KELOMPOK 6

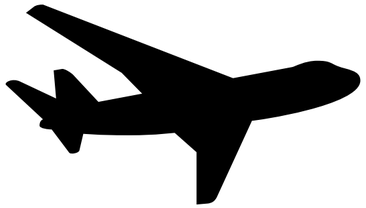
MACHINE LEARNING B

Perbandingan Kinerja Metode Klasifikasi dalam Kepuasan Penumpang Pesawat Terbang tahun 2019

Anggota Kelompok 6:

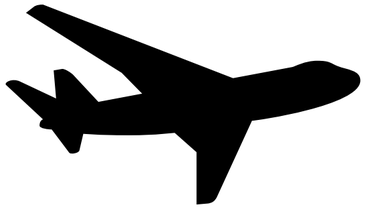
- Mohammad Nizar Riswanda 21083010015
- Meisya Vira Amalia 21083010018
- Rheinka Elyana Suprapto 21083010021
- Edina Alana Nabila 21083010022

LATAR BELAKANG



Dalam beberapa dekade terakhir, industri penerbangan mengalami pertumbuhan yang pesat dengan penumpang yang terus meningkat di seluruh dunia. Kecepatan, kenyamanan, dan jangkauan yang luas yang ditawarkan menjadi alasan mengapa banyak masyarakat yang memilih menggunakan transportasi udara. **Maskapai penerbangan tentu berlomba-lomba memberikan pengalaman terbaik kepada penumpang mereka agar mempertahankan pelanggan, memperluas pangsa pasar, dan membangun citra positif bagi maskapai penerbangan yang kuat**. Oleh karena itu, pada projek kali ini kelompok kami akan mengklasifikasikan kepuasan pelanggan pesawat terbang dengan menggunakan bantuan Machine Learning dan membandingkan beberapa metode untuk memperoleh metode terbaik.

PENELITIAN TERKAIT



Random Forest

Berdasarkan jurnal yang ditulis Herawan dkk. tahun 2021, metode Random Forest menjadi metode dengan akurasi paling baik **sebesar 99%** untuk klasifikasi kepuasan penumpang pesawat terbang dibanding metode lainnya.

Support Vector Machine

Berdasarkan jurnal yang ditulis Robbi dkk. tahun 2022, metode SVM dapat digunakan dalam mengklasifikasikan berita. Dengan hasil akurasi tertinggi didapatkan pada skenario pembagian data 90% dan 10% yaitu **sebesar 88%** dengan data yang digunakan sebanyak 510 data berita.

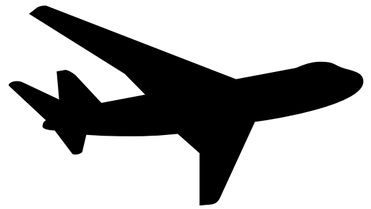
Extreme Gradient Boosting

Dalam penelitian yang ditulis Aditya dkk. tahun 2021, dibahas klasifikasi pemegang polis menggunakan metode XGBoost didapatkan bahwa metode ini dapat mengklasifikasikan klaim dengan **akurat sebesar 80,87%**. Sedangkan dalam hal presisi klasifikasi klaimnya sebesar 80,84%.

DATASET

Source: kaggle.com

129.880 BARIS
24 KOLOM

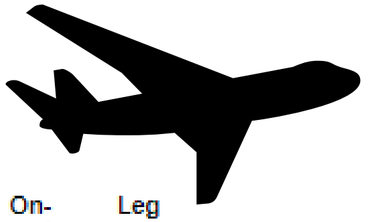


id		Inflight wifi service
satisfaction_v2	**VAR. TARGET**	Inflight entertainment
Gender		Online support
Customer Type		Ease of Online booking
Age		On-board service
Type of Travel		Leg room service
Class		Baggage handling
Flight Distance		Checkin service
Seat comfort		Cleanliness
Departure/Arrival time convenien		Online boarding
Food and drink		Departure Delay in Minutes
Gate location		Arrival Delay in Minutes

