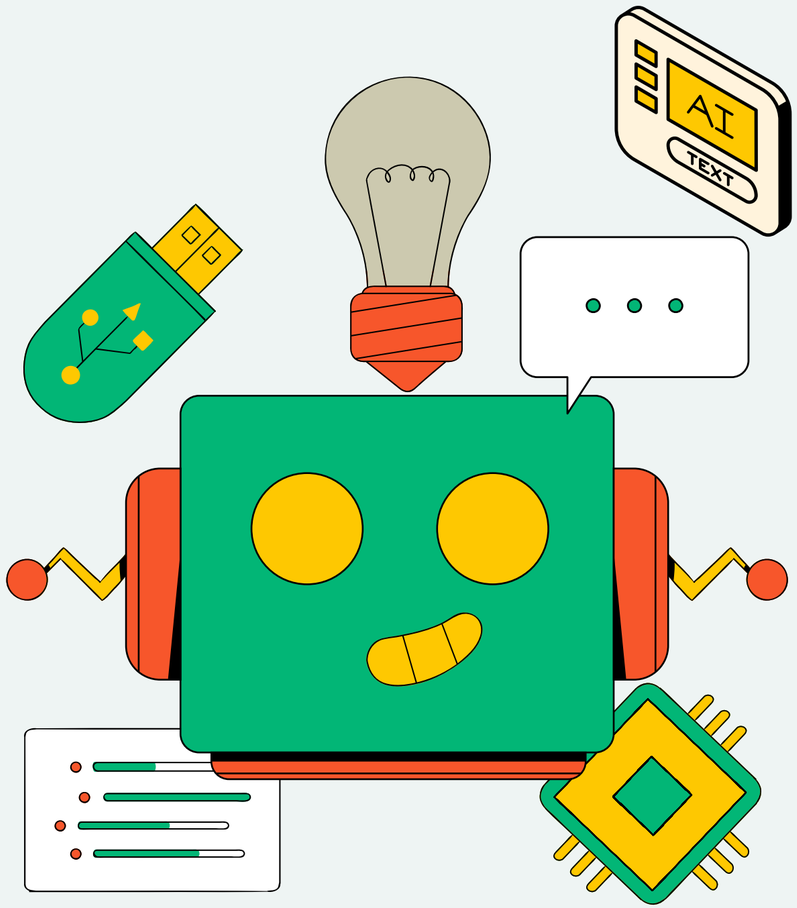
KLASIFIKASI PADA WHOLE SALE MENGGUNAKAN METODE KNN



PRESENTED BY:

LUNA AULIA – 1301223025

ADINDA LARAS – 1301223253

YULIA ADINDA – 1301223415

LATAR BELAKANG

Dalam upaya meningkatkan efektivitas strategi pemasaran dan pemahaman mengenai pola pembelian konsumen, analisis data pelanggan grosir menjadi langkah yang sangat penting. Salah satu dataset yang kita dapat manfaatkan untuk tujuan ini adalah "Wholesale Customers," yang mencakup informasi tentang tujuh fitur yaitu Channel, Fresh, Milk, Grocery, Frozen, Detergents_Paper, dan Delicassen, serta satu target yaitu Region.

Kumpulan data ini mengacu pada klien dari distributor grosir. Ini mencakup pengeluaran tahunan dalam unit moneter (m.u.) pada berbagai kategori produk. Proses ini melibatkan pembagian data menjadi data pelatihan (train) dan data pengujian (test) guna memastikan keakuratan dan keandalan model yang dikembangkan.

Link dataset : https://archive.ics.uci.edu/dataset/292/wholesale+customers

METODE K-NEAREST NEIGHBORS (KNN)

KNN (K-Nearest Neighbors) adalah salah satu algoritma dalam machine learning yang digunakan untuk melakukan klasifikasi pada data. Algoritma ini bekerja dengan mencari sejumlah K titik data (neighbors) yang terdekat dengan titik data baru yang ingin diklasifikasikan atau diprediksi nilainya.

