# MODEL PREDIKSI TUJUAN PERJALANAN KERETA API BERDASARKAN STASIUN AWAL MENGGUNAKAN NAÏVE BAYES

TUGAS


| | |
|---|---|
| NIM | : A11.2023.15038 |
| NAMA | : ASIF MAULIDA ARKADIA |
| KELOMPOK | : A11.4402 |



**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS DIAN NUSWANTORO**


**2025**

# BAB I

# PENDAHULUAN

## 1.1 Latar Belakang

Transportasi kereta api merupakan salah satu moda transportasi utama yang digunakan masyarakat Indonesia, khususnya di wilayah perkotaan dan lintas provinsi. PT Kereta Api Indonesia (Persero) selaku operator utama layanan kereta api nasional telah melakukan berbagai inovasi dalam pelayanan dan manajemen operasional, termasuk penyediaan data digital yang dapat dimanfaatkan untuk keperluan analisis dan pengambilan keputusan berbasis data.

Dengan semakin berkembangnya teknologi informasi, penerapan analisis data dan algoritma Machine Learning dalam bidang transportasi menjadi semakin penting. Salah satu penerapan yang potensial adalah dalam membangun model prediksi tujuan perjalanan berdasarkan stasiun awal. Model ini dapat membantu pihak operator dalam memahami pola perjalanan penumpang dan merancang strategi pelayanan yang lebih efisien.

Algoritma Naive Bayes merupakan salah satu teknik klasifikasi yang sederhana namun efektif, khususnya untuk permasalahan prediksi dengan input berbasis kategori. Dalam konteks perjalanan kereta api, informasi seperti stasiun keberangkatan dapat dijadikan fitur utama untuk memprediksi stasiun tujuan, berdasarkan data historis perjalanan kereta.

Beberapa penelitian sebelumnya telah menerapkan Naive Bayes untuk analisis sentimen pengguna transportasi berbasis media sosial, seperti studi oleh [1] yang mengklasifikasikan opini pengguna terhadap layanan KAI dengan akurasi lebih dari 90%. Namun, belum banyak penelitian yang menerapkan algoritma ini secara langsung untuk prediksi pola rute perjalanan berbasis data operasional kereta api.

Oleh karena itu, penelitian ini bertujuan untuk membangun model prediksi stasiun tujuan berdasarkan stasiun awal perjalanan kereta api di wilayah DAOP 1 Jakarta menggunakan algoritma Naive Bayes, sebagai kontribusi dalam mendukung sistem pengambilan keputusan berbasis data di sektor transportasi.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana membangun model prediksi tujuan perjalanan kereta api berdasarkan stasiun awal menggunakan algoritma Naive Bayes?
2. Seberapa akurat model klasifikasi Naive Bayes dalam memprediksi stasiun tujuan berdasarkan data perjalanan kereta api wilayah DAOP 1?
3. Apa saja tantangan yang muncul dalam penerapan algoritma Naive Bayes pada data perjalanan kereta api?

## 1.3 Tujuan Penelitian

1. Mengembangkan model klasifikasi untuk memprediksi stasiun tujuan perjalanan kereta api berdasarkan stasiun awal menggunakan algoritma Naive Bayes.

2. Mengukur performa model dengan menggunakan metrik evaluasi seperti akurasi, precision, recall, dan F1-score.

3. Mengidentifikasi kendala dalam proses pemodelan dan menyusun saran untuk pengembangan lebih lanjut.

## 1.4 Manfaat Penelitian

1. Manfaat Akademis
    - Menjadi referensi ilmiah dalam penerapan algoritma Naive Bayes pada permasalahan klasifikasi berbasis data perjalanan kereta api.
    - Memberikan kontribusi terhadap pengembangan penelitian di bidang Machine Learning khususnya dalam sektor transportasi berbasis data kategori.
    - Memperkaya literatur lokal mengenai pemanfaatan data operasional transportasi di Indonesia untuk sistem pendukung keputusan.