Kategori dalam Variabel Target
Satisfied
Neutral or disastified

DATASET

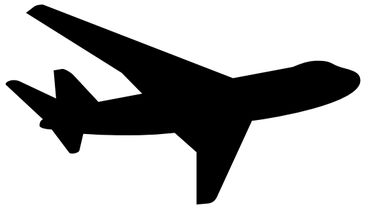
Source: kaggle.com

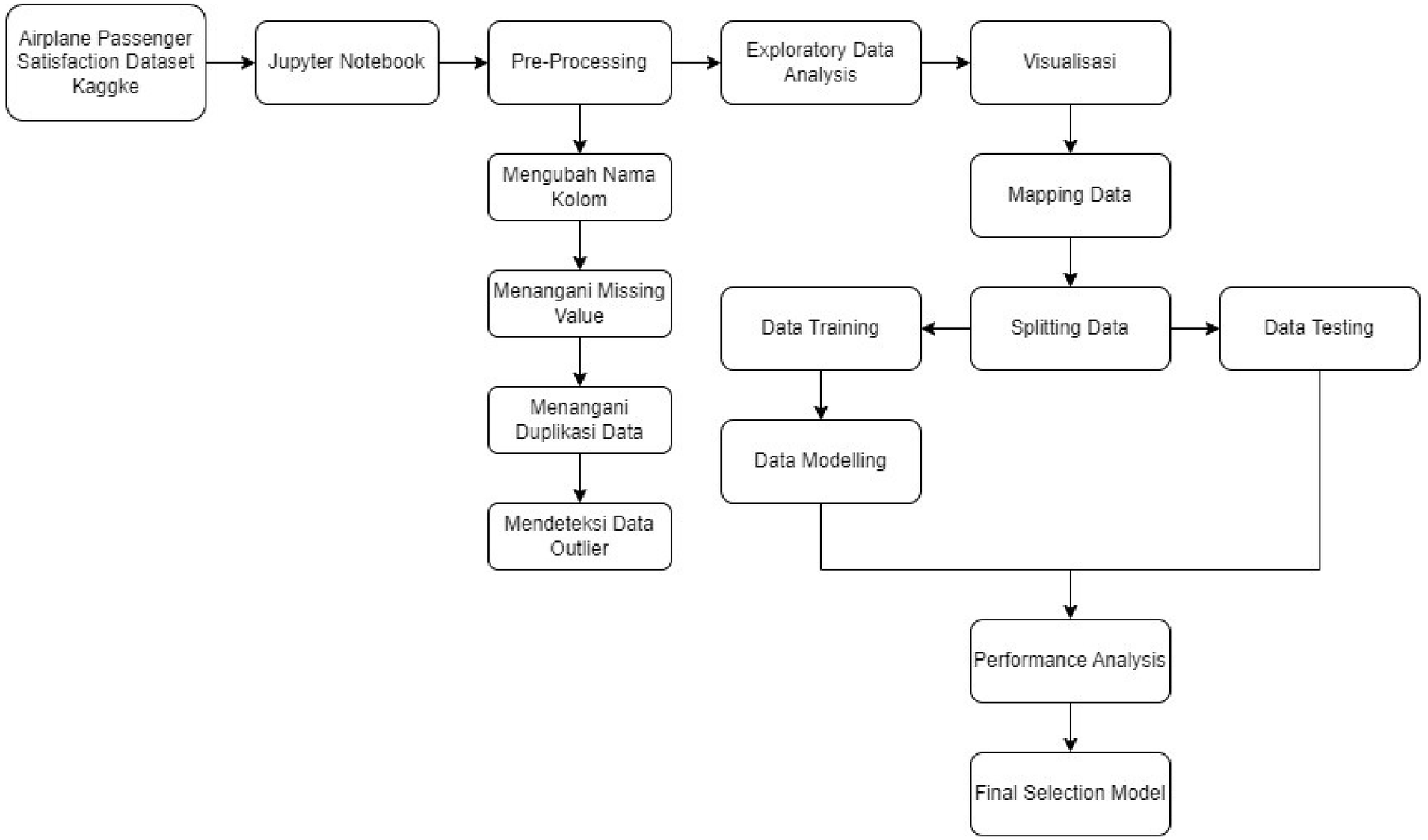


	id	satisfaction_v2	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Seat comfort	Departure/Arrival time convenient	...	Online support	Ease of Online booking	On-board service	Leg room service
0	11112	satisfied	Female	Loyal Customer	65	Personal Travel	Eco	265	0	0	...	2	3	3	0
1	110278	satisfied	Male	Loyal Customer	47	Personal Travel	Business	2464	0	0	...	2	3	4	4
2	103199	satisfied	Female	Loyal Customer	15	Personal Travel	Eco	2138	0	0	...	2	2	3	3
3	47462	satisfied	Female	Loyal Customer	60	Personal Travel	Eco	623	0	0	...	3	1	1	0
4	120011	satisfied	Female	Loyal Customer	70	Personal Travel	Eco	354	0	0	...	4	2	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
129875	119211	satisfied	Female	disloyal Customer	29	Personal Travel	Eco	1731	5	5	...	2	2	3	3
129876	97768	neutral or dissatisfied	Male	disloyal Customer	63	Personal Travel	Business	2087	2	3	...	1	3	2	3
129877	125368	neutral or dissatisfied	Male	disloyal Customer	69	Personal Travel	Eco	2320	3	0	...	2	4	4	3
129878	251	neutral or dissatisfied	Male	disloyal Customer	66	Personal Travel	Eco	2450	3	2	...	2	3	3	2
129879	84566	neutral or dissatisfied	Female	disloyal Customer	38	Personal Travel	Eco	4307	3	4	...	3	4	5	5

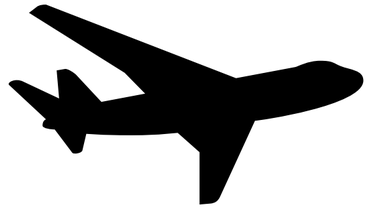
129880 rows × 24 columns

METODE PENELITIAN





METODE KLASIFIKASI



Random Forest

Random Forest adalah metode ensemble learning yang membangun beberapa pohon keputusan secara acak dan menggabungkan prediksi dari setiap pohon untuk mencapai hasil akhir yang lebih stabil dan akurat.

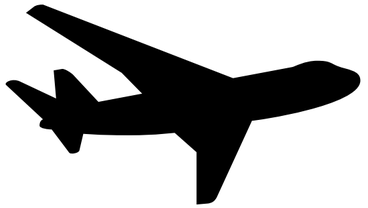
Formula Random Forest

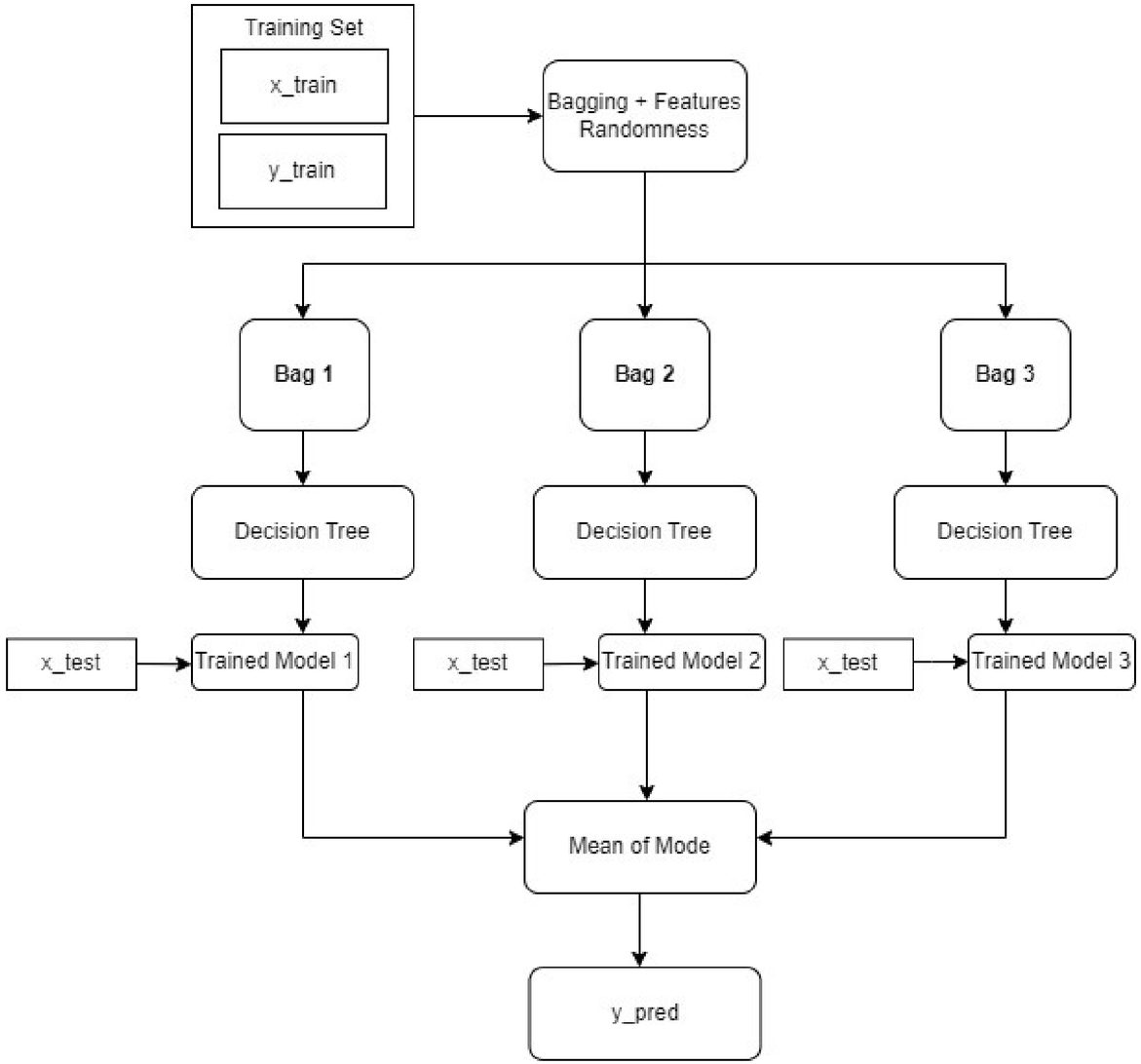
$$l(y) = argmax_c(\sum_{n=1}^{N} I_{h_n(y)=c})$$

