakan visualisasi menggunakan *boxplot*, terdapat outlier terutama pada kolom *flight distance*, *departure delay in minutes* dan *arrival delay in minutes*. *Outlier* ini merupakan angka yang memiliki nilai yang sangat berbeda jauh dengan nilai dari angka lainya pada data, seperti Gambar 2.4.



Gambar 2.5 Hasil Distribusi Data menggunakan *displot*

Dari visualisasi menggunakan *distplot* terlihat bahwa pada kolom departure delay in minute dan arrival delay in minute terlihat tidak simestris dalam penyebaran data nya, seperti yang terlihat pada Gambar 2.5.

Pada penelitian ini digunakan untuk melihat hubungan antar variabel, dimana peneliti menggunakan analisis multivariat untuk melihat hubungan antar variabel pada feature kolom yang berisi nilai numerik. Pada analisis ini memanfaatkan fitur heatmap untuk visualisasi gambar. Dihasilkan bahwa target label satisfaction memiliki hubungan positif dengan kolom *age, seat comfort, departure/arrival time convient, food and drink, gate location, inflight wifi service, inflight entertainment, online support, ease of online booking, on-board service, leg room service, baggage handling, checkin service, cleanliness,* dan *online boarding*. Sedangkan target label memiliki hubungan negatif dengan kolom *flight distance, departure delay in minutes* dan *arrival delay in minutes.* Pada kolom *departure delay in minutes* memiliki hubungan positif yang kuat dengan *arrival delay in minutes*, ada kemungkinan bahwa dua *feature* ini redundant seperti Gambar 2.6.



Gambar 2.6 Hubungan antar kolom *feature*

Pada hasil *exploratory data analysis* dihasilkan bahwa, data kepuasan penumpang merupakan data valid yang tidak memiliki kecacatan, namun terdapat data kosong dan ada beberapa distribusi data yang tidak simetris, ini perlu di proses ketika *data cleaning*.

- Proses *Data Cleaning*
  Dalam proses exploratory data analysis menggunakan statistik deskritif ditemukan bahwa ada data yang kosong, dalam proses ini peneliti mengecek jumlah baris data yang hilang tersebut dan juga apakah ada data yang duplikat menggunakan fungsi baris kode isnull().sum() dan duplicated().sum(). Seperti yang terlihat pada Gambar 3.13, ditemukan bahwa ada 393 data yang hilang pada feature arrival delay in minutes. Sedangkan tidak ditemukan data yang duplikat. Peneliti menghilangkan *data null* dengan fungsi baris kode *dropna(inplace=True).*

```
satisfaction                          0
Gender                                0
Customer Type                         0
Age                                   0
Type of Travel                        0
Class                                 0
Flight Distance                       0
Seat comfort                          0
Departure/Arrival time convenient     0
Food and drink                        0
Gate location                         0
Inflight wifi service                 0
Inflight entertainment                0
Online support                        0
Ease of Online booking                0
On-board service                      0
Leg room service                      0
Baggage handling                      0
Checkin service                       0
Cleanliness                           0
Online boarding                       0
Departure Delay in Minutes            0
Arrival Delay in Minutes            393
dtype: int64
```

Gambar 2.7 Jumlah *Missing Value*

Dalam proses *exploratory data analysis* menggunakan analisis univariat ditemukan bahwa pada *feature flight distance*, *departure delay in minutes* dan *arrival delay in minutes* memiliki data yang *outlier*, pada proses ini digunakan untuk menghapus data *outlier* tersebut dengan menggunakan fungsi *method interquartile range (IQR). IQR* sendiri didapatkan dengan rumus IQR = *Q3 -Q1*. Dengan menggunakan metode IQR, peneliti dapat mengetahui *outlier* pada *feature flight distance*, *departure delay in minutes* dan *arrival delay in minutes* melalui suatu nilai batas yang ditentukan dengan rumusan batas bawah = *Q1 – (1.5 * IQR)* dan batas atas = *Q3 + (1.5 * IQR)*. Sebelum melakukan *filtering* terhadap *outlier* jumlah data adalah 129.487 dan sesudah melakukan filtering data terdahap outlier jumlah data menjadi 106.922.