KNN adalah algoritma non-parametrik yang digunakan untuk klasifikasi dan regresi. Dalam konteks klasifikasi, KNN akan mengklasifikasikan data baru berdasarkan mayoritas tabel dari tetangga terdekatnya. Sedangkan dalam konteks regresi, KNN akan memprediksi nilai berdasarkan rata-rata nilai dari tetangga terdekatnya

ALASAN MENGGUNAKAN METODE KNN

01

SEDERHANA DAN MUDAH DIIMPLEMENTASIKAN

Tidak memerlukan asumsi yang rumit tentang distribusi data, sangat cocok untuk pemula atau untuk situasi di mana interpretabilitas model sangat diutamakan.

02

NON-PARAMETRIC

Tidak memerlukan asumsi tentang distribusi data. Hal ini membuat KNN menjadi algoritma yang fleksibel dan dapat digunakan pada berbagai jenis data.

03

DAPAT DIADAPTASI DENGAN MUDAH

Nilai k yang digunakan dalam algoritma KNN dapat disesuaikan dengan kebutuhan. Selain itu, metode KNN dapat dikombinasikan dengan algoritma lain untuk meningkatkan performa.

04

EFEKTIVITAS UNTUK DATA YANG TIDAK LINEAR:

Fleksibel dalam menangani berbagai bentuk dan pola data yang kompleks, yang mungkin ditemukan dalam dataset penjualan grosir.

DATA FRAME

	Channel	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
266	2	572	9763	22182	2221	4882	2563
294	1	21273	2013	6550	909	811	1854
31	1	2612	4339	3133	2088	820	985
84	2	11867	3327	4814	1178	3837	120
301	2	5283	13316	20399	1809	8752	172
..	...	...	...	...	...	...	...
106	2	1454	6337	10704	133	6830	1831
270	1	4720	1032	975	5500	197	56
348	1	3428	2380	2028	1341	1184	665
435	1	29703	12051	16027	13135	182	2204
102	2	2932	6459	7677	2561	4573	1386

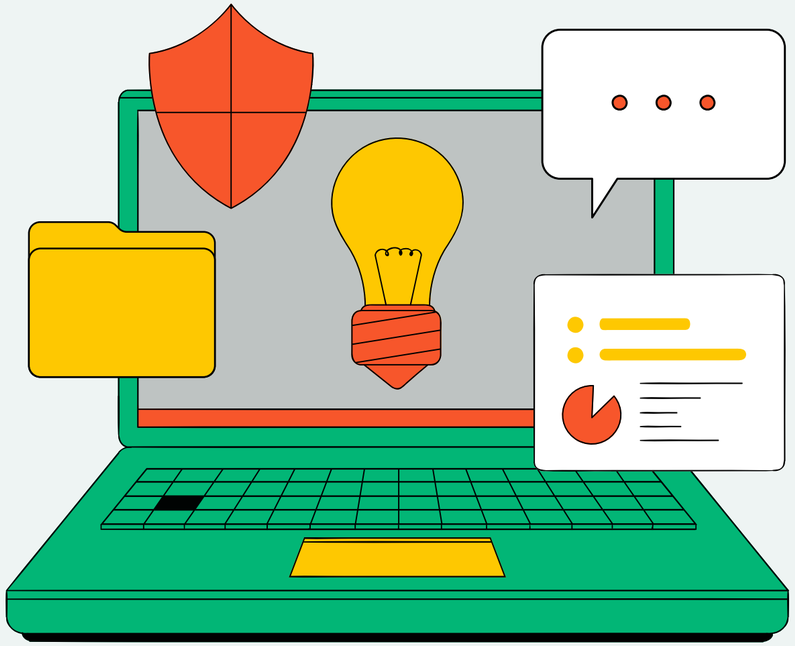
[352 rows x 7 columns]

DATA SPLIT

```
[ ]  # membagi data menjadi data train dan data test

     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.
```