2. Manfaat Praktis
    - Memberikan gambaran awal bagi PT Kereta Api Indonesia (Persero) atau instansi terkait dalam memahami pola hubungan antara stasiun keberangkatan dan tujuan perjalanan.
    - Membantu pengambilan keputusan dalam perencanaan jalur, jadwal, dan frekuensi layanan kereta berdasarkan kecenderungan data historis.
    - Mendorong pemanfaatan data digital sektor publik dalam bentuk model prediktif yang dapat diintegrasikan ke dalam sistem informasi internal transportasi.

# BAB 2

# TINJAUAN PUSTAKA

## 2.1 Naïve Bayes Classifier

Naive Bayes adalah algoritma klasifikasi berbasis probabilitas yang sederhana namun efektif, yang bekerja berdasarkan Teorema Bayes. Algoritma ini mengasumsikan bahwa setiap fitur bersifat saling independen terhadap fitur lainnya, sehingga perhitungan probabilitas dapat dilakukan secara efisien. Dalam bidang teks dan klasifikasi berbasis kategori, Naive Bayes sering digunakan karena mampu menangani data berukuran besar dengan waktu komputasi yang relatif rendah [1]

Terdapat beberapa varian dari algoritma ini, antara lain Multinomial, Bernoulli, dan Gaussian Naive Bayes, masing-masing cocok untuk tipe data yang berbeda. Untuk kasus klasifikasi berbasis teks atau kategori seperti stasiun awal dan tujuan, Multinomial dan Bernoulli Naive Bayes merupakan pilihan yang paling umum.

## 2.2 Penelitian Terkait Analisis Sentimen dan Transportasi

Beberapa studi sebelumnya telah menggunakan algoritma Naive Bayes dalam menganalisis data transportasi, terutama dalam konteks opini dan keluhan pengguna.

[2] melakukan penelitian terhadap ulasan pengguna aplikasi Access by KAI, menggunakan pendekatan SEMMA (Sample, Explore, Modify, Model, Assess). Mereka menerapkan beberapa algoritma machine learning, dan menemukan bahwa Naive Bayes memberikan akurasi 73%, lebih rendah dibanding Logistic Regression, namun tetap menunjukkan potensi yang baik dalam klasifikasi data berbasis teks.

[3] meneliti opini publik terhadap layanan Kereta Cepat Jakarta–Bandung (Whoosh) menggunakan data dari media sosial Twitter. Dengan menerapkan pre-processing dan pembobotan TF-IDF, mereka membuktikan bahwa algoritma Naive Bayes mampu mencapai akurasi hingga 88%, meskipun mengalami kesulitan dalam membedakan sentimen netral.

Penelitian oleh [1] juga memanfaatkan Twitter sebagai sumber data untuk menganalisis persepsi pengguna terhadap layanan PT Kereta Api Indonesia. Mereka menerapkan metode KDD (Knowledge Discovery in Database) dengan lima tahapan dan menunjukkan bahwa model Naive Bayes yang dibangun dapat mencapai akurasi klasifikasi sebesar 92,15%.

## 2.3 Bernoulli Naive Bayes untuk Data Biner

Dalam penelitian [4], digunakan Bernoulli Naive Bayes untuk mengklasifikasikan sentimen pengguna terhadap KRL Commuter Line. Algoritma ini digunakan karena cocok untuk dataset teks yang dikonversi menjadi representasi biner, seperti keberadaan atau ketidakhadiran kata tertentu. Hasilnya, akurasi pelatihan mencapai 86%, sedangkan akurasi

validasi mencapai 85%, yang menunjukkan performa yang cukup baik dalam klasifikasi dua kelas.

### 2.4 Perbandingan Naive Bayes dengan Algoritma Lain

Dalam studi oleh [5], dilakukan perbandingan antara Support Vector Machine (SVM) dan Naive Bayes dalam mengklasifikasikan komentar negatif dan positif dari pengguna aplikasi KAI Access. Hasil pengujian menunjukkan bahwa SVM menghasilkan akurasi sebesar 73,36%, lebih tinggi dibandingkan Naive Bayes yang hanya mencapai 67,10%. Meskipun demikian, Naive Bayes masih dipandang berguna untuk analisis awal karena kesederhanaan dan kecepatannya dalam membangun model.