peneliti melakukan standarisasi terhadap skala nilai data pada *feature flight distance, departure delay in minutes* dan *arrival delay in minutes*. Fungsi *normalization and standardization* digunakan untuk menurunkan nilai pada *feature* tersebut menjadi antara angka 0 hingga angka 1, *feature* yang berpengaruh dan mengesampingkan *feature* yang tidak berpengaruh dalam penelitian ini, pada *feature gender, customer type, type of travel* dan *class* dilakukan *one-hot encoding* untuk membuat *feature* baru dari nilai variabel kategorik tersebut menggunakan fungsi baris kode *get.dummies()*. Pada *feature arrival delay in minutes* dihilangkan dari *dataframe* karena pada *feature* tersebut memiliki korelasi yang sangat tinggi dengan *feature departure delay in minutes.*

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 106922 entries, 0 to 129875
Data columns (total 23 columns):
 #   Column                            Non-Null Count   Dtype
---  ------                            --------------   -----
 0   satisfaction                      106922 non-null  int64
 1   Gender                            106922 non-null  object
 2   Customer Type                     106922 non-null  object
 3   Age                               106922 non-null  int64
 4   Type of Travel                    106922 non-null  object
 5   Class                             106922 non-null  object
 6   Flight Distance                   106922 non-null  float64
 7   Seat comfort                      106922 non-null  int64
 8   Departure/Arrival time convenient 106922 non-null  int64
 9   Food and drink                    106922 non-null  int64
 10  Gate location                     106922 non-null  int64
 11  Inflight wifi service             106922 non-null  int64
 12  Inflight entertainment            106922 non-null  int64
 13  Online support                    106922 non-null  int64
 14  Ease of Online booking            106922 non-null  int64
 15  On-board service                  106922 non-null  int64
 16  Leg room service                  106922 non-null  int64
 17  Baggage handling                  106922 non-null  int64
 18  Checkin service                   106922 non-null  int64
 19  Cleanliness                       106922 non-null  int64
 20  Online boarding                   106922 non-null  int64
 21  Departure Delay in Minutes        106922 non-null  float64
 22  Arrival Delay in Minutes          106922 non-null  float64
dtypes: float64(3), int64(16), object(4)
memory usage: 19.6+ MB
```

Gambar 2.8 Sebelum Proses *Feature Selection*

Peneliti juga menghapus *feature gender, customer type, class* dan *type of travel.* Ini dilakukan karena pada proses *one-hot encoding feature* nilai kategori pada *feature* ini sudah dibagi menjadi beberapa *feature* baru. Hasil sebelum dan sesudah proses *feature selection* dapat dilihat pada Gambar 2.8 dan Gambar 2.9.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 106922 entries, 0 to 129875
Data columns (total 27 columns):
 #   Column                             Non-Null Count   Dtype
---  ------                             --------------   -----
 0   satisfaction                       106922 non-null  int64
 1   Age                                106922 non-null  int64
 2   Flight Distance                    106922 non-null  float64
 3   Seat comfort                       106922 non-null  int64
 4   Departure/Arrival time convenient  106922 non-null  int64
 5   Food and drink                     106922 non-null  int64
 6   Gate location                      106922 non-null  int64
 7   Inflight wifi service              106922 non-null  int64
 8   Inflight entertainment             106922 non-null  int64
 9   Online support                     106922 non-null  int64
 10  Ease of Online booking             106922 non-null  int64
 11  On-board service                   106922 non-null  int64
 12  Leg room service                   106922 non-null  int64
 13  Baggage handling                   106922 non-null  int64
 14  Checkin service                    106922 non-null  int64
 15  Cleanliness                        106922 non-null  int64
 16  Online boarding                    106922 non-null  int64
 17  Departure Delay in Minutes         106922 non-null  float64
 18  Gender_Female                      106922 non-null  uint8
 19  Gender_Male                        106922 non-null  uint8
 20  Customer Type_Loyal Customer       106922 non-null  uint8
 21  Customer Type_disloyal Customer    106922 non-null  uint8
 22  Type of Travel_Business travel     106922 non-null  uint8
 23  Type of Travel_Personal Travel     106922 non-null  uint8
 24  Class_Business                     106922 non-null  uint8
 25  Class_Eco                          106922 non-null  uint8
 26  Class_Eco Plus                     106922 non-null  uint8
dtypes: float64(2), int64(16), uint8(9)
memory usage: 20.4 MB
```

