**Tugas Data Mining**

Nama : Oktavian Andi Cahya Nugraha

NIM : P31.2024.02647

Mata Kuliah : Data Mining

**Tugas Regression - Data Mining**

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

1. 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().

1. Eksperimen Model

Motode :

* KNN Linear dengan K = 3, 5, 7, 9
* Linear Regression

Validasi Split :

* 90% data latih / 10% data uji
* 80% data latih / 20% data uji

Metrik Evaluasi

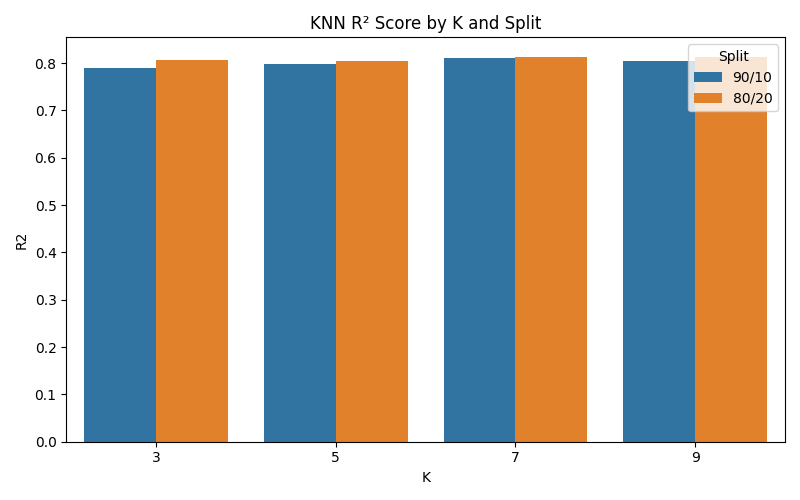
* MAE (Mean Absolute Error)
* RMSE (Root Mean Aquared Error)
* R^2 Score (Koefisien Determinasi)

1. Hasil Eksperimen

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

1. Visualisasi Hasil Eksperimen

KNN R^2 Score nerdasarkan K dan Split



1. 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 : KNN - K = 7
* Stabilitas : Linear Regression
* Sederhana & Cepat : Linear Regression
* Responsif terhadap data local : KNN

1. 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.

1. 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