**Selamat pagi/siang, Bapak/Ibu pembimbing serta penguji dan teman-teman semua.**

Perkenalkan, saya **Luh Sukma Mulyani** dengan **NIM 2108541027**. Pada kesempatan ini, saya akan memaparkan proposal tugas akhir saya yang berjudul: **"Komparasi Akurasi Klasifikasi Data Teks Menggunakan Jarak Manhattan dan Jaccard pada Algoritma KNN"**.

Latar belakang penelitian ini berangkat dari pentingnya pemilihan metode klasifikasi dalam analisis data. Dalam proses klasifikasi, hasil yang diperoleh sangat bergantung pada bagaimana algoritma menghitung kedekatan atau jarak antar data. Salah satu algoritma yang menggunakan perhitungan jarak untuk menentukan hasil klasifikasinya adalah algoritma K Nearest Neighbors.

Algoritma K Nearest Neighbors merupakan metode pembelajaran mesin terawasi. KNN bekerja dengan mengasumsikan bahwa data yang berdekatan dalam ruang fitur memiliki karakteristik yang serupa. Algoritma ini menentukan kelas atau nilai baru berdasarkan mayoritas label dari sejumlah tetangga terdekat yang telah diketahui sebelumnya.

Keberhasilan algoritma KNN sangat bergantung pada pemilihan nilai *k* dan metrik jarak yang digunakan terutama pada klasifikasi data teks. Metrik jarak memiliki peran penting dalam menentukan tetangga terdekat, di mana setiap metrik memiliki karakteristik unik yang dapat memengaruhi hasil klasifikasi. Metrik seperti **Euclidean**, **Manhattan**, Minkowski, Cosine dan **Jaccard** sering digunakan karena pendekatan perhitungan jaraknya yang berbeda.

Pemilihan metrik jarak yang sesuai dapat meningkatkan akurasi klasifikasi data teks, sementara metrik yang kurang tepat dapat menurunkan performa model. Oleh karena itu, penelitian ini difokuskan pada perbandingan dua metrik jarak, yaitu **Manhattan** dan **Jaccard**, untuk mengevaluasi mana yang memberikan akurasi terbaik dalam klasifikasi data teks.

Berdasarkan latar belakang tersebut, rumusan masalah yang ingin saya jawab melalui penelitian ini adalah:

Sejalan dengan rumusan masalah tersebut, tujuan dari penelitian ini adalah:

**Batasan Masalah:** Agar penelitian ini tetap terfokus dan terarah, terdapat beberapa batasan masalah sebagai berikut:

Adapun penelitian terdahulu yang menjadi dasar dari penelitian ini.

Selanjutnya, Matode penelitian

Data yang digunakan dalam penelitian ini merupakan **data sekunder** yang diambil dari Kaggle. Dataset ini berisi **448 komentar YouTube** pada video.

Penelitian ini menggunakan dua variabel utama, yaitu:

- **Variabel Dependen (Y):** CLASS (0: HAM, 1: SPAM)
- **Variabel Independen (X):** CONTENT (jumlah kata dalam komentar YouTube)

Adapun tahapan penelitiannya yaitu

1. **Pengumpulan Data:**
    - Mengunduh dataset dari Kaggle yang terdiri dari 448 komentar YouTube.
2. Sebelum analisis, dilakukan **preprocessing** untuk membersihkan data teks. Proses ini meliputi:
3. **Pembobotan TF-IDF**

   Setelah preprocessing, dibentuk **matriks TF-IDF** untuk mewakili bobot kata dalam dokumen. Untuk **jarak Manhattan**, digunakan matriks TF-IDF asli, sedangkan untuk **jarak Jaccard**, matriks TF-IDF diubah menjadi **biner** (nilai 1 jika bobot > 0, dan 0 jika sebaliknya).
4. **Pembagian Data Latih dan Uji (20 Kali)**

   Data dibagi menjadi **80% data latih** dan **20% data uji** secara acak. Proses ini dilakukan sebanyak **20 kali**.

5. **Penerapan Metode KNN**

   Klasifikasi dilakukan menggunakan algoritma **K-Nearest Neighbors (KNN)** dengan dua metrik jarak: **Manhattan** dan **Jaccard**.

   "Langkah pertama dalam algoritma *k-Nearest Neighbors* (k-NN) adalah menentukan nilai **k**, yaitu jumlah tetangga terdekat yang akan digunakan dalam proses prediksi. Selanjutnya, dihitung jarak antara setiap **data baru atau data uji** dengan seluruh data berlabel di **data latih**. Setelah itu, dipilih **k pengamatan** dari data latih yang memiliki jarak terdekat dengan **data baru atau data uji**. Untuk memprediksi nilai dari data tersebut, digunakan **kelas yang paling sering muncul** di antara **k tetangga terdekat** sebagai hasil prediksi"

   Setelah mengimplementasikan algoritma K-Nearest Neighbors (KNN) dan menyelesaikan proses klasifikasi, tahap selanjutnya adalah mengevaluasi performa model menggunakan confusion matrix. Evaluasi ini bertujuan untuk mengukur sejauh mana model mampu mengklasifikasikan data uji secara akurat. Dari hasil evaluasi, diperoleh 200 nilai akurasi untuk masing-masing metrik jarak. Selanjutnya, dilakukan perhitungan rata-rata akurasi pada setiap nilai k untuk menentukan parameter k yang menghasilkan performa terbaik.

   Setelah proses evaluasi model menggunakan confusion matrix, tahap berikutnya adalah menguji apakah terdapat perbedaan signifikan secara statistik antara dua metrik jarak, yaitu Jaccard dan Manhattan, menggunakan **uji-t beda berpasangan**.

   **Langkah-langkah Pengujian:**

1. **Memastikan Asumsi Distribusi Normal** Uji normalitas dilakukan menggunakan **uji Shapiro-Wilk**. Jika hasil uji menunjukkan bahwa data berdistribusi normal, maka dilanjutkan dengan **uji-t beda berpasangan**.
2. **Menentukan Hipotesis Statistik**
3. **Menghitung Statistik Uji** Statistik uji dihitung menggunakan persamaan 2.8
4. **Membandingkan nilai t hitung dan t kritis**

   Dengan mengikuti langkah-langkah di atas, diperoleh kesimpulan mengenai ada atau tidaknya perbedaan signifikan secara statistik antara akurasi yang dihasilkan oleh metrik jarak Manhattan dan Jaccard dalam proses klasifikasi menggunakan algoritma KNN.