Gambar 2.9 Sesudah Proses *Feature Selection*

- Proses Data Splitting
  Pada proses *data splitting*, data akan dibagi menjadi dua bagian dengan skala 70% untuk *data training dan 30%* untuk data testing, *Data training* digunakan untuk melatih algoritma dalam mencari model yang sesuai untuk prediksi penilaian kepuasan dan *data testing* digunakan untuk menguji performa algoritma yang didapatkan pada tahapan *training*. *Data splitting* ini dilakukan dengan memanfaatkan *library* dari *scikit-learn* yaitu *train_test_split.* Proporsi pembagian jumlah *data spitting* dapat dilihat pada Tabel 2.2.

Tabel 2.2 Proporsi *Data Splitting*

Data Splitting	**Jumlah Data**
Data Training	74.845
Data Testing	32.077

- Proses Implementasi Model
  Untuk menggunakan algoritma *decision tree* dilakukan dengan memanfaatkan *library dari scikit-learn* dengan fungsi baris kode *from sklearn.tree import DecisionTreeClassifier*. Dalam melakukan pelatihan model algoritma *decision tree* menggunakan baris kode *dtree =DecisionTreeClassifier(random_state=42) dtree=dtree.fit (Xtrain,ytrain)*.
  Untuk menggunakan algoritma *random forest* dilakukan dengan memanfaatkan *library dari scikit-learn* dengan fungsi baris kode *from sklearn.ensemble import RandomForest Classifier*. Dalam melakukan pelatihan model algoritma *random forest* menggunakan baris kode *rf = RandomForest Classifier (random_state=42)rf.fit(Xtrain,ytrain))*.
  Untuk menggunakan algoritma *naïve bayes* dilakukan dengan memanfaatkan *library dari scikit-learn* dengan fungsi baris kode *from sklearn.naive_bayes import GaussianNB.* Dalam melakukan pelatihan model algoritma *naïve bayes* menggunakan baris kode *nbc = GaussianNB()nbc.fit(Xtrain, ytrain).*

- Proses Pengujian Hasil Performa Model
  Pada tahap ini peneliti melakukan pengujian terhadap performa model yang dihasilkan dari algoritma *decision tree*, algoritma *random forest* dan algoritma *naive bayes* yang digunakan untuk melakukan prediksi terhadap penilaian kepuasan. Dalam melakukan pengujian peneliti menggunakan metode pengukuran *confusion matrix* dilakukan dengan memanfaatkan *library* dari *scikit-learn* untuk mendapatkan hasil *testing* dari performa algoritma. Dengan menggunakan baris kode *from sklearn.metrics import plot_confusion_matrix* menghasilkan tabel hasil pengujian nilai yang diprediksi secara akurat dan tidak akurat. Dari tabel hasil pengujian tersebut digunakan untuk menghitung nilai akurasi sesuai dengan rumus akurasi $\frac{TP+TN}{TP+TN+FP+FN}$.

c. Analisis Hasil
Nilai akurasi yang didapat dari proses pengujian hasil performa model sebelumnya akan di analisis. Hasil pada proses ini digunakan untuk membuat laporan dalam menjawab studi kasus yang telah ditentukan diawal penelitian.

d. Laporan
Pada tahap ini, peneliti merangkum hasil dari seluruh kegiatan penilitian, dari mulai hingga analisis hasil. Tujuan dibuat nya rangkuman ini sebagai syarat dalam menyelesaikan tugas akhir atau skripsi.