Kelebihan Random Forest

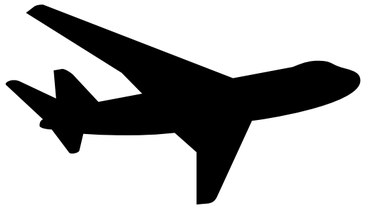
Kelebihan Random Forest adalah kemampuannya dalam mengatasi overfitting, dapat menghandle data yang memiliki banyak fitur, dan mampu mengidentifikasi fitur yang penting dalam prediksi.

RANDOM FOREST





METODE KLASIFIKASI



Extreme Gradient Boosting

XGBoost adalah algoritma gradient boosting yang menggunakan kombinasi dari beberapa pohon keputusan dengan memperhatikan bobot dan kesalahan sebelumnya untuk meningkatkan performa prediksi, sehingga menghasilkan model yang kuat dan akurat.

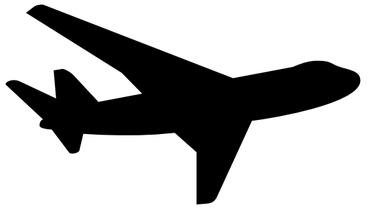
Formula XGBoost

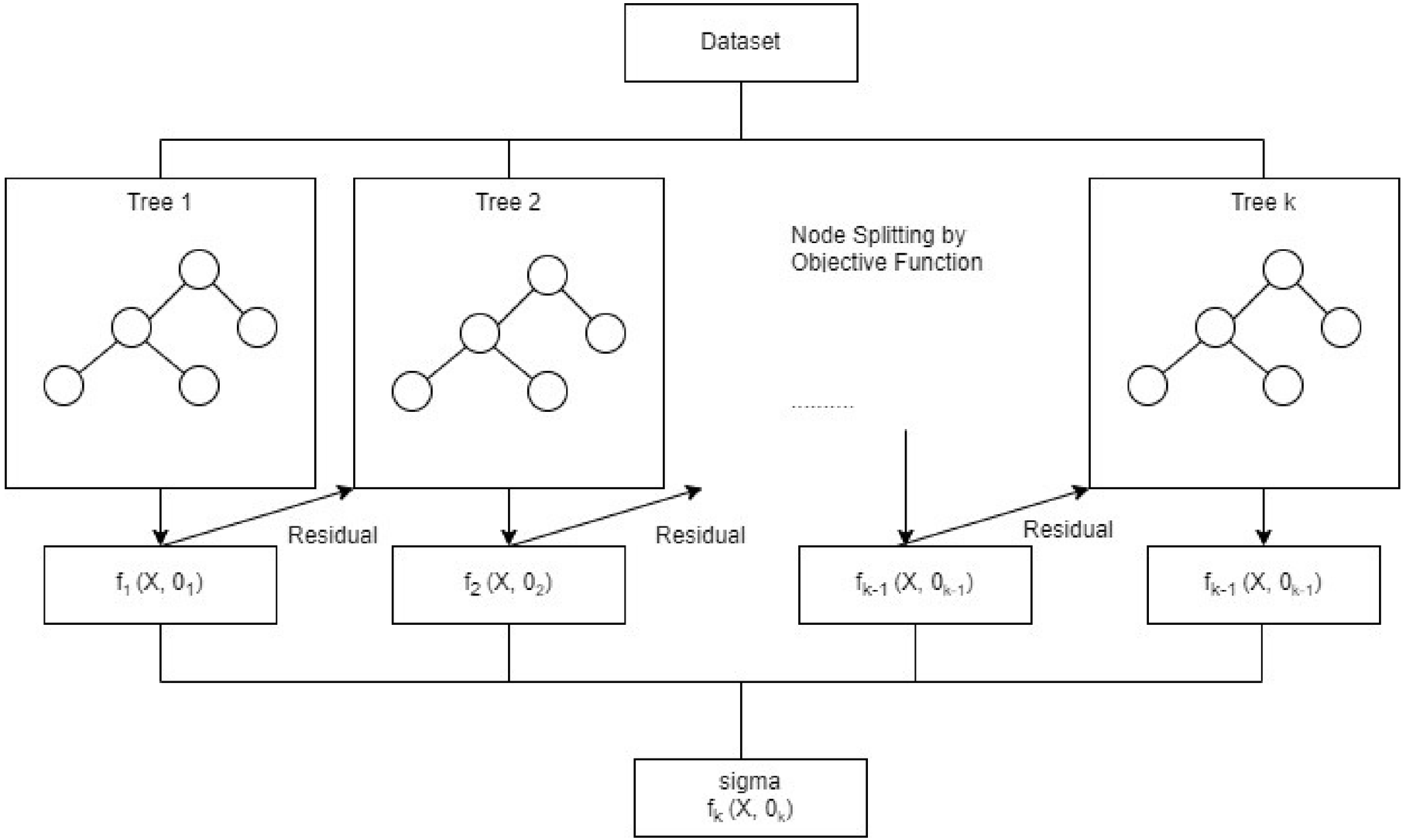
$$Obj^{(t)} = \sum_{i=1}^{t} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=i}^{t} \Omega(f_i)$$

Kelebihan XGBoost

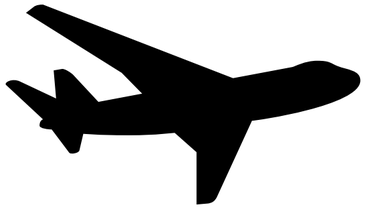
Kelebihan XGBoost adalah kemampuannya dalam mengatasi masalah overfitting, dapat menghandle fitur yang beragam dengan baik, serta memberikan performa prediksi yang cepat dan akurat dengan menggunakan teknik ensemble learning.

