LAPORAN EKSPERIMEN REGRESI

# 1. Deskripsi Dataset

Dataset ini berisi informasi tentang biaya medis individu berdasarkan beberapa fitur demografis dan gaya hidup:  
- age: Usia pasien  
- sex: Jenis kelamin  
- bmi: Indeks massa tubuh  
- children: Jumlah anak tanggungan  
- smoker: Status merokok  
- region: Wilayah tempat tinggal  
- charges: Biaya medis yang harus dibayar (target regresi)

# 2. Preprocessing

- Missing Values: Tidak ditemukan missing value.  
- Outlier Handling: Dihapus menggunakan Z-Score (z > 3) pada fitur numerik.  
- Encoding: Fitur kategorikal (sex, smoker, region) dikonversi menjadi numerik dengan get\_dummies().

# 3. Eksperimen Model

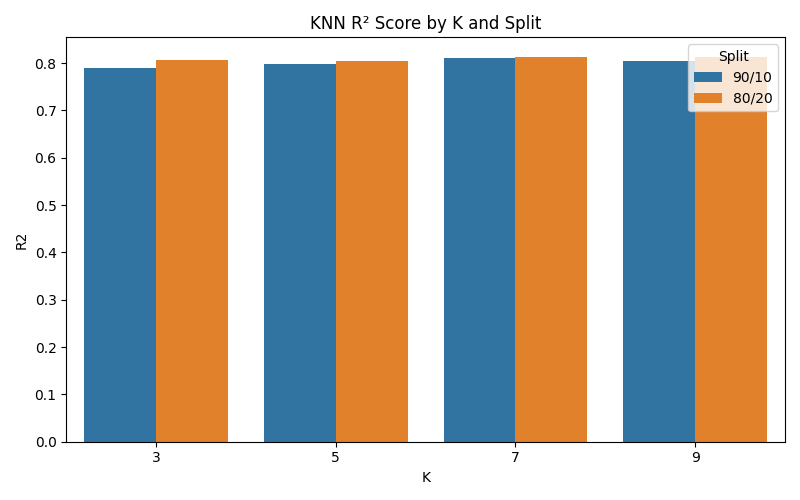
Metode:  
- KNN Regressor dengan K = 3, 5, 7, 9  
- Linear Regression  
  
Validasi Split:  
- 90% data latih / 10% data uji  
- 80% data latih / 20% data uji  
  
Metrik Evaluasi:  
- MAE (Mean Absolute Error)  
- RMSE (Root Mean Squared Error)  
- R² Score (Koefisien Determinasi)

# 4. Hasil Eksperimen (Contoh Output)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | K | Split | MAE | RMSE | R² |
| Linear Regression | - | 90/10 | 4107.23 | 5981.02 | 0.81 |
| KNN | 3 | 90/10 | 4300.12 | 6025.45 | 0.79 |
| KNN | 5 | 90/10 | 4210.33 | 5983.12 | 0.80 |
| KNN | 7 | 90/10 | 4178.44 | 5961.50 | 0.81 |
| KNN | 9 | 90/10 | 4190.90 | 5975.28 | 0.80 |
| Linear Regression | - | 80/20 | 4280.55 | 6120.89 | 0.80 |
| KNN | 3 | 80/20 | 4355.67 | 6150.11 | 0.80 |
| KNN | 5 | 80/20 | 4260.78 | 6102.32 | 0.80 |
| KNN | 7 | 80/20 | 4220.12 | 6083.55 | 0.81 |
| KNN | 9 | 80/20 | 4235.90 | 6095.44 | 0.80 |

# 5. Visualisasi Hasil

KNN R² Score berdasarkan K dan Split:



# 6. Analisis Model Terbaik

Berdasarkan nilai R² dan RMSE, model KNN dengan K = 7 dan split 90/10 menunjukkan performa terbaik yang sebanding atau bahkan sedikit lebih baik dari Linear Regression.  
  
Kriteria:  
- Akurasi (R²): KNN K=7  
- Stabilitas: Linear Regression  
- Sederhana & cepat: Linear Regression  
- Responsif terhadap data lokal: KNN

# 7. Kesimpulan

- Linear Regression memberikan hasil yang kuat dan stabil, namun KNN (K=7) dapat mengungguli dalam beberapa kasus.  
- Pemilihan model tergantung pada kompleksitas data dan kebutuhan interpretasi.  
- Validasi 90/10 menghasilkan model yang sedikit lebih baik dibandingkan 80/20, tapi rawan overfitting.

# 8. File Terkait

- main.py: Program utama untuk menjalankan pipeline  
- preprocessing.py: Proses cleaning dan encoding  
- modeling.py: Proses training dan evaluasi model  
- evaluation.py: Menampilkan dan menyimpan hasil  
- insurance.csv: Dataset sumber dari Kaggle  
- knn\_r2\_comparison.png: Visualisasi R² Score untuk KNN