2.2. Instrumen Penelitian

Dalam melakukan penelitian, berikut merupakan perangkat keras maupun perangkat lunak yang digunakan :

a. Spesifikasi Perangkat Keras
- Processor 8th Gen Intel Core i5-8250U
- Memory 12 GB DDR4 (2133Mhz)
- Graphic Card Intel UHD Graphics 620, Nvidia GeForce GT 930MX 2 GB
- Storage HDD 1 TB dan SSD 512 GB

b. Spesifikasi Perangkat Lunak
- Operating System Windows 10 home
- PyCharm Community Edition 2020.3.3 x64
- Python version 3.10.2
- Microsoft Word 2019

III. HASIL DAN PEMBAHASAN

Proses serangkaian pembersihan data menghasilkan sebuah dataset yang siap olah untuk selanjutnya digunakan sebagai pengujian terhadap algoritma *decision tree*, *random forest* dan *naive bayes*. Seperti yang terdapat pada Tabel 3.1, dataset kepuasan penumpang maskapai pesawat memiliki total 26 *feature* dan memiliki tipe data numerik.

Tabel 3.1 Kategori Penilaian Dataset Kepuasan Penumpang

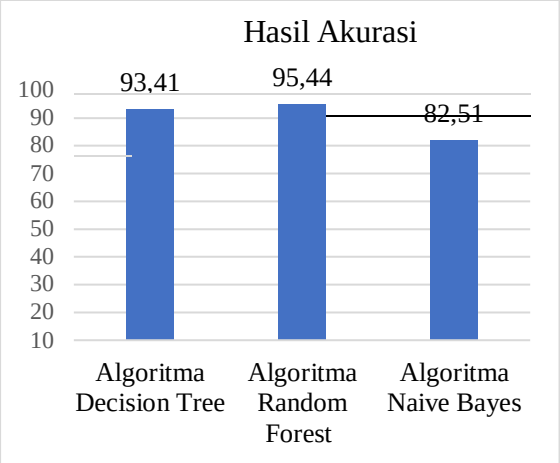
No	Variabel	Nilai Data
1	*Satisfaction*	0 & 1
2	*Gender_female*	0 & 1
3	*Gender_male*	0 & 1
4	*Customer Type_Loyal Customer*	0 & 1
5	*Customer Type_Disloyal Customer*	0 & 1
6	*Age*	0-80
7	*Type of Travel_Business Travel*	0 & 1
8	*Type of Travel_Personal Travel*	0 & 1
9	*Class_Business*	0 & 1
10	*Class_Business*	0 & 1
11	*Class_Eco*	0 & 1
12	*Class_Eco Plus*	0 & 1
13	*Flight Distance*	0 - 1
14	*Seat Comfort*	0 - 5
15	*Departure/Arrival*	0 - 5
16	*Food and Drink*	0 - 5
17	*Gate Location*	0 - 5
18	*Inflight Wifi Service*	0 - 5
19	*Inflight Entertaint*	0 - 5
20	*Online Support*	0 - 5
21	*Ease of Online Booking*	0 - 5
22	*On-board Service*	0 - 5
23	*Leg Room Service*	0 - 5
24	*Baggage Handling*	0 - 5
25	*Checkin Service*	0 - 5
26	*Cleanliness*	0 - 5
27	*Online Boarding*	0 - 5
28	*Departure Delay*	0 - 1

Setelah melakukan proses klasifikasi terhadap tiga algoritma yang di uji peneliti melakukan analisis hasil dari proses yang telah dilakukan sebelumnya, mulai dari proses *data loading*, proses *data understanding*, proses *exploratory data analysis*, proses *data cleaning*, proses *data splitting*, proses implementasi model dan proses pengujian hasil performa model. Setelah melakukan serangkaian proses di atas, ketiga algoritma menghasilkan nilai dari proses pengujian menggunakan *confusion matrix*, matriks tersebut digunakan untuk bahan analisis terhadap performa algoritma.

