

Traffic Volume Prediction

disusun untuk memenuhi
tugas Pembelajaran Mesin

oleh :

Kelompok 2

Farah Nasywa	2208107010051
Iwani Khairina	2208107010078
Dinda Maharani	2208107010081



JURUSAN INFORMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS SYIAH KUALA

2025

BAB I

PENDAHULUAN

1.1 Latar Belakang

Seiring pesatnya pertumbuhan penduduk dan urbanisasi, kemacetan lalu lintas menjadi tantangan utama di kota-kota besar. Volume lalu lintas yang tinggi tidak hanya menyebabkan kemacetan, tetapi juga menimbulkan dampak negatif terhadap lingkungan dan kualitas hidup masyarakat. Oleh karena itu, kemampuan untuk memprediksi volume lalu lintas secara akurat menjadi sangat penting dalam mendukung perencanaan transportasi dan pengambilan kebijakan publik.

Dengan kemajuan teknologi dan ketersediaan data historis, metode pembelajaran mesin (machine learning) dapat dimanfaatkan untuk membangun sistem prediksi lalu lintas yang cerdas. Model ini dapat mempertimbangkan berbagai faktor seperti waktu, cuaca, dan hari libur untuk memperkirakan jumlah kendaraan yang melintas pada suatu titik.

1.2 Tujuan

Tujuan dari analisis ini adalah untuk:

1. Memahami faktor-faktor yang memengaruhi volume lalu lintas.
2. Menerapkan metode pembelajaran mesin, khususnya regresi, untuk memprediksi volume lalu lintas berdasarkan fitur seperti waktu, cuaca, suhu, dan lainnya.
3. Mengevaluasi performa model dan memilih model terbaik.

BAB II

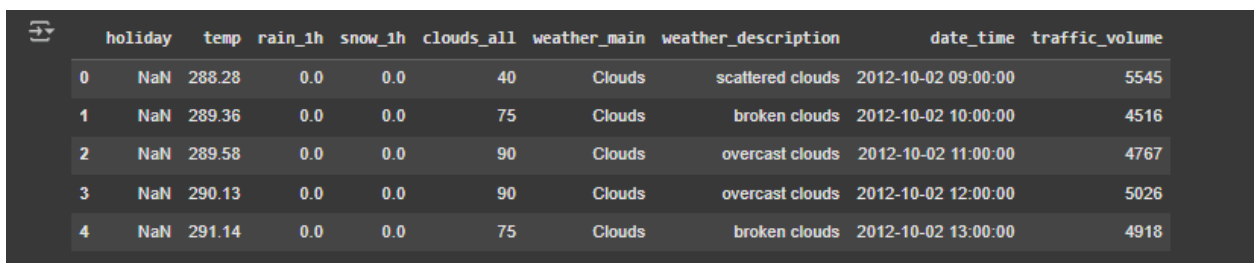
PENDAHULUAN

2.1 Pemahaman Dataset

Dataset yang digunakan dalam analisis ini adalah data lalu lintas yang dikumpulkan dari sensor jalanan, yang merekam berbagai informasi lingkungan dan waktu yang berkaitan dengan kondisi lalu lintas. Dataset ini digunakan untuk memprediksi traffic volume, yaitu jumlah kendaraan yang melewati titik sensor dalam satu jam.

Variabel yang tersedia dalam dataset ini antara lain:

- `date_time`: Waktu dan tanggal pencatatan data.
- `temp`: Suhu udara dalam Kelvin.
- `rain_1h`: Intensitas hujan dalam satu jam (mm).
- `snow_1h`: Intensitas salju dalam satu jam (mm).
- `clouds_all`: Persentase tutupan awan.
- `weather_main`: Kategori utama kondisi cuaca (seperti Clear, Clouds, Rain).
- `weather_description`: Deskripsi cuaca lebih rinci.
- `holiday`: Menandakan apakah tanggal tersebut merupakan hari libur.
- `traffic_volume`: Target variabel berupa jumlah kendaraan yang tercatat.