XGBOOST




Dataset
Tree 1
Tree 2
Node Splitting by
Objective Function
Tree k
Residual
Residual
Residual
$f_1(X, 0_1)$
$f_2(X, 0_2)$
$f_{k-1}(X, 0_{k-1})$
$f_{k-1}(X, 0_{k-1})$
sigma
$f_k(X, 0_k)$

METODE KLASIFIKASI



Support Vector Machine

SVM (Support Vector Machine) adalah algoritma yang mencari hyperplane terbaik yang memisahkan dua kelas data dengan margin terbesar, sehingga memungkinkan klasifikasi atau regresi yang optimal.

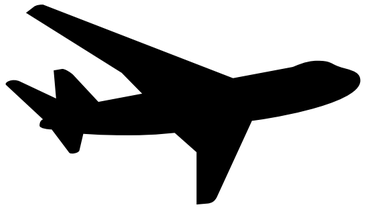
Formula Support Vector Machine

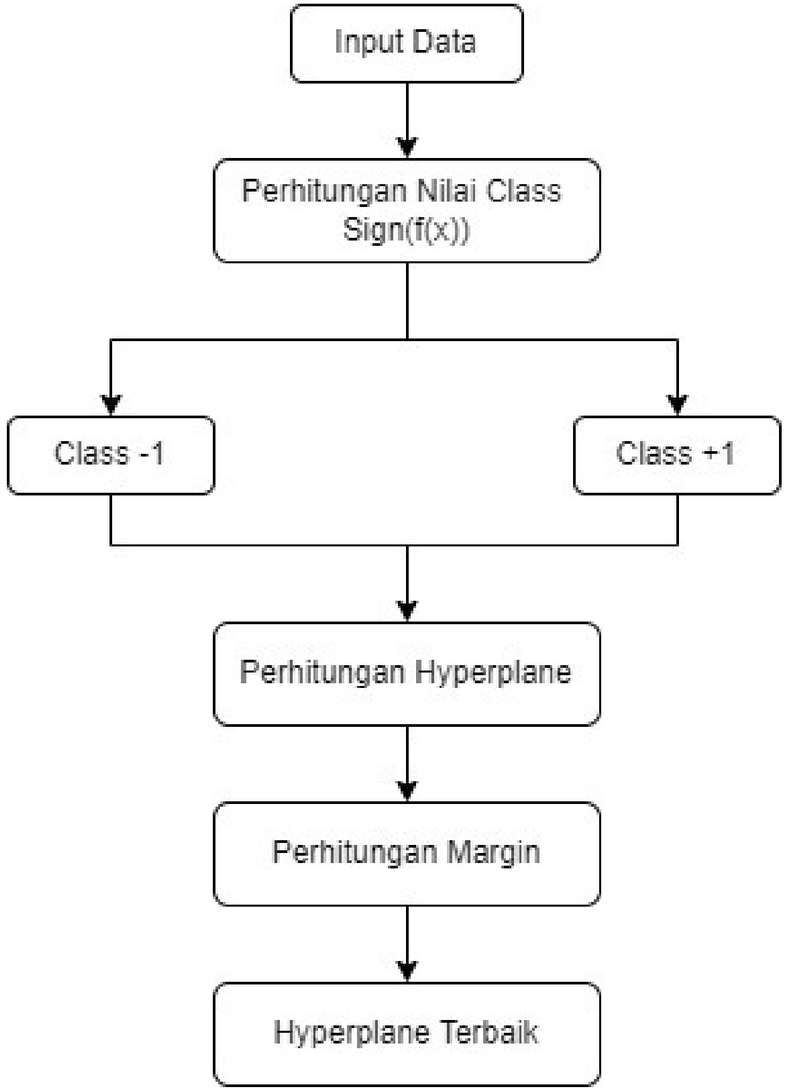
$$\alpha_i \geq 0 (i = 1,2, \dots, l) \qquad \sum_{i=1}^{l} \alpha_i y_i = 0$$

Kelebihan Support Vector Machine

Kelebihan SVM adalah kemampuannya dalam menangani data yang tidak linier dan memiliki dimensi tinggi, serta memiliki keakuratan yang tinggi dalam klasifikasi atau regresi pada dataset yang kompleks.

SVM




Input Data
Perhitungan Nilai Class
Sign(f(x))
Class -1
Class +1
Perhitungan Hyperplane
Perhitungan Margin
Hyperplane Terbaik

PRE-PROCESSING

Mengubah nama kolom

Menangani Missing Value

```
df.isna().sum().sort_values(ascending=False)
```

```
arrival_delay_minutes      393
```

```
df.shape
```

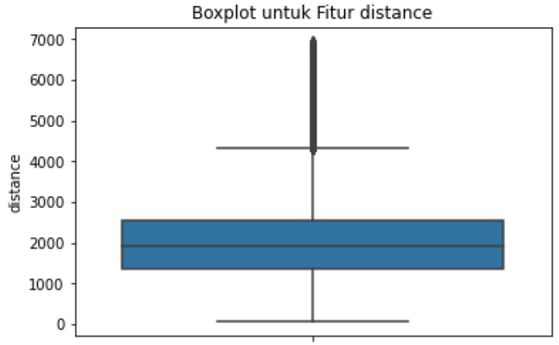
```
(129487, 23)
```

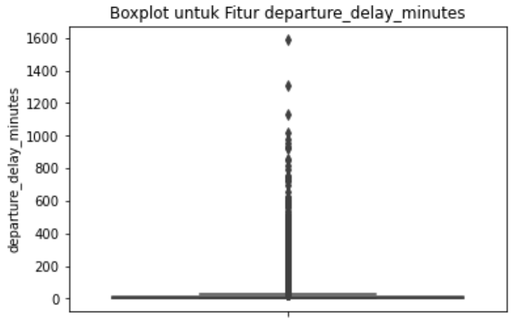
Menangani Duplikasi Data

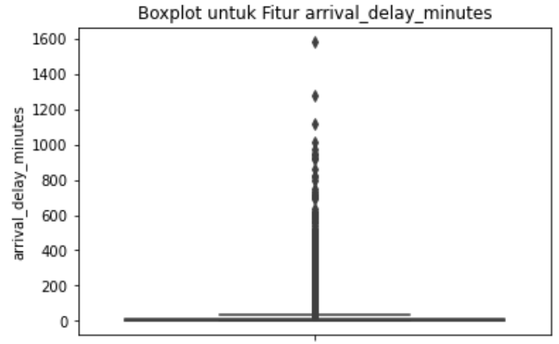
```
df.drop_duplicates(keep = 'first', inplace = True)
```