Tabel 3.2 Hasil Pengujian Pada Tabel Confusion Matrix

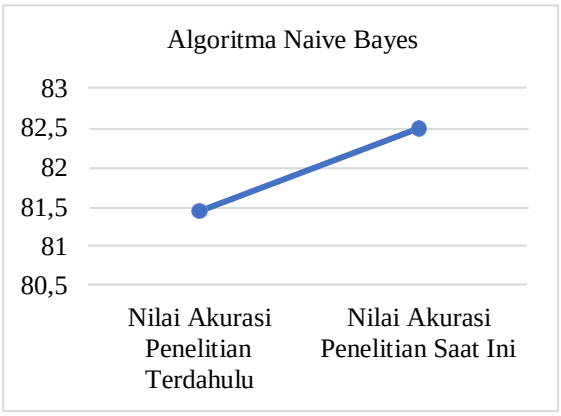
Nilai Aktual	Algoritma Decision Tree	
	Prediksi Penilaian Tidak Puas	Prediksi Penilaian Puas
Penilaian Tidak Puas	13.015	1.107
Penilaian Puas	1.006	16.949
Nilai Aktual	Algoritma Random Forest	
	Prediksi Penilaian Tidak Puas	Prediksi Penilaian Puas
Penilaian Tidak Puas	13.486	636
Penilaian Puas	825	17.130
Nilai Aktual	Algoritma Naive Bayes	
	Prediksi Penilaian Tidak Puas	Prediksi Penilaian Puas
Penilaian Tidak Puas	11.539	2.583
Penilaian Puas	3.027	14.928

Hasil akurasi seperti pada Gambar 3.1. Peneliti melakukan analisis dari hasil pengujian terhadap algoritma *decision tree* nilai akurasi yang dihasilkan sebesar 93.41%. Sedangkan pada algoritma *random forest* nilai akurasi yang dihasilkan merupakan tertinggi diantara algoritma lainnya sebesar 95.44%. Algoritma *naive bayes* memiliki nilai akurasi yang paling rendah sebesar 82.51%. Berdasarkan nilai akurasi yang dihasilkan setelah melewati tahap pengujian, algoritma *random forest* merupakan algoritma yang memiliki performa lebih baik dalam digunakan untuk melakukan prediksi terhadap penilaian kepuasan penumpang maskapai pesawat dibandingkan dengan algoritma *decision tree* dan algoritma *naive bayes.*



Gambar 3.1 Hasil Akurasi Algoritma Decision Tree, Random Forest dan Naive Bayes

Hasil pengujian yang didapat dalam peneilitan ini, peneliti menyimpulkan hasil perbandingan performa ketiga algoritma bahwa algoritma *random forest* memiliki kelebihan nilai akurasi yang didapat pada algoritma ini merupakan yang terbesar diantara dua algoritma pembandingnya. Algoritma *decision tree* memiliki kelebihan nilai akurasi yang dihasilkan lebih tinggi dari algoritma *naïve bayes*, kekurangan dari algoritma *decision tree* adalah nilai akurasi yang dihasilkan masih lebih rendah dari algoritma *random forest*. Algoritma *naïve bayes* memiliki kelebihan nilai akurasi yang dihasilkan masih dapat lebih baik seiring dengan pemodelan yang dilakukan, kekurangan dari algoritma ini nilai akurasi yang dihasilkan jauh lebih rendah dari algoritma *random forest* dan *decision tree*.



Gambar 3.2 Perbandingan Nilai Akurasi Dengan Penelitian Terdahulu

Berdasarkan perbandingan akurasi yang dihasilkan dengan penelitian terdahulu, proses *exploratory data analysis* dan *data cleaning* pada penelitian saat ini efektif dalam meminimalisir *data noise* sehingga performa yang dihasilkan pada model algoritma memiliki peningkatan akurasi. Hal ini dapat dilihat pada Gambar 3.2.

IV. PENUTUP

5.1 Kesimpulan

Dari hasil pengujian yang telah dijalankan dapat diambil beberapa kesimpulan sebagai berikut:

1. Model *machine learning* dengan algoritma *random forest* memiliki nilai akurasi paling tinggi sebesar 95,44%.
2. Model *machine learning* dengan algoritma *decision tree* memiliki nilai akurasi sebesar 93,41%.
3. Model *machine learning* dengan algoritma *naive bayes* memiliki nilai akurasi paling rendah sebesar 82,51%.
4. Nilai akurasi pada algoritma *random forest* dan algoritma *naive bayes* memiliki selisih yang cukup jauh, pada studi kasus serupa lebih disarankan menggunakan algoritma *random forest*.