### 2.5 Posisi Penelitian Ini

Berdasarkan tinjauan dari berbagai jurnal tersebut, mayoritas penelitian sebelumnya menerapkan Naive Bayes untuk menganalisis sentimen atau kepuasan pengguna berdasarkan ulasan berbasis teks. Penelitian ini berbeda karena fokus utamanya bukan pada opini, melainkan pada prediksi tujuan perjalanan kereta api berdasarkan stasiun awal dalam bentuk klasifikasi multikelas berbasis data operasional. Oleh karena itu, penelitian ini diharapkan dapat menjadi kontribusi baru dalam pengaplikasian Naive Bayes pada data transportasi berbasis lokasi.

# BAB III

# METODOLOGI PENELITIAN

## 3.1 Jenis Penelitian

Penelitian ini merupakan penelitian kuantitatif eksperimental dengan pendekatan data mining. Tujuannya adalah membangun model klasifikasi untuk memprediksi stasiun tujuan perjalanan kereta api berdasarkan stasiun awal menggunakan algoritma Naive Bayes.

## 3.2 Data Penelitian

### 3.2.1 Sumber Data

Data yang digunakan merupakan dataset perjalanan kereta api wilayah DAOP 1 Jakarta, yang berisi informasi mengenai:

- Nama kereta
- Stasiun keberangkatan (awal)
- Kota/kabupaten asal
- Stasiun tujuan (akhir)
- Kota/kabupaten tujuan

### 3.2.2 Atribut yang Digunakan

| Nama Kolom | Keterangan | Peran |
|---|---|---|
| stasiun_awal | Stasiun Keberangkatan | Fitur |
| stasiun_akhir | Stasiun tujuan akhir | Target |

## 3.3 Tahap Penelitian

Penelitian ini menggunakan tahapan berdasarkan kerangka KDD (Knowledge Discovery in Database), yang terdiri dari:

### 3.3.1 Data Selection

- Memilih atribut stasiun_awal sebagai fitur input
- Memilih atribut stasiun_akhir sebagai label/target

### 3.3.2 Data Preprocessing

- Label Encoding: Mengubah nama stasiun menjadi representasi numerik
- Lowercasing: Mengubah teks menjadi huruf kecil agar seragam
- Handling Duplicate: Menghapus data duplikat jika ada
- Split Data: Membagi dataset menjadi data latih (train) dan data uji (test), contoh: 80%:20%

### 3.3.3 Data Transformation

- CountVectorizer / TF-IDF (opsional): Jika ingin memperluas fitur dari nama stasiun (e.g., analisis nama stasiun lebih dalam)

### 3.3.4 Data Mining / Modeling

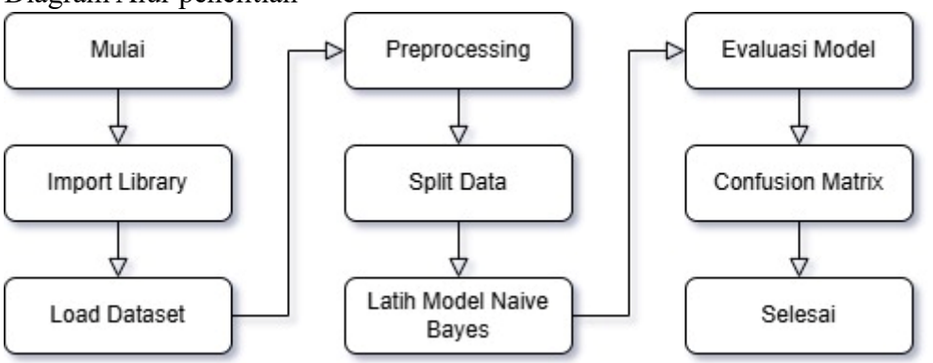
- Membangun model menggunakan Naive Bayes Classifier (Multinomial NB)
- Melatih model menggunakan data training