Tidak ditemukan adanya duplikasi data

Mendeteksi Data Outlier



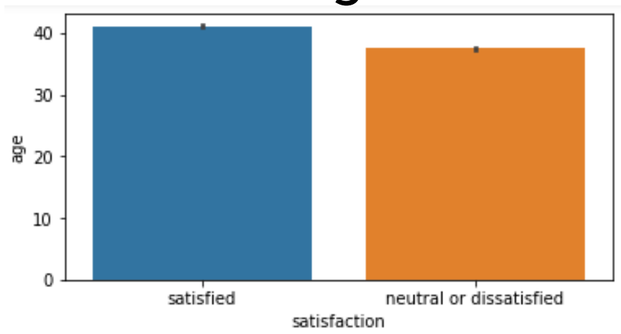




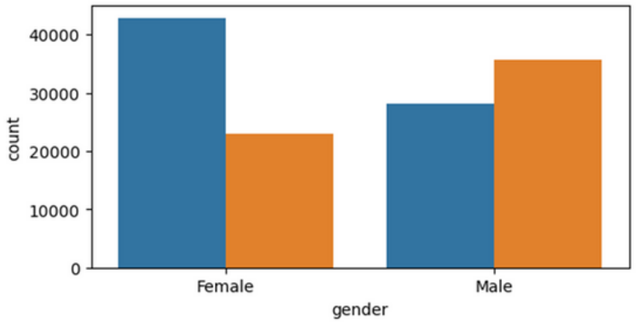
Outlier tersebut kita pertahankan karena jika dihapus atau diisi tidak merepresentasikan data yang sebenarnya