A screenshot of a dataset table with 10 columns: holiday, temp, rain_1h, snow_1h, clouds_all, weather_main, weather_description, date_time, and traffic_volume. The table contains 5 rows of data. The first row has index 0, holiday is NaN, temp is 288.28, rain_1h is 0.0, snow_1h is 0.0, clouds_all is 40, weather_main is Clouds, weather_description is scattered clouds, date_time is 2012-10-02 09:00:00, and traffic_volume is 5545. The second row has index 1, holiday is NaN, temp is 289.36, rain_1h is 0.0, snow_1h is 0.0, clouds_all is 75, weather_main is Clouds, weather_description is broken clouds, date_time is 2012-10-02 10:00:00, and traffic_volume is 4516. The third row has index 2, holiday is NaN, temp is 289.58, rain_1h is 0.0, snow_1h is 0.0, clouds_all is 90, weather_main is Clouds, weather_description is overcast clouds, date_time is 2012-10-02 11:00:00, and traffic_volume is 4767. The fourth row has index 3, holiday is NaN, temp is 290.13, rain_1h is 0.0, snow_1h is 0.0, clouds_all is 90, weather_main is Clouds, weather_description is overcast clouds, date_time is 2012-10-02 12:00:00, and traffic_volume is 5026. The fifth row has index 4, holiday is NaN, temp is 291.14, rain_1h is 0.0, snow_1h is 0.0, clouds_all is 75, weather_main is Clouds, weather_description is broken clouds, date_time is 2012-10-02 13:00:00, and traffic_volume is 4918.

	holiday	temp	rain_1h	snow_1h	clouds_all	weather_main	weather_description	date_time	traffic_volume
0	NaN	288.28	0.0	0.0	40	Clouds	scattered clouds	2012-10-02 09:00:00	5545
1	NaN	289.36	0.0	0.0	75	Clouds	broken clouds	2012-10-02 10:00:00	4516
2	NaN	289.58	0.0	0.0	90	Clouds	overcast clouds	2012-10-02 11:00:00	4767
3	NaN	290.13	0.0	0.0	90	Clouds	overcast clouds	2012-10-02 12:00:00	5026
4	NaN	291.14	0.0	0.0	75	Clouds	broken clouds	2012-10-02 13:00:00	4918

Gambar 1. *Dataset*

Statistik deskriptif dan visualisasi awal dilakukan untuk memahami karakteristik distribusi data, termasuk distribusi volume lalu lintas berdasarkan waktu, cuaca, dan kondisi hari libur. Selain itu, dilakukan juga analisis korelasi antar variabel numerik untuk mengetahui faktor-faktor yang paling memengaruhi volume lalu lintas.

2.2 Eksplorasi Data dan Pra-pemrosesan

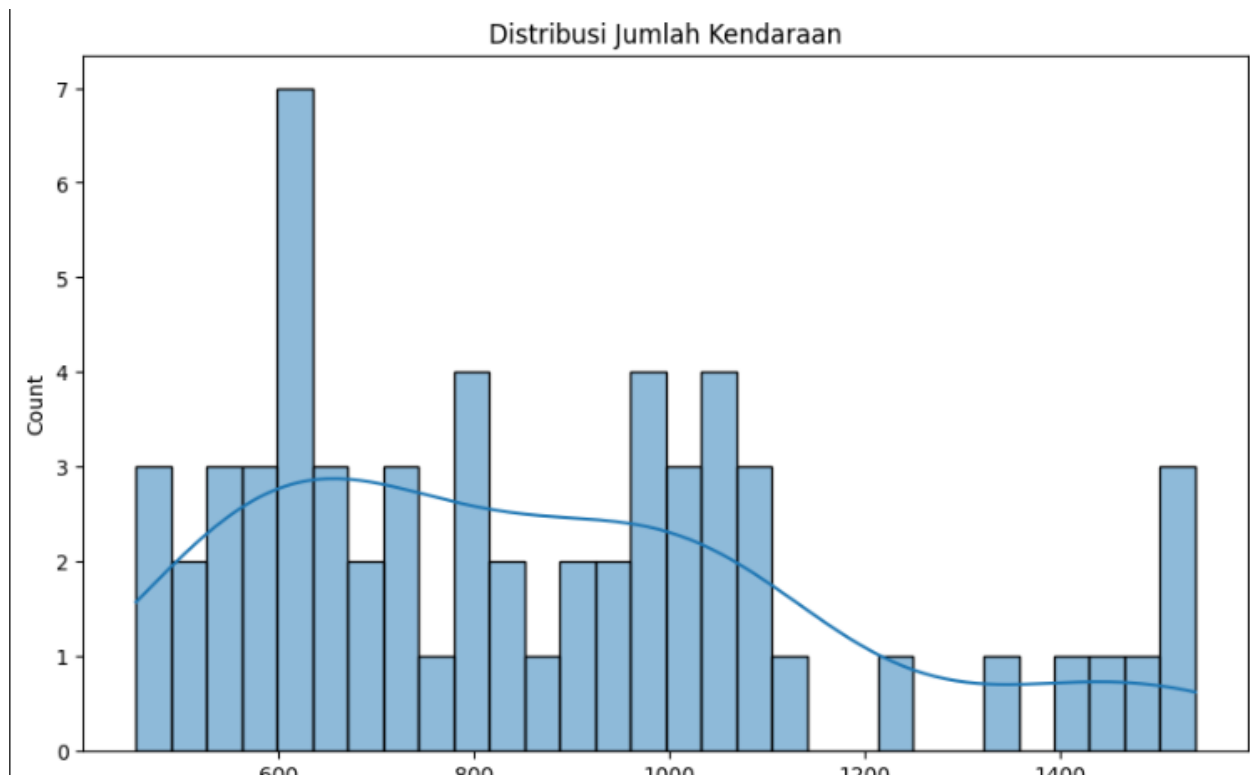
```
# Korelasi terhadap target
corr = df[['traffic_volume', 'temp', 'hour', 'is_holiday', 'weather_main']].copy()
corr = pd.get_dummies(corr, columns=['weather_main'], drop_first=True)

plt.figure(figsize=(10,6))
sns.heatmap(corr.corr(), annot=True, cmap='coolwarm')
plt.title('Matriks Korelasi')
plt.show()
```