5.2 Saran

Berdasarkan penelitian yang telah dilakukan terdapat saran untuk dijadikan referensi untuk penelitian selanjutnya, adapun saran sebagai berikut:

1. Dalam melakukan pengukuran terhadap pengujian performa algoritma hanya menggunakansatu metrik yaitu akurasi, dalam penelitian lebih lanjut disarankan agar menggunakan metode pengukuran lain untuk mengukur performa dari algoritma.
2. Dalam melakukan pengujian terhadap studi kasus serupa agar menggunakan algoritma lain seperti *K-Nearest Neighbor, Neural Networks* atau *Support Vector Machine.*
3. Menggunakan lebih banyak data agar akurasi yang dihasilkan lebih beragam.

REFERENSI

Aji Prasetya Wibawa, Muhammad Guntur Aji Purnama, Muhammad Fathony Akbar, F. A. D. (2018). Metode-metode Klasifikasi. *Prosiding Seminar Ilmu Komputer Dan Teknologi Informasi*, *3*(1), 134.

Alfarizi, M. R. sirfatullah, Al-farish, M. Z., Taufiqurrahman, M., Ardiansah, G., & Elgar, M. (2023). Penggunaan Python Sebagai Bahasa Pemrograman Untuk Machine Learning Dan Deep Learning. *Karimah Tauhid*, *2*(1), 1–6.

Dharma, A. S., & Tambunan, V. (2021). Penerapan Model Pembelajaran dengan Metode Reinforcement Learning Menggunakan Simulator Carla. *Jurnal Media Informatika Budidarma*, *5*(4), 1405. https://doi.org/10.30865/mib.v5i4.3169

Heliyanti Susana. (2022). Penerapan Model Klasifikasi Metode Naive Bayes Terhadap Penggunaan Akses Internet. *Jurnal Riset Sistem Informasi Dan Teknologi Informasi (JURSISTEKNI)*, *4*(1), 1–8. https://doi.org/10.52005/jursistekni.v4i1.96

Husnul Khatimi, Muhammad Alkaff, & Dewi Rizqia Najipah. (2017). Penerapan Support Vector Regression (Svr) Untuk Peramalan Inflasi Bulanan Nasional. *Jurnal Teknologi Informasi Universitas Lambung Mangkurat (JTIULM)*, *2*(2), 59–64. https://doi.org/10.20527/jtiulm.v2i2.21

Ibnu Daqiqil Id. (2021). *MACHINE LEARNING : Teori, Studi Kasus dan Implementasi Menggunakan Python - Ibnu Daqiqil Id - Google Buku*. *July*. https://doi.org/10.5281/zenodo.5113507

Indrasari, D. M. (2019). *Pemasaran dan Kepuasan Pelanggan* (Pertama). Unitomo Press. http://repository.unitomo.ac.id/2773/1/PEMASARAN DAN KEPUASAN PELANGGAN.pdf

Kesuma, M. E.-K., & Iskandar, R. (2022). Analisis Toko dan Asal Toko Fashion Pria di Shopee Menggunakan Data Scrapping dan Exploratory Data Analysis. *Majalah Ilmiah Teknologi Elektro*, *21*(1), 127. https://doi.org/10.24843/mite.2022.v21i01.p17

Kurnia, J. D., Retnaningsih, S. M., & Aridinanti, L. (2013). Analisis kapabilitas proses produksi monosodium glutamat (MSG) di PT. Ajinomoto Indonesia. *Jurnal Sains Dan Seni Pomits*, *2*(1), 2337–3520.

Lubis, A. I., Erdiansyah, U., & Siregar, R. (2022). Komparasi Akurasi pada Naive Bayes dan Random Forest dalam Klasifikasi Penyakit Liver. *Journal of Computing Engineering, System and Science (CESS)*, *7*(1), 81–89.