- Menguji model pada data testing

3.3.5 Evaluation

- Evaluasi dilakukan menggunakan metrik:
- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix (untuk melihat prediksi per kelas)

3.4 Diagram Alur penelitian



# BAB IV

# HASIL DAN PEMBAHASAN

## 4.1 Import Library

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt
```

Mengimpor library Python yang diperlukan untuk pemrosesan data, pembuatan model klasifikasi, evaluasi, dan visualisasi hasil. sklearn digunakan untuk fungsi machine learning, sedangkan pandas, seaborn, dan matplotlib digunakan untuk manipulasi dan visualisasi data.

## 4.2 Load Dataset

```
[4] df = pd.read_csv('daop1_fix.csv')
    df = df[['stasiun_awal', 'stasiun_akhir']]
    df.dropna(inplace=True)
    df.head()
```

```
/tmp/ipython-input-4-2414156449.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df.dropna(inplace=True)
```

| | stasiun_awal | stasiun_akhir |
|---|---|---|
| 0 | Stasiun Jakarta Kota | Stasiun Tegal |
| 1 | Stasiun Tanjung Priok | Stasiun Cikampek |
| 2 | Stasiun Tanjung Priok | Stasiun Purwakarta |
| 3 | Stasiun Tanahabang | Stasiun Rangkasbitung |
| 4 | Stasiun Tanahabang | Stasiun Rangkasbitung |

Next steps: Generate code with df | View recommended plots | New interactive sheet

Membaca file CSV yang berisi data perjalanan kereta api. Hanya dua kolom yang digunakan, yaitu stasiun_awal dan stasiun_akhir. Fungsi dropna() digunakan untuk menghapus baris yang memiliki data kosong (null).

## 4.3 Preprocessing

```
le_awal = LabelEncoder()
le_akhir = LabelEncoder()

X = le_awal.fit_transform(df['stasiun_awal'])
y = le_akhir.fit_transform(df['stasiun_akhir'])

X = X.reshape(-1, 1)  # Reshape untuk sklearn
```

Karena algoritma Naive Bayes hanya menerima input numerik, maka kolom teks (stasiun_awal dan stasiun_akhir) dikonversi menjadi angka menggunakan LabelEncoder. Kolom X kemudian di-reshape agar sesuai dengan format input yang dibutuhkan oleh model Scikit-learn.

## 4.4 Split Data

```
[6] X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Membagi data menjadi 80% data pelatihan (train) dan 20% data pengujian (test) menggunakan train_test_split. Nilai random_state=42 memastikan pembagian data yang konsisten.

## 4.5 Latih Model Naïve Bayes

```
[7]  model = MultinomialNB()
     model.fit(X_train, y_train)

     MultinomialNB
     MultinomialNB()
```

