Matematika

SEMINAR PROPOSAL

KOMPARASI AKURASI KLASIFIKASI DATA TEKS MENGGUNAKAN JARAK MANHATTAN DAN JACCARD PADA ALGORITMA KNN

Luh Sukma Mulyani - 2108541027

Kompetensi Statistika

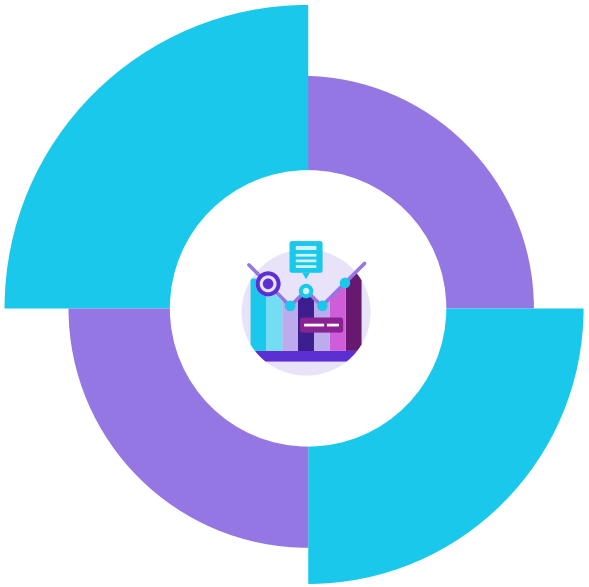
Daftar Isi

BAB 1 - Pendahuluan

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard

BAB 2 - Tinjauan Pustska

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard



BAB 3 - Metode Penelitian

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard

BAB 1 PENDAHULUAN

Latar Belakang

Lorem ipsum dolor sit amet, consectetur adipiscing

Rumusan Maasalah

Lorem ipsum dolor sit amet, consectetur adipiscing

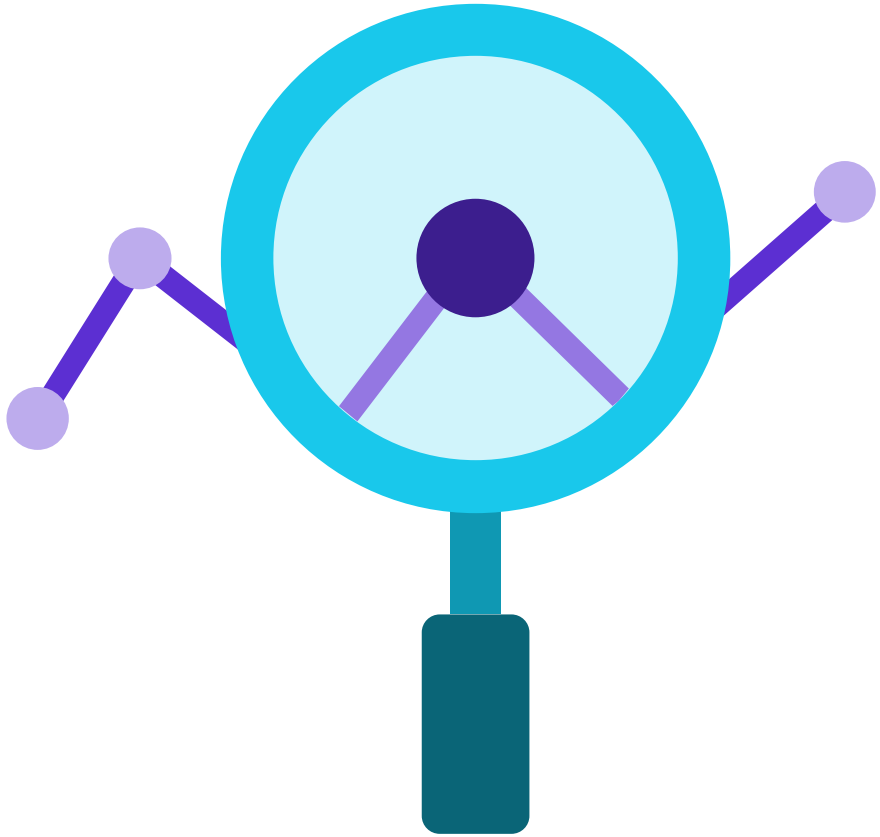
Tujuan Penelitian

Lorem ipsum dolor sit amet, consectetur adipiscing

Batasan Permasalahan

Lorem ipsum dolor sit amet, consectetur adipiscing

BAB 1 PENDAHULUAN



Latar Belakang

Pada proses analisis data, pemilihan metode klasifikasi sangat bergantung pada bagaimana algoritma menghitung kedekatan antar data. Berbagai pendekatan berbasis jarak telah dikembangkan untuk menyelesaikan masalah klasifikasi.