EXPLORATORY DATA ANALYSIS

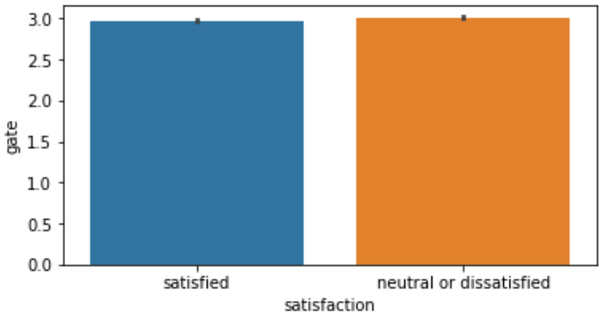
Age



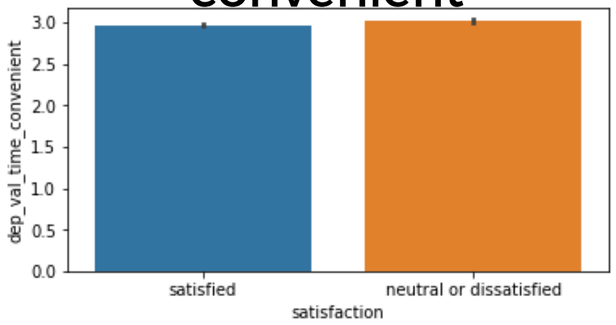
Gender



Gate

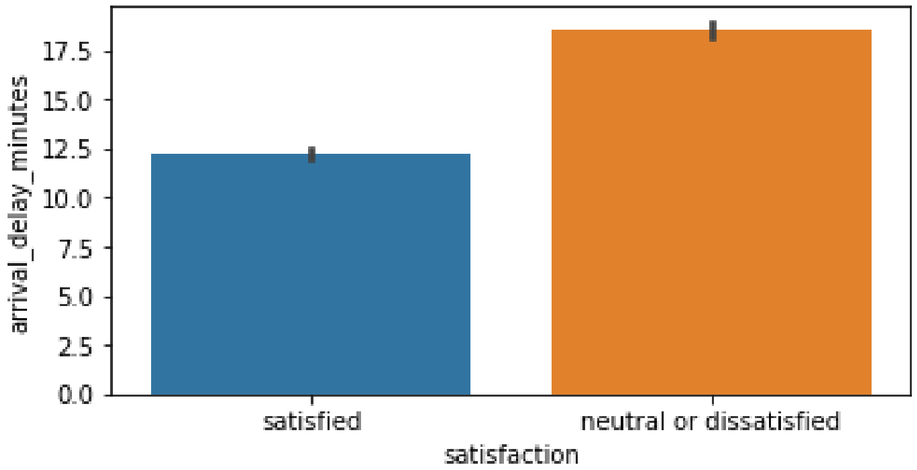


Dept. /arrival time convenient

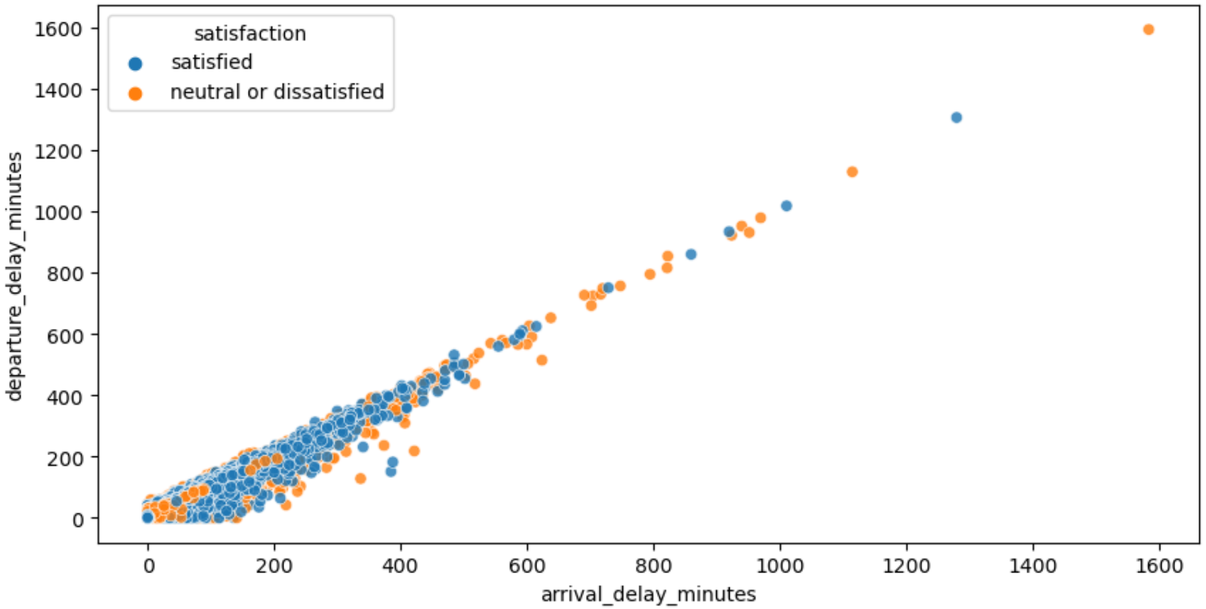


Keempat fitur tersebut akan dihapus karena tidak terlalu mempengaruhi kepuasan penerbangan dan juga tidak terlalu memberi banyak informasi. Kolom tinggal 19.

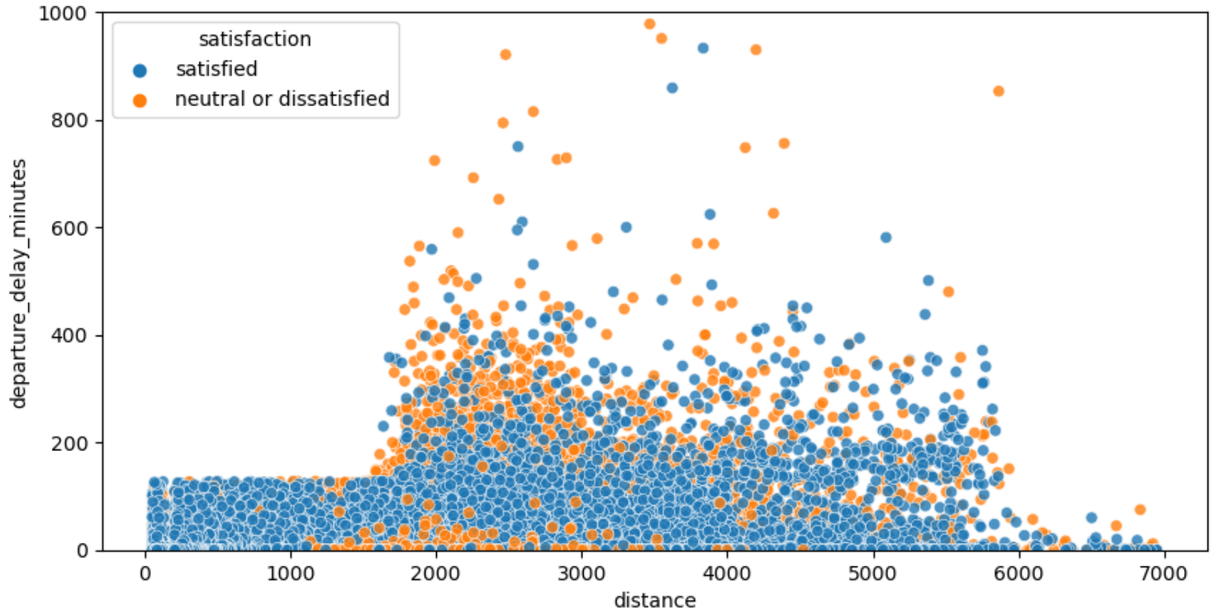
Arrival delay



VISUALISASI

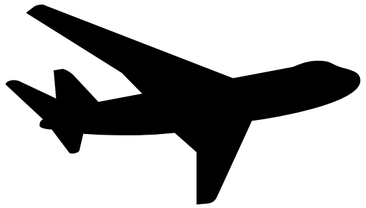


Penundaan kedatangan dan keberangkatan memiliki hubungan yang linear.



Semakin jauh jarak penerbangan, kebanyakan penumpang tidak keberatan dengan penundaan sebentar dalam keberangkatan, sedangkan penumpang yang memiliki penerbangan jarak dekat tidak terlalu puas. Dari hal ini bisa diketahui bahwa penundaan keberangkatan bukan menjadi faktor kepuasan.

MAPPING DATA



```
df['satisfaction'] = df['satisfaction'].map({'neutral or dissatisfied':0 , 'satisfied':1})
df['customer_type'] = df['customer_type'].map({'Loyal Customer':1, 'disloyal Customer':0})
df['travel_type'] = df['travel_type'].map({'Personal Travel':0, 'Business travel':1})
df['class'] = df['class'].map({'Eco':0, 'Eco Plus':1, 'Business':2})
```

SPLITTING DATA

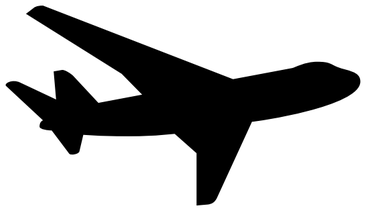
Rasio
80 : 20

Jumlah data training:

Jumlah data testing:

```
#Scaling fitur dengan metode pipeline, dan scaler standar
pipeline = Pipeline([
            ('std_scaler',StandardScaler()),
                    ])
scaled_X_train = pipeline.fit_transform(X_train)
scaled_X_test = pipeline.transform(X_test)
```

CONFUSION MATRIX



Confusion Matrix merupakan pengukuran performa untuk klasifikasi machine learning dimana keluaranya dapat berupa dua kelas atau lebih.

Tabel 1. *Confusion Matrix*

Confusion Matrix		Kelas Aktual	
		1(Positive)	0 (Negative)
Kelas Prediksi	1(Positive	TP	FP
	0 (Negative)	FN	TN