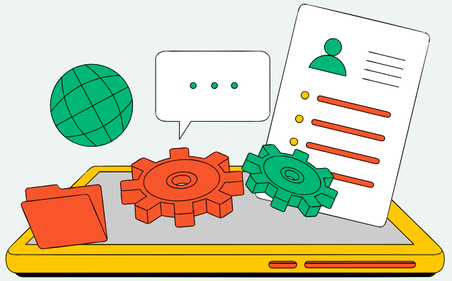
```
     # menampilkan bentuk dari data train dan data test

     print(f"X_train shape: {X_train.shape}")
     print(f"X_test shape: {X_test.shape}")
     print(f"y_train shape: {y_train.shape}")
     print(f"y_test shape: {y_test.shape}")
```

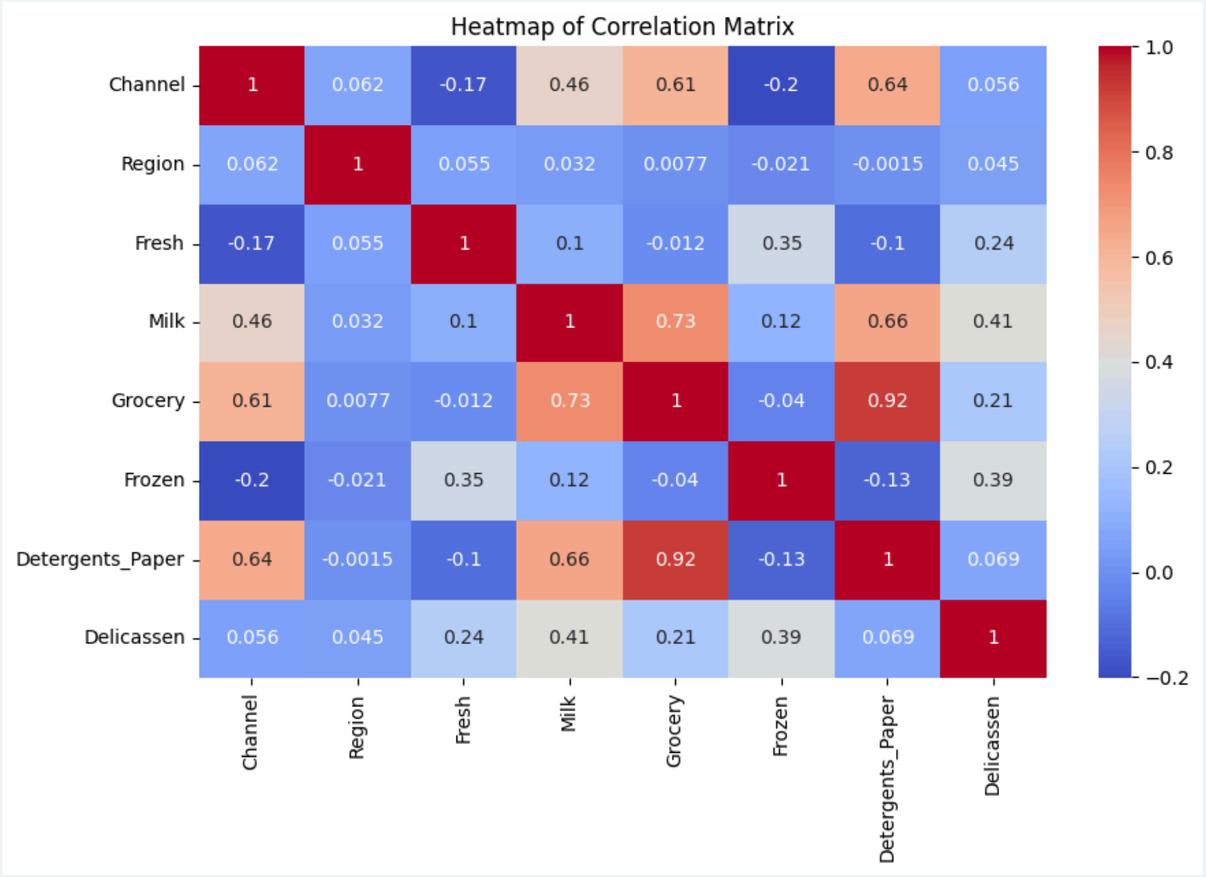
```
X_train shape: (352, 7)
X_test shape: (88, 7)
y_train shape: (352,)
y_test shape: (88,)
```