BAB 1 PENDAHULUAN

Latar Belakang

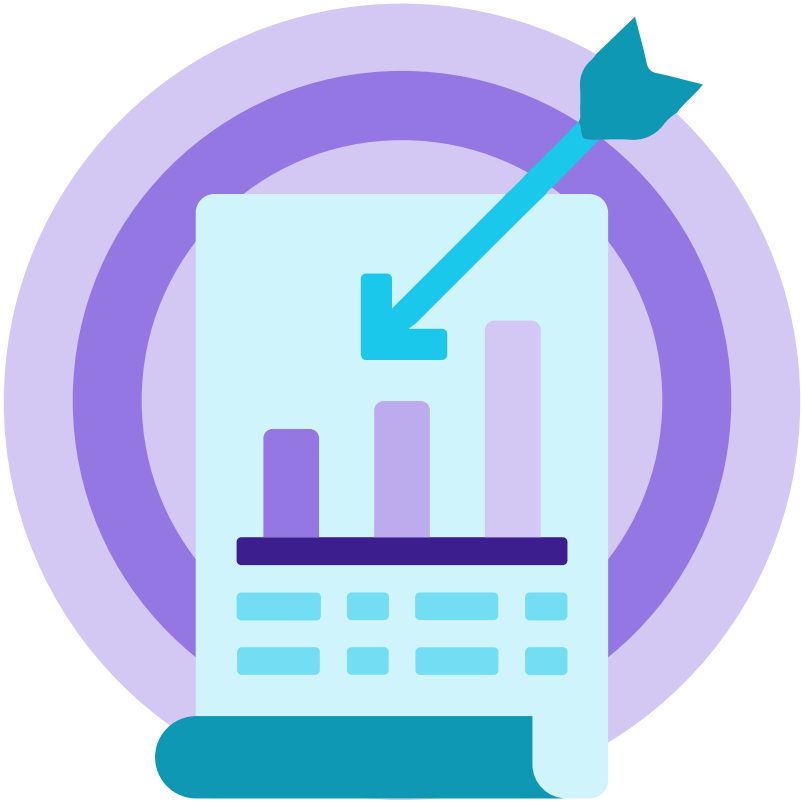
Algortima *K Nearest Neighbors* merupakan metode pembelajaran mesin terawasi yang menggunakan pendekatan jarak untuk menentukan hasil klasifikasi.

Pemilihan nilai k

Metrik Jarak

BAB 1 PENDAHULUAN

Metrik Jarak memiliki karakteristik yang berbeda dan dapat memberikan hasil yang bervariasi



Jarak Manhattan

Metrik jarak yang menghitung perbedaan absolut antara dua titik amatan dalam ruang multidimensi

Jarak Jaccard

Metrik jarak yang mengukur kesamaan antara dua amatan dengan membandingkan jumlah elemen yang sama dengan total elemen gabungan

BAB 1 PENDAHULUAN

How Smart Analysis

Rumusan Masalah

1. Bagaimana Perbandingan Akurasi Algoritma KNN Menggunakan Mertrik Jarak Manhattan dan Jaccard dalam Klasifikasi Data Teks?
2. Apakah Perbedaan Akurasi Klasifikasi Antara Kedua Metrik Jarak Signifikan Secara Statistik?

Batasan Masalah

1. Data teks direpresentasikan menggunakan metode TF-IDF;
2. Penelitian ini difokuskan pada perbandingan performa dua metrik jarak, yaitu Manhattan dan Jaccard;
3. Nilai k yang diuji dari 1 hingga 19 (ganjil)

Tujuan Masalah

1. Menganalisis perbandingan akurasi klasifikasi data teks antara metrik jarak Manhattan dan Jaccard dalam algoritma KNN.
2. Untuk menguji signifikansi statistik dari perbedaan akurasi klasifikasi antara kedua metrik jarak.