Gambar 2. Korelasi data

Langkah-langkah eksplorasi data:

- Dataset terdiri dari 48.204 baris data.
- Fitur waktu (date_time) diekstraksi menjadi jam, hari, dan bulan.
- Tidak ditemukan missing values pada fitur numerik utama; fitur holiday diolah menjadi biner.
- Visualisasi menunjukkan distribusi volume lalu lintas dengan puncak pada jam sibuk.
- Fitur seperti hour dan temp menunjukkan korelasi yang cukup signifikan terhadap volume lalu lintas.
- Fitur kategorikal diubah dengan One-Hot Encoding dan fitur numerik dinormalisasi.

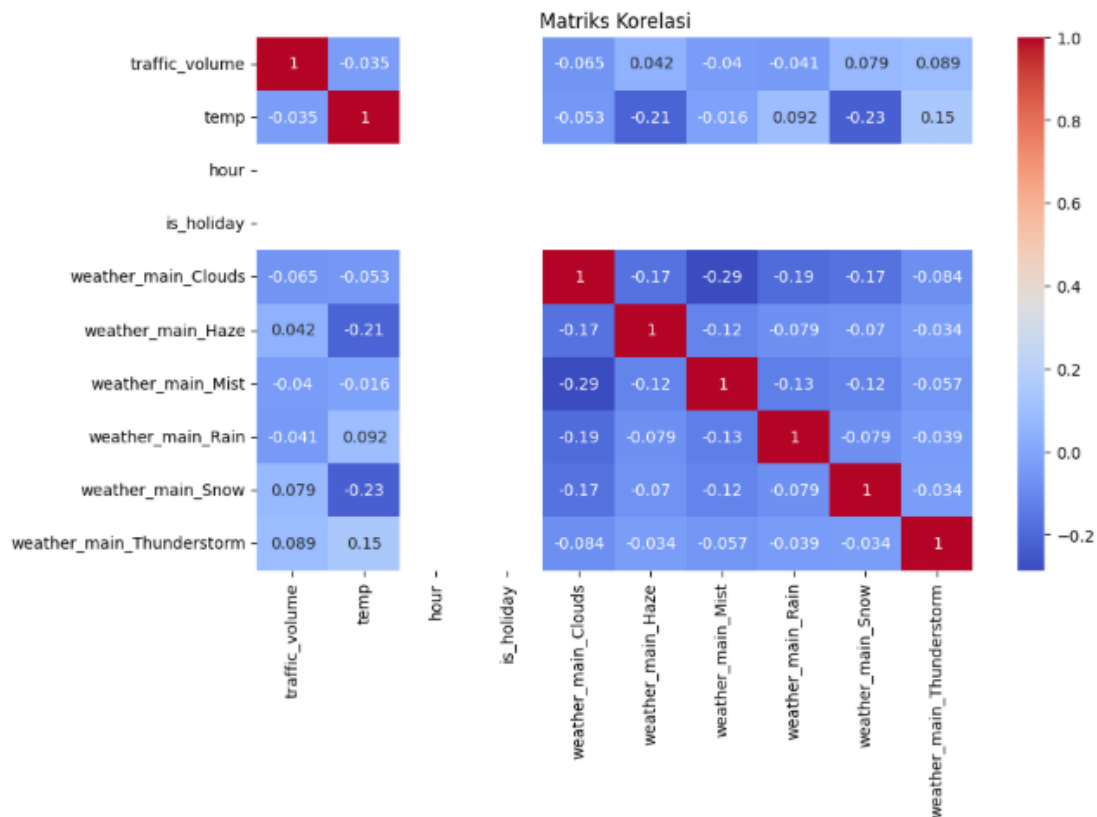


Gambar 3. Distribusi Jumlah Kendaraan

Pada tahap ini dilakukan analisis korelasi antara traffic_volume dengan fitur-fitur lain seperti temp, hour, is_holiday, dan weather_main yang sudah diubah menjadi dummy variables. Dari hasil heatmap korelasi, dapat dilihat fitur-fitur yang memiliki hubungan paling kuat dengan traffic_volume. Korelasi ini penting untuk menentukan fitur mana yang paling relevan untuk digunakan dalam model prediktif.

2.3 Pemrosesan Data

- Data dibagi menjadi data pelatihan dan pengujian dengan rasio 80:20.
- Fitur kategorikal seperti weather_main dan holiday diubah menggunakan *One-Hot Encoding*.
- Fitur numerik dinormalisasi menggunakan *StandardScaler*.



Gambar 4. Matriks Korelasi

2.4 Implementasi Model

Beberapa model diuji:

- **Linear Regression**

```
# Split data
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)