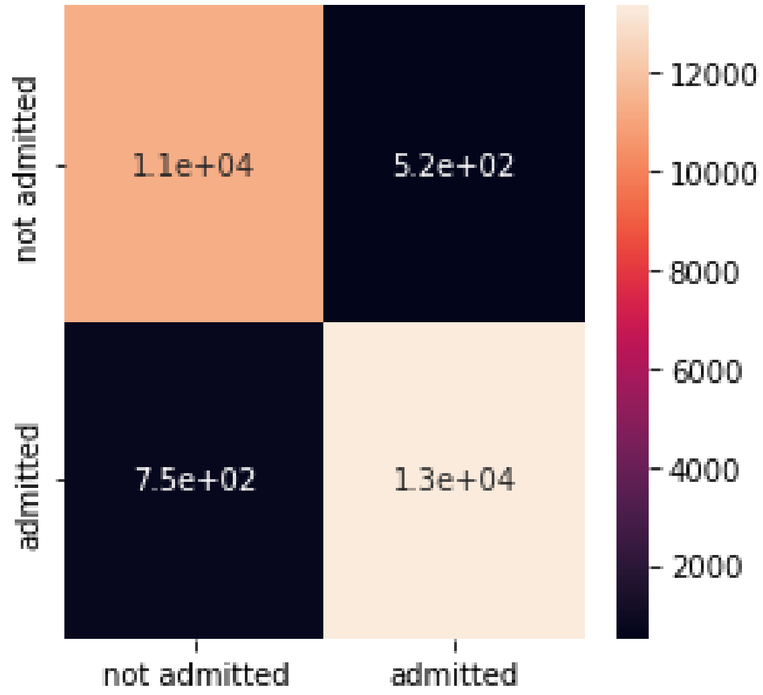
dengan:

- True Positive (TP) : Merupakan data positif yang diprediksi benar
- True Negative (TN) : Merupakan data negatif yang diprediksi benar
- False Positive (FP): Merupakan data negatif namun diprediksi sebagai data positif
- False Negative (FN): Merupakan data positif namun diprediksi sebagai data negatif

- $Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$
- $Precision = \frac{TP}{TP+FP}$
- $Sensitivity/Recall = \frac{TP}{TP+FN}$
- $Specificity = \frac{TN}{TN+FP}$
- $F1 - Score = \frac{2\ x\ precision\ x\ recall}{precision + recall}$

PERFORMANCE ANALYSIS

XGBOOST



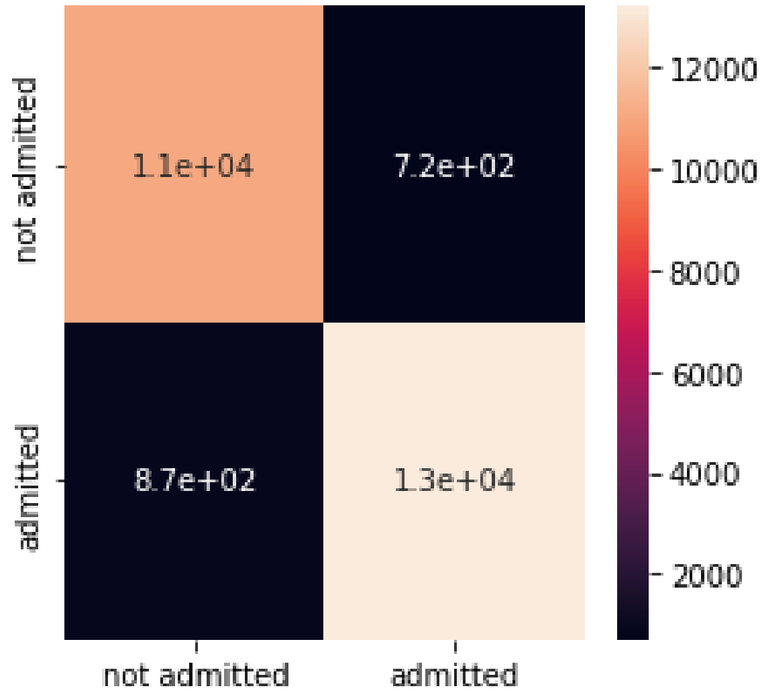
- TP = 13329
- TN = 11303
- FP = 518
- FN = 748

- Akurasi = (TN + TP) / (TN + FP + FN + TP)
  = (24632) / (25898)
  = 0.951

- Presisi = TP / (TP + FP) = 0.963
- Recall = TP / (TP + FN) = 0.947
- F1-Score = 2 * (Presisi * Recall) / (Presisi + Recall)
  = 0.955

PERFORMANCE ANALYSIS

SVM



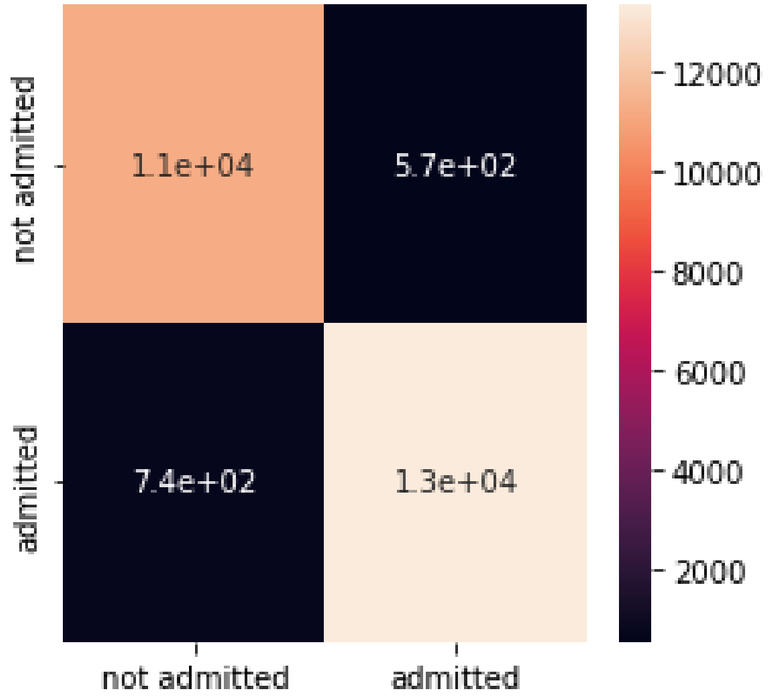
- TP = 13205
- TN = 11103
- FP = 718
- FN = 872

- Akurasi = (TN + TP) / (TN + FP + FN + TP)
  = (24310 / (25900)
  = 0.939

- Presisi = TP / (TP + FP) = 0.948
- Recall = TP / (TP + FN) = 0.938
- F1-Score = 2 * (Presisi * Recall) / (Presisi + Recall)
  = 0.943

PERFORMANCE ANALYSIS

RANDOM FOREST

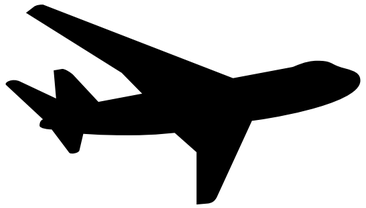


- TP = 13338
- TN = 11252
- FP = 569
- FN = 739

- Akurasi = (TN + TP) / (TN + FP + FN + TP)
  = (24590) / (25898)
  = 0.949

- Presisi = TP / (TP + FP) = 0.959
- Recall = TP / (TP + FN) = 0.948
- F1-Score = 2 * (Presisi * Recall) / (Presisi + Recall)
  = 0.953