Penelitian Terdahulu

Wahyono et al. (2019)

Membandingkan metode perhitungan jarak pada KNN untuk klasifikasi teks menggunakan Euclidean, Chebyshev, Manhattan, dan Minkowski. Hasilnya, jarak Euclidean dan Minkowski mencapai akurasi tertinggi 85,5%, pada k = 3, Manhattan (85,05%), Chebyshev (61.87%)

Prasath et al. (2019)

Menguji 54 metrik jarak dengan 8 keluarga pada KNN menggunakan 28 dateset UCI. Dengan k=1, hasilnya menunjukkan bahwa Hassanat Distance memiliki performa terbaik. Manhattan dan Minkowski stabil, sementara Cosine dan Jaccard efektif untuk teks.

Jain et al. (2020)

Membandingkan metrik jarak pada KNN untuk klasifikasi ulasan Amazon (2000 data). Menunjukkan hasil Manhattan memiliki akurasi tertinggi 97,03%, diikuti Cosine, Dice, dan Jaccard (96,43%). Euclidean dan Chebyshev memiliki performa lebih rendah.

BAB 3 Metode Penelitian

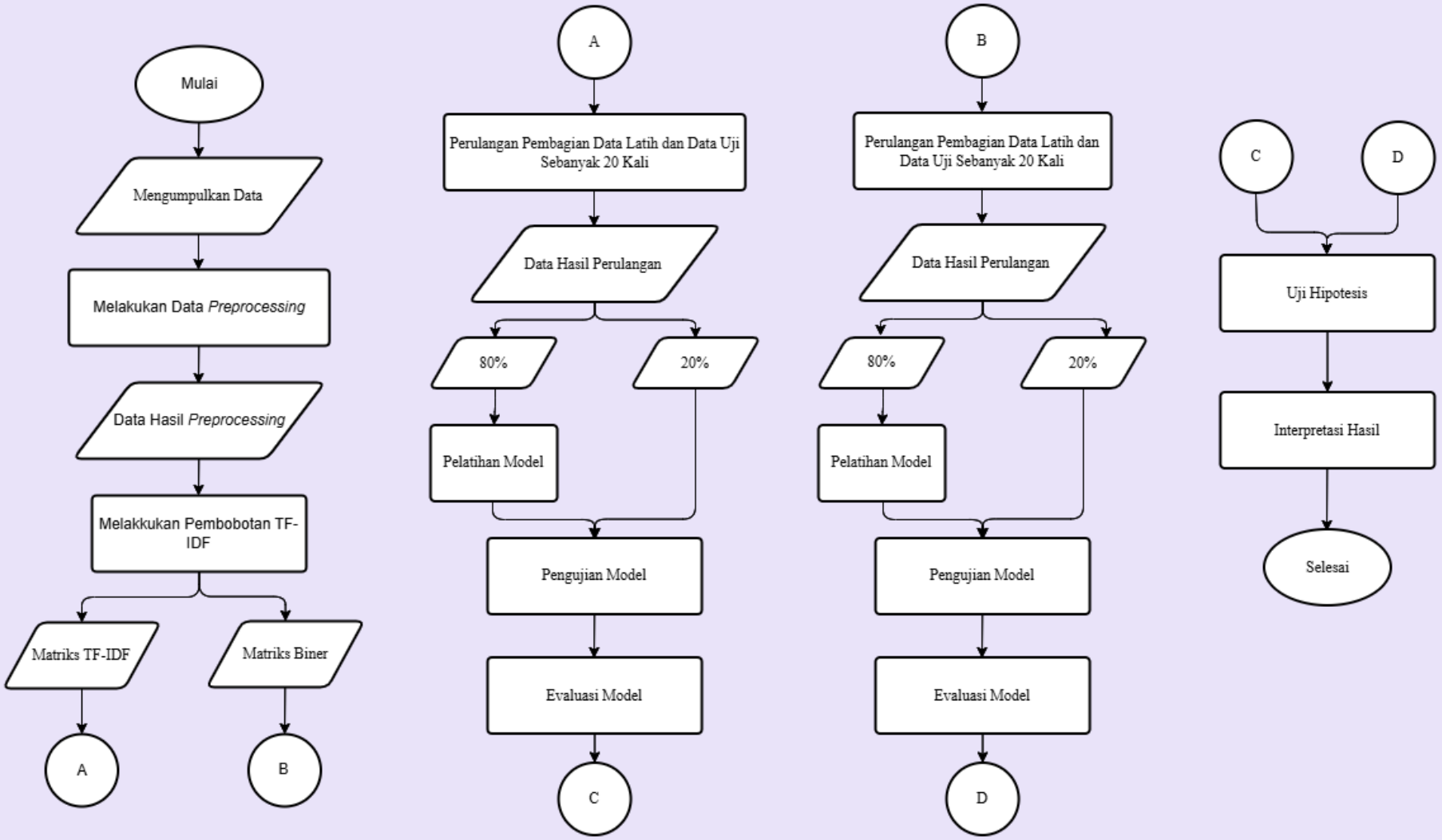
Jenis dan Sumber Data

Penelitian ini menggunakan data sekunder yang diperoleh dari Kaggle berupa 448 komentar Youtube pada video Eminem (6 Mei 2015-29 Mei 2015). Data diklasifikasikan menjadi SPAM (245) dan HAM (203) berdasarkan atribut CONTENT dan CLASS.

Variabel Penelitian

Penelitian ini menggunakan CLASS (Y) sebagai variabel target (HAM: 0, SPAM: 1) dan CONTENT (X) sebagai variabel prediktor yang diukur berdasarkan jumlah kata dalam komentar Youtube.

Pelaksana Penelitian

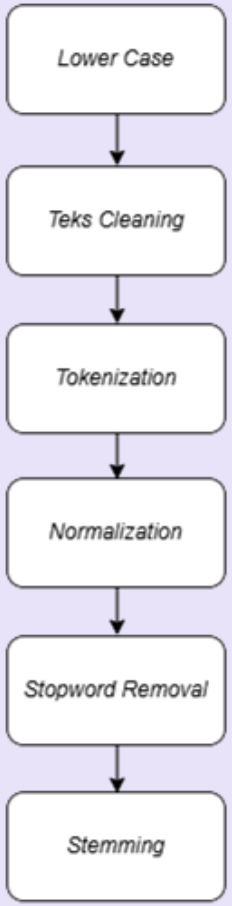


Pelaksana Penelitian

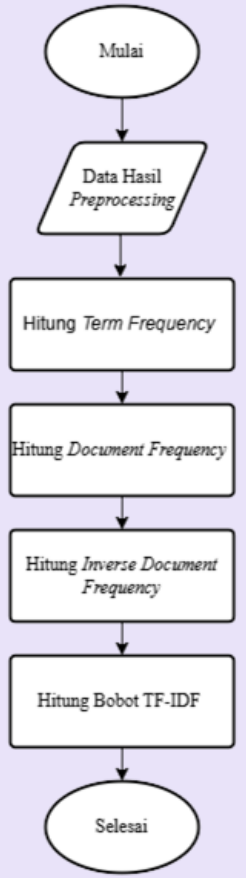
Pengumpulan Data

Penelitian ini menggunakan 448 komentar Youtube yang diperoleh dari Kaggle.