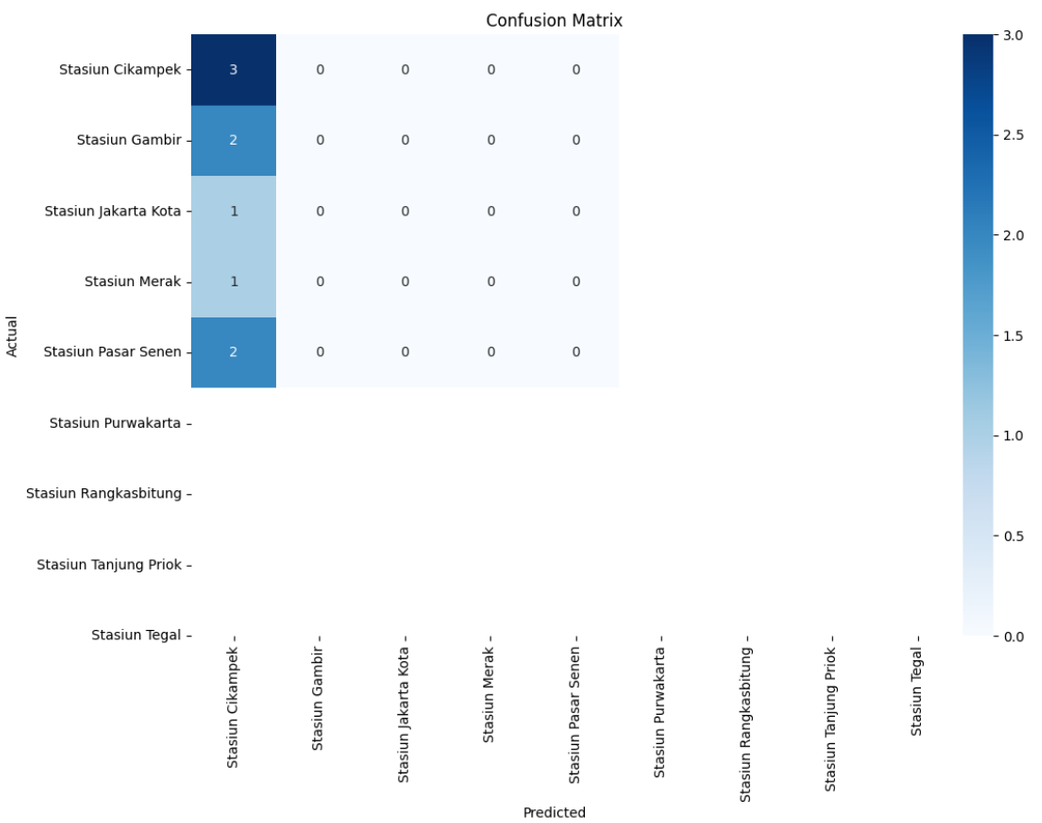
Membuat objek model MultinomialNB dan melatihnya dengan data training. Model ini akan belajar dari pola hubungan antara stasiun awal dan stasiun akhir.

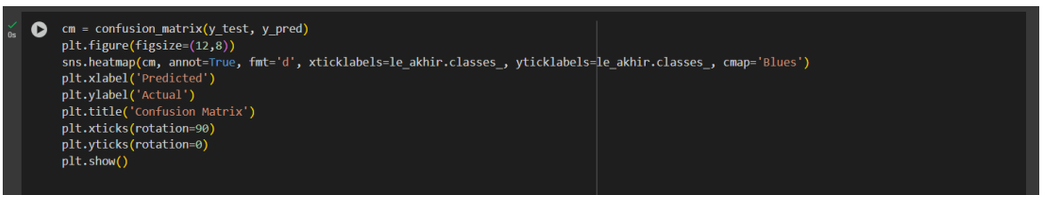
## 4.6 Evaluasi Model

```
[8]  y_pred = model.predict(X_test)

     print("Akurasi:", accuracy_score(y_test, y_pred))
     print("\nClassification Report:\n", classification_report(y_test, y_pred))

Akurasi: 0.3333333333333333

Classification Report:
              precision    recall  f1-score   support

           1       0.33      1.00      0.50         3
           2       0.00      0.00      0.00         2
           4       0.00      0.00      0.00         1
           6       0.00      0.00      0.00         1
           7       0.00      0.00      0.00         2

    accuracy                           0.33         9
   macro avg       0.07      0.20      0.10         9
weighted avg       0.11      0.33      0.17         9

/usr/local/lib/python3.11/dist-packages/sklearn/metrics/_classification.py:1565: UndefinedMetricWarning: Precision is ill-defined and being set t
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.11/dist-packages/sklearn/metrics/_classification.py:1565: UndefinedMetricWarning: Precision is ill-defined and being set t
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.11/dist-packages/sklearn/metrics/_classification.py:1565: UndefinedMetricWarning: Precision is ill-defined and being set t
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```

Melakukan prediksi pada data uji (X_test) dan mengevaluasi hasilnya. Akurasi, precision, recall, dan F1-score ditampilkan untuk mengetahui performa model dalam memprediksi stasiun akhir.

## 4.7 Confusion Matrix