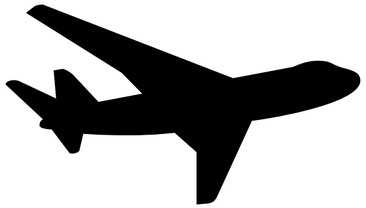
HASIL & PEMBAHASAN



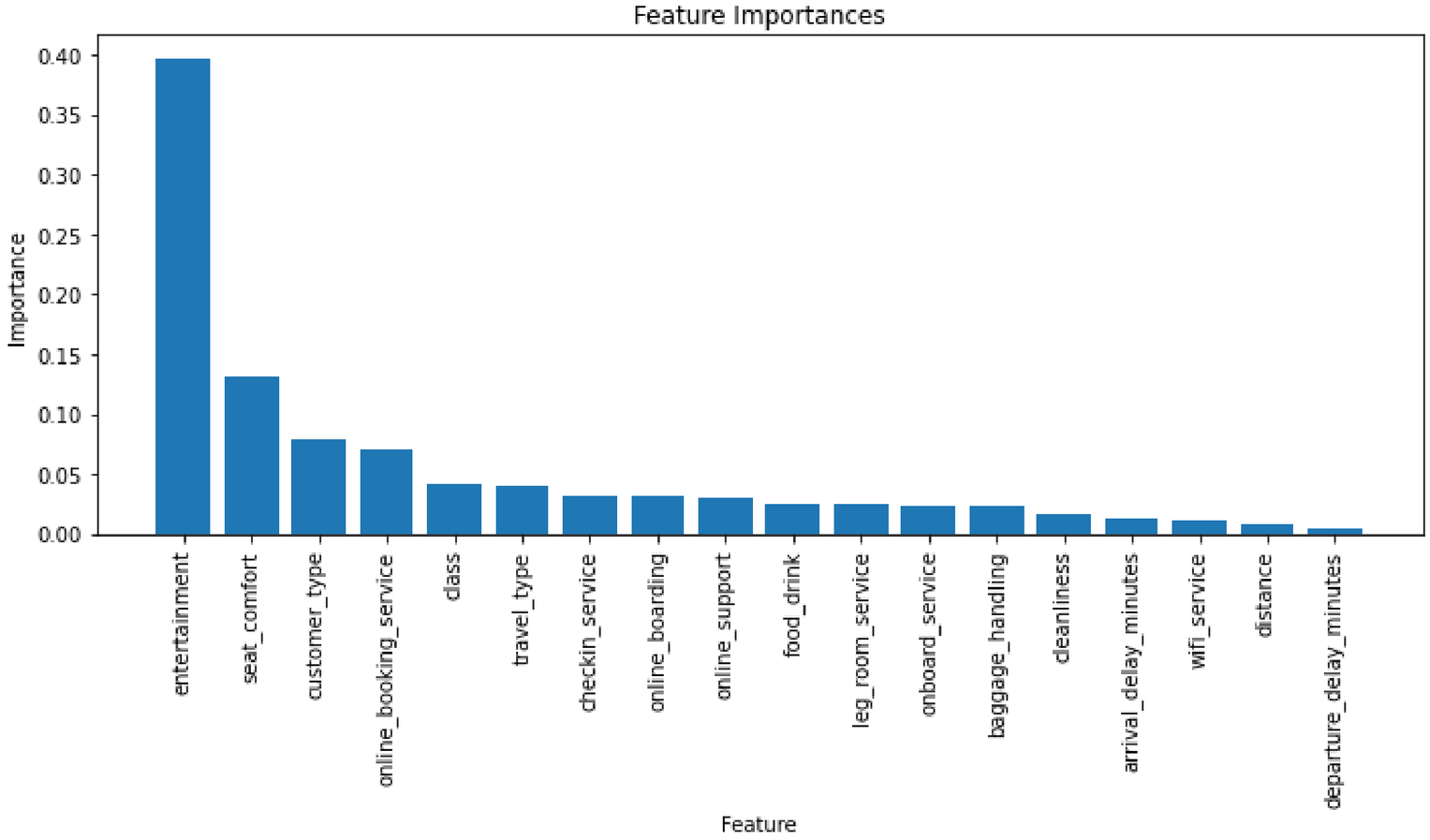
Algoritma	Akurasi
XGBOOST	95,1%
SVM	93,9%
Random Forest	94,9%

Algoritma terbaik: XGBOOST

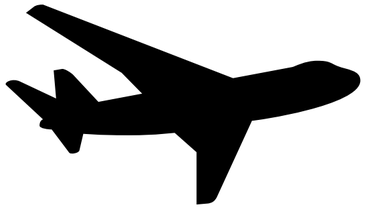
INSIGHT BARU



```
entertainment: 0.3964584767818451
seat_comfort: 0.13088694214820862
customer_type: 0.07929284125566483
online_booking_service: 0.07049394398927689
class: 0.041308820247650146
travel_type: 0.0394633375108242O3
checkin_service: 0.03234986960887909
online_boarding: 0.03106897883117199
online_support: 0.029797013849020004
food_drink: 0.024572933092713356
leg_room_service: 0.024479830637574196
onboard_service: 0.023614173755049706
baggage_handling: 0.022785391658544454
cleanliness: 0.017145197838544846
arrival_delay_minutes: 0.012166867032647133
wifi_service: 0.011433177627623081
distance: 0.007983206771314144
departure_delay_minutes: 0.004699028097093105
```



KESIMPULAN



Berdasarkan penelitian yang telah kami lakukan dapat disimpulkan bahwa metode pemodelan klasifikasi dapat diterapkan dengan baik pada dataset Kepuasan Penumpang Pesawat Terbang tahun 2019.

Berdasarkan hasil analisis ketiga metode klasifikasi, didapatkan perbedaan hasil akurasinya yang tidak terlalu signifikan. Diperoleh bahwa metode XGBoost merupakan metode terbaik pada dataset ini karena mampu mencapai nilai akurasi tertinggi di antara ketiga metode lainnya dengan nilai akurasi sebesar 95,1%. Metode XGBoost juga memberikan kinerja terbaik berdasarkan nilai precision, recall, dan F1-score yang di mana merupakan nilai terbesar di antara metode klasifikasi lainnya. Sedangkan metode SVM memiliki akurasi paling rendah sebesar 93,9% yang merupakan nilai terendah di antara ketiga metode klasifikasi lainnya.

KELOMPOK 6

Terima Kasih!

Thank you for making attention to us.

MACHINE LEARNING B