DATASET DIBAGI MENJADI 2 YAITU DATA UJI(TEST) DAN DATA LATIH(TRAIN)

VISUALISASI DATA





Warna pada heatmap menunjukkan tingkat korelasi antar variabel.
. Semakin gelap warnanya, semakin kuat tingkat korelasinya

Terdapat korelasi yang kuat antara grocery dengan detergent_paper ssekitar 0.97, yang menunjukkan bahwa kedua variabel ini sangat erat kaitannya. Dapat diartikan Ketika pembelian deterjen dan kertas meningkat, pembelian bahan makanan juga meningkat.

Ada juga korelasi yang lemah antara Frozen dengan channel yaitu -0.2, yang menunjukkan bahwa kedua variabel ini tidak terlalu erat kaitannya.

IMPLEMENTASI KNN

Google Colab

google.com

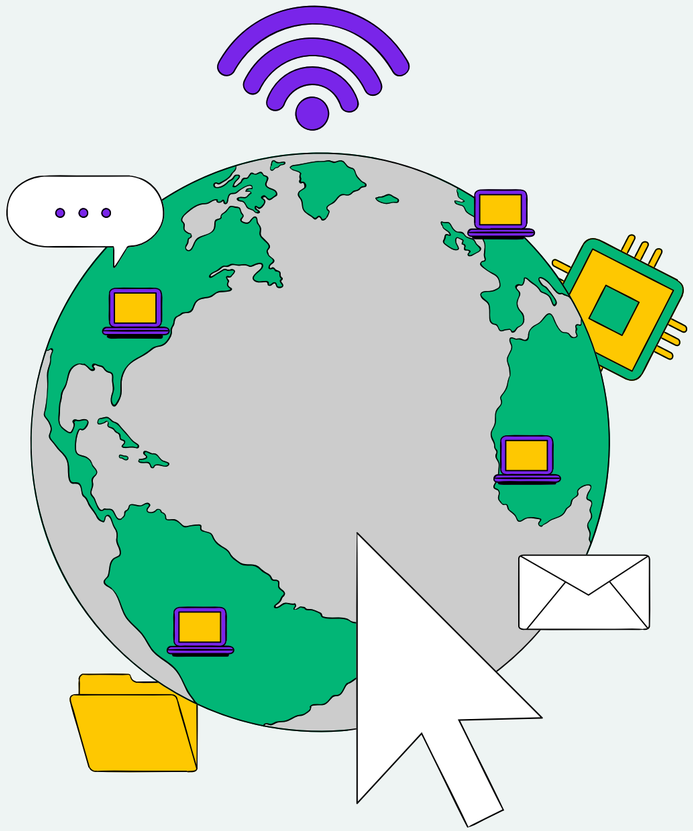
HASIL DAN ANALISIS

- K = 3
- MSE (Y1) = 1.0454545454545454
- R-SQUARED(Y1) = -1.629425138031828
- AKURASI = 67 %

- K = 5
- MSE(Y1) = 0.863636363636363636
- R-SQUARED(Y1) = -1.172133809678467
- AKURASI = 71 %

- K = 14
- MSE(Y1) = 0.82954
- R-SQUARED(Y1) = -0.3544170356314569
- AKURASI = 71.59%

- K = 7
- MSE(Y1) = 0.7045454545454546
- R-SQUARED(Y1) = -0.7720038973692755
- AKURASI = 77 %



KESIMPULAN

Nilai K yang memberikan hasil terbaik dalam hal akurasi dan performa keseluruhan adalah K = 7. Pada K = 7, model memiliki Mean Squared Error terendah dan akurasi tertinggi dibandingkan dengan nilai K lainnya yang diuji. Meskipun R-squared masih negatif, nilainya mendekati nol, menunjukkan bahwa model KNN dengan K = 7 memiliki potensi terbaik dalam memprediksi target dengan data yang tersedia.