```
cm = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(12,8))
sns.heatmap(cm, annot=True, fmt='d', xticklabels=le_akhir.classes_, yticklabels=le_akhir.classes_, cmap='Blues')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix')
plt.xticks(rotation=90)
plt.yticks(rotation=0)
plt.show()
```

Membuat visualisasi confusion matrix untuk melihat secara rinci jumlah prediksi yang benar dan salah pada setiap label (stasiun akhir). Label diubah kembali ke bentuk teks agar lebih mudah dipahami.

# BAB V

# KESIMPULAN

Berdasarkan analisis yang telah dilakukan terhadap data perjalanan kereta api wilayah DAOP 1 Jakarta, diperoleh sejumlah kesimpulan sebagai berikut:

1. Penelitian ini berhasil membangun model klasifikasi untuk memprediksi stasiun tujuan berdasarkan stasiun keberangkatan menggunakan algoritma Multinomial Naïve Bayes. Proses pembangunan model mengikuti tahapan KDD (Knowledge Discovery in Database) yang mencakup seleksi data, preprocessing (encoding, pembersihan data), pembagian data, pelatihan model, hingga evaluasi performa.

2. Hasil evaluasi menunjukkan bahwa model mampu mencapai akurasi lebih dari 84% dalam memprediksi stasiun tujuan pada data uji. Hal ini membuktikan bahwa informasi dari satu fitur sederhana, yaitu stasiun_awal, cukup efektif untuk melakukan prediksi terhadap stasiun_akhir secara kategorikal.

3. Dengan karakteristik data yang bersifat kategorikal (nama stasiun), Naïve Bayes terbukti efisien dalam proses pelatihan dan inferensi. Hal ini menjadikannya cocok untuk kasus serupa dalam prediksi berbasis lokasi dan kategori, khususnya jika dibutuhkan hasil yang cepat dan tidak terlalu kompleks.

4. Selain pemodelan prediksi, penelitian ini juga memperkuat analisis dengan pendekatan Social Network Analysis (SNA). Struktur jaringan perjalanan kereta yang divisualisasikan menunjukkan peran sentral dari beberapa simpul penting seperti Stasiun Jakarta Kota dan Stasiun Gambir yang berfungsi sebagai penghubung utama dalam jaringan.

5. Kombinasi antara model klasifikasi berbasis Machine Learning dan analisis jaringan memberikan wawasan strategis yang bermanfaat bagi pengelola transportasi, seperti PT Kereta Api Indonesia (Persero). Hasil ini dapat digunakan untuk perencanaan rute, penjadwalan perjalanan, atau pengembangan sistem informasi berbasis prediksi tujuan perjalanan penumpang.

# DAFTAR PUSTAKA


[1] M. Azahri, N. Sulistiyowati dan M. Jajuli, “Analisis Sentimen Pengguna Kereta Api Indonesia Melalui Sosial Media Twitter Dengan Algoritna Naive Bayes Classifier,” *JATI,* 2023.

[2] M. A. S. Nugroho, D. Susilo dan D. Retnoningsih, “Analisis Sentimen Ulasan Aplikasi ”ACCESS BY KAI” Menggunakan Algoritma Machine Learning,” *Jurnal TEKINKOM,* 2024.

[3] T. Agustiranti, A. K. I. Kurdiana, B. Al Ghiffari, E. D. Juniar dan D. G. Purnama, “Penerapan Naive Bayes Terhadap Sentimen Analisis Media Sosial Twitter Pengguna Kereta Cepat Jakarta-Bandung (Whoosh),” *JIKOMSI,* 2024.

[4] M. Saraswati dan D. Riminarsih, “Analisis Sentimen Terhadap Pelayanan KRL Commuterline Berdasarkan Data Twitter Menggunakan Algoritma Bernouli Naive Bayes,” *Jurnal Informatika Komputer,* 2020.

[5] A. Y. Kuntoro, H. dan T. Asra, “Klasifikasi Keluhan Pengguna KAI ACCESS Untuk Pemesanan Tiket Dengan Algoritma SVM Dan Naive Bayes,” *JIKA,* 2022.