Data Prepocessing

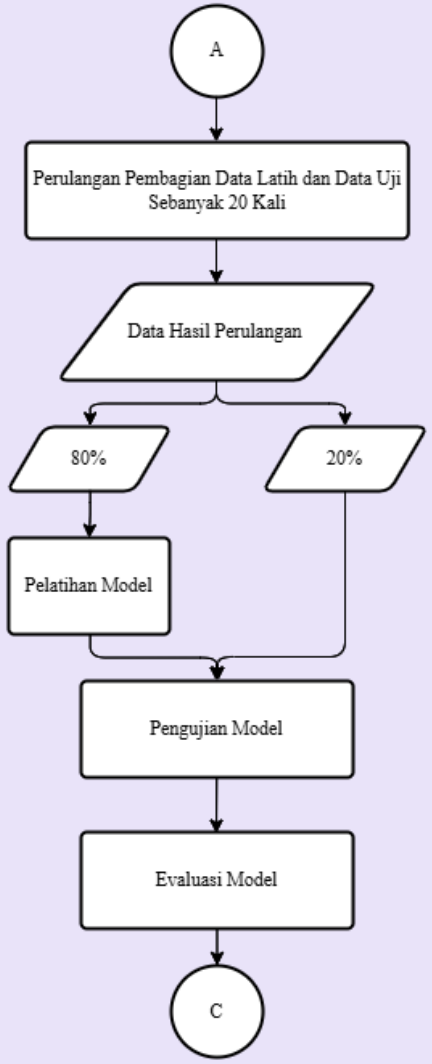


Pembobotan TF-IDF

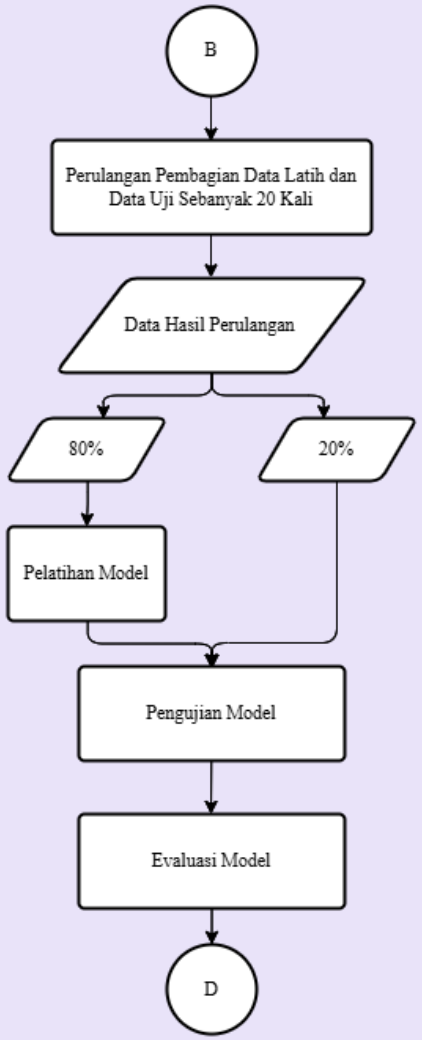


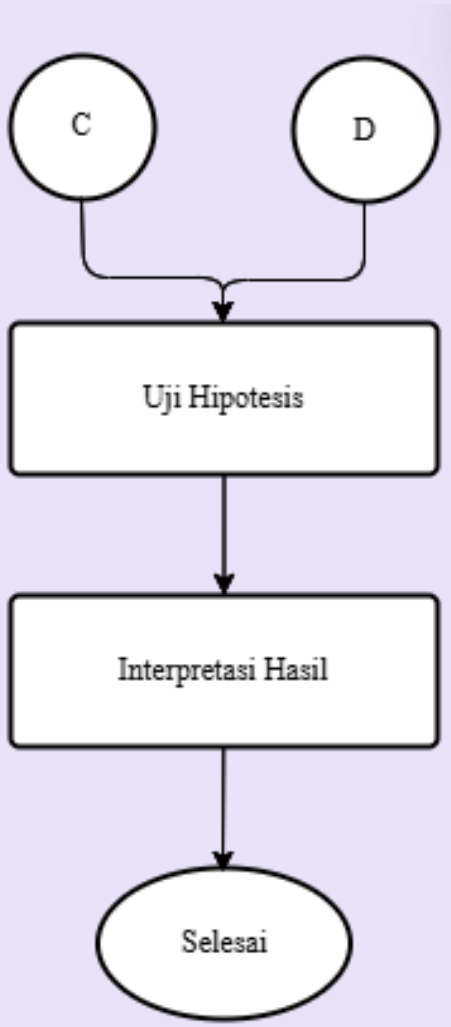
Pelaksana Penelitian

Manhattan



Jaccard





UNIVERSITAS UDAYANA

Thank You

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make.

BAB 2 Tinjauan Pustaka

Klasifikasi

Klasifikasi (Han et al., 2012) merupakan metode pembelajaran terawasi yang bertujuan membangun model (*classifier*) untuk memprediksi label kelas berdasarkan pola dalam data sebelumnya. Proses klasifikasi terdiri dua tahap, yaitu *Learning Phase* dan *Classification Phase*.

Klasifikasi Teks merupakan proses pengelompokan dokumen ke dalam label tertentu berdasarkan isi teks (Muliono & Tanzil, 2018). Metode ini sering digunakan dalam *spam detection*, analisis sentiment, pengelompokan berita dan sistem rekomendasi. Prosesnya meliputi *pre-processing* dan representasi teks.