Maksum, A., & Swanjaya, D. (2021). Perbandingan Antara Metode Decision Tree Dan Support Vector Machine Pada Model Rekomendasi Mobil Bekas. *Prosiding SEMNAS INOTEK …*. https://proceeding.unpkediri.ac.id/index.php/inotek/article/view/1098

Martias, L. D. (2021). Statistika Deskriptif Sebagai Kumpulan Informasi. *Fihris: Jurnal Ilmu Perpustakaan Dan Informasi*, *16*(1), 40. https://doi.org/10.14421/fhrs.2021.161.40-59

Nanda, A. P., Pramono, D. E. H., & Hartati, S. (2020). Menentukan Tingkat Kepuasan Mahasiswa Terhadap Pelayanan Akademik Menggunakan Metode Algoritma K-Means. *Jurnal Sistem Informasi Dan Telematika*, *11*(1), 23–28.

Nasution, M. R. A., & Hayaty, M. (2019). Perbandingan Akurasi dan Waktu Proses Algoritma K-NN dan SVM dalam Analisis Sentimen Twitter. *Jurnal Informatika*, *6*(2), 226–235. https://doi.org/10.31311/ji.v6i2.5129

Novian, A. (2014). FAKTOR YANG BERHUBUNGAN DENGAN KEPATUHAN DIIT PASIEN HIPERTENSI (Studi Pada Pasien Rawat Jalan di Rumah SakitIslam Sultan Agung Semarang Tahun 2013). *Unnes Journal of Public Health*, *3*(3), 1–9.

Prasetyo, B., Suryani, V., & Anbiya, D. R. (2021). *Analisis Deteksi Malware pada Aplikasi Android Fintech berdasarkan Permissions dengan menggunakan Naive Bayes dan Random Forest*. *8*(5), 9885–9897.

Ramli, R. G., & Sibaroni, Y. (2022). *Klasifikasi Topik Twitter menggunakan Metode Random Forest dan Fitur Ekspansi*. *9*(1), 79–92.

Saputro, I. W., & Sari, B. W. (2020). Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa. *Creative Information Technology Journal*, *6*(1), 1. https://doi.org/10.24076/citec.2019v6i1.178

Sial, A. H., Yahya, S., & Rashdi, S. (2021). Comparative Analysis of Data Visualization Libraries Matplotlib and Seaborn in Python. *International Journal of Advanced Trends in Computer Science and Engineering*, *10*(1), 277–281. https://doi.org/10.30534/ijatcse/2021/391012021

SOTARJUA, L. M., & SANTOSO, D. B. (2022). Perbandingan Algoritma Knn, Decision Tree,* Dan Random* Forest Pada Data Imbalanced Class Untuk Klasifikasi Promosi Karyawan. *... Informatika Sains Dan ...*, *7*, 192–200. https://journal3.uin-alauddin.ac.id/index.php/instek/article/view/31385%0Ahttps://journal3.uin-alauddin.ac.id/index.php/instek/article/download/31385/15560

Syamsul, B., Dwi, M., & Rahmi, H. (2018). Perbandingan Algoritma Naive Bayes dan C4.5 Untuk Klasifikasi Penyakit Anak. *Seminar Nasional Aplikasi Teknologi Informasi (SNATi)*, B24–B31.

Widjaja, K., & Oetama, R. S. (2020). K-Means Clustering Video Trending di Youtube Amerika Serikat. *Ultima InfoSys : Jurnal Ilmu Sistem Informasi*, *11*(2), 78–84. https://doi.org/10.31937/si.v11i2.1508

Wijayanto, A., Bernardo, J. F. A., & Pamungkas, S. (2021). Analisis Klasifikasi Kepuasan Penumpang Maskapai Penerbangan Menggunakan Algoritma Naïve Bayes. *Jurnal Sains Komputer Dan Teknologi Informasi*, *3*(2), 97–103. https://doi.org/10.33084/jsakti.v3i2.2041