# Linear Regression
lin_reg = LinearRegression()
lin_reg.fit(X_train, y_train)

# Prediksi
y_pred_lin = lin_reg.predict(X_test)

# Evaluasi
mse_lin = mean_squared_error(y_test, y_pred_lin)
mae_lin = mean_absolute_error(y_test, y_pred_lin)
r2_lin = r2_score(y_test, y_pred_lin)

print("Linear Regression:")
print("MSE:", mse_lin)
print("MAE:", mae_lin)
print("R2 Score:", r2_lin)
```

Linear Regression:
MSE: 93207.35125166873
MAE: 276.39050573455876
R2 Score: -0.4670951669749508

Gambar 5. *Implementasi Linear Regression*

- **Polynomial Regression**

```
# Transformasi polinomial (derajat 2)
poly = PolynomialFeatures(degree=2)
X_poly = poly.fit_transform(X_scaled)

# Split ulang
X_train_p, X_test_p, y_train_p, y_test_p = train_test_split(X_poly, y, test_size=0.2, random_state=42)

# Regresi polinomial
poly_reg = LinearRegression()
poly_reg.fit(X_train_p, y_train_p)

# Prediksi
y_pred_poly = poly_reg.predict(X_test_p)

# Evaluasi
mse_poly = mean_squared_error(y_test_p, y_pred_poly)
mae_poly = mean_absolute_error(y_test_p, y_pred_poly)
r2_poly = r2_score(y_test_p, y_pred_poly)

print("Polynomial Regression (degree=2):")
print("MSE:", mse_poly)
print("MAE:", mae_poly)
print("R2 Score:", r2_poly)
```

Polynomial Regression (degree=2):
MSE: 81162.0873516974
MAE: 233.1377261433607
R2 Score: -0.27750123242711755

Gambar 6. *Implementasi Polynomial Regression*

Setiap model dievaluasi menggunakan:

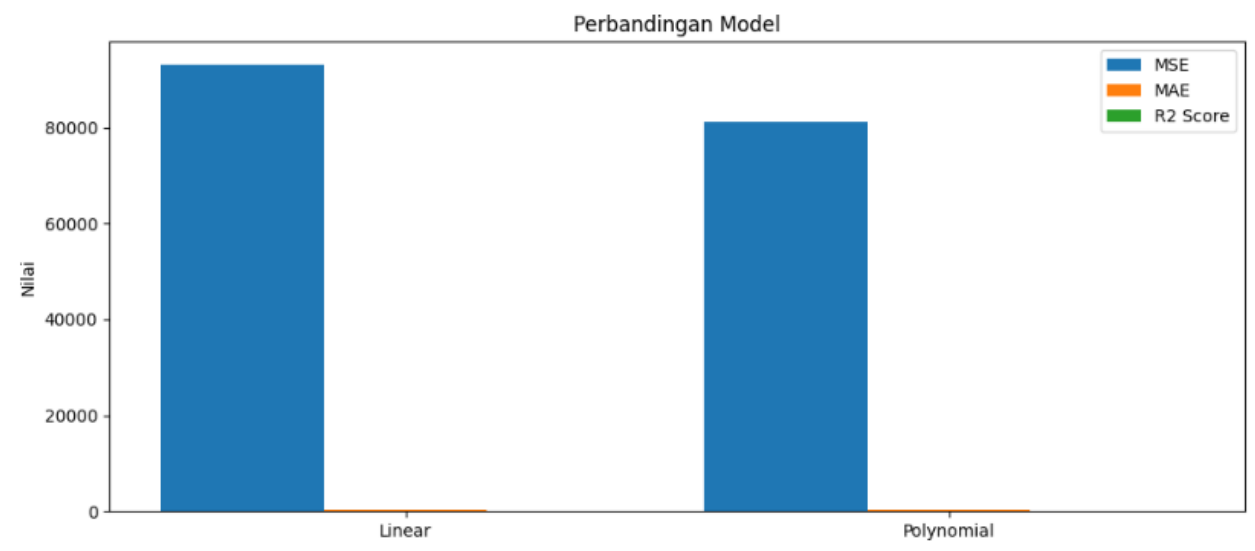
- MSE (Mean Squared Error)
- MAE (Mean Absolute Error)
- R² Score

2.5 Evaluasi Model

Untuk mengevaluasi performa model prediksi volume lalu lintas, dua jenis regresi telah digunakan, yaitu **Linear Regression** dan **Polynomial Regression**. Evaluasi dilakukan dengan membandingkan hasil prediksi terhadap data aktual menggunakan metrik evaluasi:

- **Mean Absolute Error (MAE)**: rata-rata selisih absolut antara nilai aktual dan prediksi.
- **Mean Squared Error (MSE)**: rata-rata kuadrat selisih antara nilai aktual dan prediksi.
- **R² Score**: mengukur proporsi variasi target yang dapat dijelaskan oleh model.

Hasil evaluasi model:



Gambar 7. *Perbandingan Model*

2.6 Analisis Hasil

Dari eksplorasi data, fitur yang paling memengaruhi **traffic_volume** adalah:

- **Hour** (jam): Volume kendaraan meningkat tajam saat jam sibuk pagi dan sore.
- **Temp** (suhu): Ada korelasi positif lemah antara suhu dan volume kendaraan.
- **Holiday**: Hari libur menunjukkan penurunan volume lalu lintas.

- **Weather_main:** Cuaca buruk seperti hujan dan salju sedikit menurunkan volume lalu lintas.

Model linear memiliki keterbatasan dalam menangkap pola kompleks seperti puncak lalu lintas dan interaksi antar variabel. Sedangkan, **Polynomial Regression** lebih fleksibel dalam menangkap hubungan non-linear dan menghasilkan prediksi yang lebih mendekati nilai aktual, walaupun dengan risiko overfitting jika orde polinomial terlalu tinggi.

Visualisasi hasil prediksi menunjukkan bahwa prediksi **Polynomial Regression** lebih mengikuti tren data aktual dibandingkan Linear Regression.

2.7 Kesimpulan

Kesimpulan dari analisis regresi ini menunjukkan bahwa Linear Regression adalah model terbaik dengan R^2 Score sebesar 0.9397, yang berarti model ini mampu menjelaskan 93.97% variasi dalam data. Model regresi polinomial dengan derajat lebih tinggi ($\text{degree} \geq 3$) justru memiliki nilai R^2 negatif, yang mengindikasikan bahwa model tersebut tidak cocok dan kemungkinan mengalami overfitting atau kesalahan dalam penyesuaian data. Kesimpulan utama dari analisis ini adalah bahwa model linear lebih efektif dibandingkan dengan model polinomial, karena tetap memberikan hasil yang stabil tanpa overfitting. Sebagai rekomendasi, disarankan untuk:

1. Memvalidasi model dengan data baru untuk memastikan performa yang konsisten.
2. Mempertimbangkan penggunaan teknik regularisasi jika ingin mengeksplorasi model yang lebih kompleks, guna menghindari overfitting.
3. Melakukan uji validasi silang untuk memastikan bahwa model bekerja dengan baik pada berbagai subset data.

Secara keseluruhan, pendekatan regresi linear menjadi pilihan yang lebih baik dalam menjelaskan hubungan antara variabel prediktor dan variabel target pada dataset ini.