Klasifikasi Teks

BAB 2 Tinjauan Pustaka

Term Frequency Inverse Document Frequency (TF-IDF)

Term Frequency (TF)

Mengukur seberapa sering kata *i* muncul dalam dokumen *j*

$$TF_{ij} = \frac{f_{ij}}{X_j} \quad (2.1)$$

Inverse Document Frequency (IDF)

Mengukur kelangkaan suatu kata dalam Kumpulan dokumen

$$IDF_i = \log\left(\frac{X}{df_i}\right) \quad (2.2)$$

TF-IDF

$$TF - IDF_{ij} = TF_{ij} \times IDF_i \quad (2.3)$$

BAB 2 Tinjauan Pustaka

K Nearest Neighbor

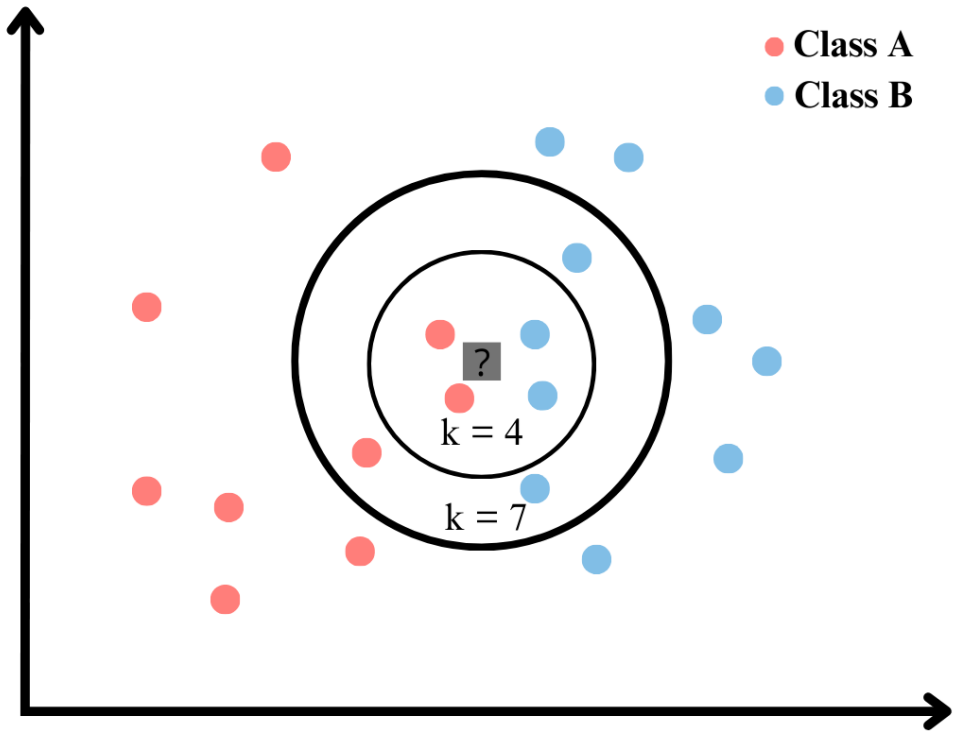
K Nearest Neighbors adalah algoritma non-parametrik berbasis jarak yang digunakan dalam kasus klasifikasi dan regresi. Konsep utama KNN adalah menentukan kelas suatu data berdasarkan mayoritas tetangga terdekat.

Beragam Metrik Jarak

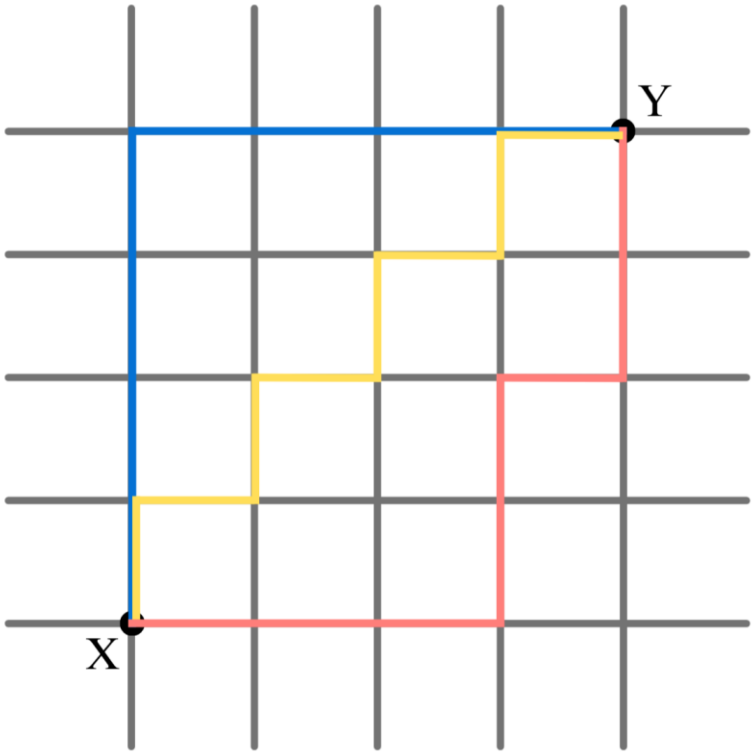
- Jarak Manhattan
- Jarak Jaccard

Pemilihan k:

- Jika k terlalu kecil, rentan terhadap outlier.
- Jika k terlalu besar, prediksi menjadi kurang spesifik.



BAB 2 Tinjauan Pustaka



Jarak Manhattan

Mengukur jarak antara dua titik berdasaekan jumlah perbedaan absolut di setiap dimensi

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

BAB 2 Tinjauan Pustaka

Jarak Jaccard

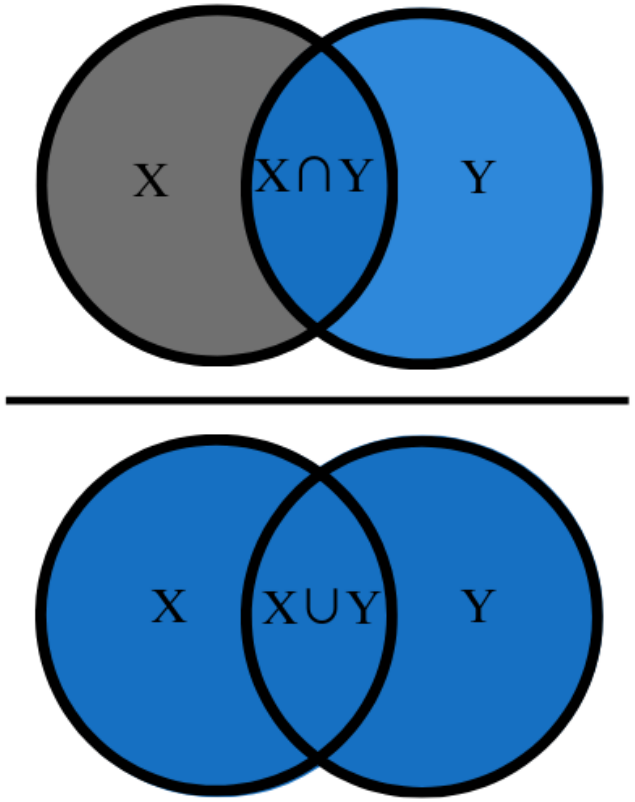
Digunakan untuk mengukur kemiripan aantara dua himpunan kata dalam dokumen

Jaccard Similarity

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

Jaccard Distance

$$d_J(X,Y) = 1 - J(X,Y)$$



BAB 2 Tinjauan Pustaka

Confusion Matrix

Kelas Aktual (y)	Kelas Prediksi $(f(x))$	
	-1	+1
-1	*True Negative* (TN)	*False Positive* (FP)
+1	*False Negative* (FN)	*True Positive* (TP)

Confusion matrix digunakan untuk menilai akurasi model klasifikasi dengan menyajikan hasil prediksi dalam bentuk matriks berukuran $n \times n$.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

BAB 2 Tinjauan Pustaka

Uji Hipotesis

Untuk mengetahui apakah perbedaan akurasi model dalam mengklasifikasi teks signifikan atau hanya kebetulan, digunakan uji t beda berpasangan. Uji ini membandingkan rata-rata perbedaan akurasi dari dua metrik jarak.

1. Menyusun Hipotesis

- H_0 : Tidak terdapat perbedaan yang signifikan antara rata-rata perbedaan akurasi yang dihasilkan oleh kedua metrik jarak ($\mu_d = 0$)
- H_1 : Terdapat perbedaan signifikan antara rata-rata perbedaan akurasi yang dihasilkan oleh kedua metrik jarak. ($\mu_d \neq 0$)

2. Menghitung Statistik Uji

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

3. Membandingkan t-hitung dan t-kritis

- Jika $|t_{hitung}| > t_{kritis}$ atau $p - value < \alpha$, maka H_0 ditolak
- Jika $|t_{hitung}| \leq t_{kritis}$ atau $p - value \geq \alpha$, maka H_0 gagal ditolak