**LAPORAN TUGAS KLASIFIKASI DAN KLASTERING**

**MATA KULIAH TEXT MINING & NATURAL LANGUAGE PROCESSING**

Logo

Description automatically generated

Tim Penyusun:

1. <5231811022> < Lathif Ramadhan >
2. <5231811029> < Andini Angel M >
3. <5231811033> < Rama Panji N >
4. <5231811036> < Giffari Riyanda P>

**PROGRAM STUDI SAINS DATA PROGRAM SARJANA**

**FAKULTAS SAINS & TEKNOLOGI**

**UNIVERSITAS TEKNOLOGI YOGYAKARTA**

**2025**

**KLASIFIKASI MENGGUNAKAN DECISION TREE**

**Link Proyek Notebook:** <https://github.com/LatiefDataVisionary/text-mining-and-natural-language-processing-college-task/blob/main/notebooks/decision_tree_model.ipynb>

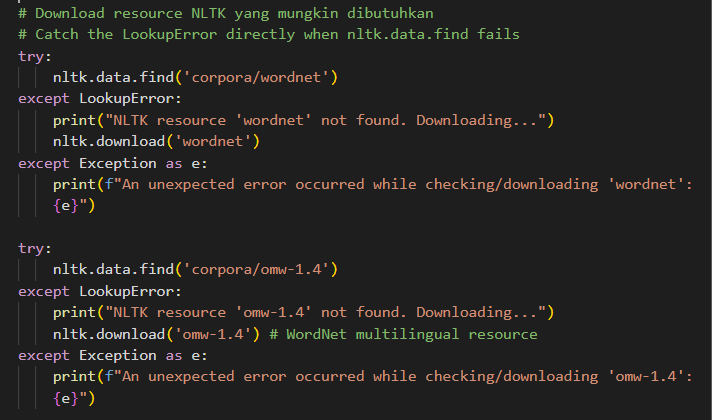
1. **Tulis/Screenshot Import Library yang digunakan**

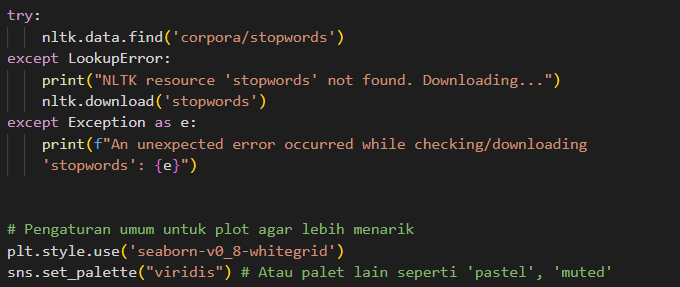
**Jawab:**

Berikut adalah library yang digunakan dalam proyek klasifikasi teks ini. Setiap library memiliki peran penting dalam tahapan analisis data, mulai dari pemrosesan data hingga pemodelan dan evaluasi.

**A computer screen shot of a program code

AI-generated content may be incorrect.**

****

****

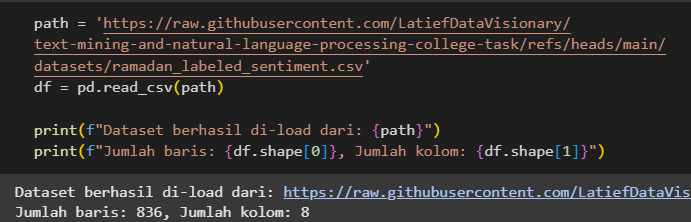
Berikut adalah daftar library Python yang di-import untuk menjalankan analisis sentimen menggunakan model Decision Tree ini. Setiap library memainkan peran spesifik dalam alur kerja machine learning, mulai dari pemuatan data, pra-pemrosesan, pemodelan, hingga evaluasi dan visualisasi.

* pandas (pd): Digunakan untuk manipulasi dan analisis data tabular, terutama untuk memuat dan mengelola dataset.
* numpy (np): Menyediakan dukungan untuk array dan matriks multidimensi besar, bersama dengan kumpulan fungsi matematika tingkat tinggi untuk beroperasi pada array ini.
* sklearn.model\_selection:
  + train\_test\_split: Untuk membagi dataset menjadi set pelatihan dan pengujian.
  + GridSearchCV: Untuk melakukan pencarian hyperparameter terbaik secara sistematis menggunakan cross-validation.
  + cross\_val\_score: Untuk mengevaluasi performa model menggunakan cross-validation.
* sklearn.feature\_extraction.text:
  + TfidfVectorizer: Untuk mengubah koleksi dokumen teks mentah menjadi matriks fitur TF-IDF.
* sklearn.tree:
  + DecisionTreeClassifier: Implementasi algoritma Decision Tree untuk klasifikasi.
  + plot\_tree: Untuk memvisualisasikan pohon keputusan secara langsung menggunakan matplotlib.
  + export\_graphviz: Untuk mengekspor pohon keputusan dalam format DOT, yang kemudian dapat dirender oleh Graphviz.
* sklearn.metrics:
  + accuracy\_score: Untuk menghitung akurasi klasifikasi.
  + confusion\_matrix: Untuk menghitung confusion matrix guna mengevaluasi akurasi klasifikasi.
  + classification\_report: Untuk membangun laporan teks yang menunjukkan metrik klasifikasi utama (precision, recall, F1-score).
  + ConfusionMatrixDisplay: Untuk memvisualisasikan confusion matrix.
  + roc\_auc\_score dan roc\_curve: Untuk mengevaluasi performa model klasifikasi biner menggunakan kurva ROC dan AUC.
* matplotlib.pyplot (plt): Library plotting 2D yang komprehensif untuk membuat visualisasi statis, animasi, dan interaktif.
* seaborn (sns): Library visualisasi data Python berbasis matplotlib yang menyediakan antarmuka tingkat tinggi untuk menggambar grafik statistik yang menarik dan informatif.
* graphviz: Digunakan untuk merender output dari export\_graphviz (file DOT) menjadi representasi grafis dari pohon keputusan.
* nltk (Natural Language Toolkit): Platform utama untuk membangun program Python untuk bekerja dengan data bahasa manusia.
  + WordNetLemmatizer: Untuk melakukan lemmatisasi (mengubah kata ke bentuk dasarnya/lemma).
  + nltk\_stopwords: Menyediakan daftar stop words umum.

1. **Tulis/screenshot import dataset yang digunakan**

**Jawab:**

Dataset yang digunakan dalam analisis sentimen ini adalah ***ramadan\_labeled\_sentiment.csv*** (Dataset yang kami gunakan pada tugas-tugas sebelumnya). Dataset ini berisi tweet yang berkaitan dengan Ramadan beserta label sentimennya (positif atau negatif) dan skor sentimen lainnya.

****

1. **Pemrosesan pembagian (splitdata) Data Training dan Data Testing yang digunakan**

Dataset akan dibagi menjadi dua bagian: data training (untuk melatih model) dan data testing (untuk menguji performa model pada data yang belum pernah dilihat sebelumnya).

1. **Tulis/screenshot codingnya**

**Jawab:**

**A computer screen with colorful text

AI-generated content may be incorrect.**

1. **Tulis/screenshot hasilnya**

**Jawab:**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**4. Pemodelan Decision Tree (perhitungan entropy sampai dengan information gain):**

Model Decision Tree akan dilatih menggunakan data training. Scikit-learn secara internal menangani perhitungan entropy/information gain saat membangun pohon. Kita akan menggunakan GridSearchCV untuk menemukan hyperparameter terbaik untuk model Decision Tree.

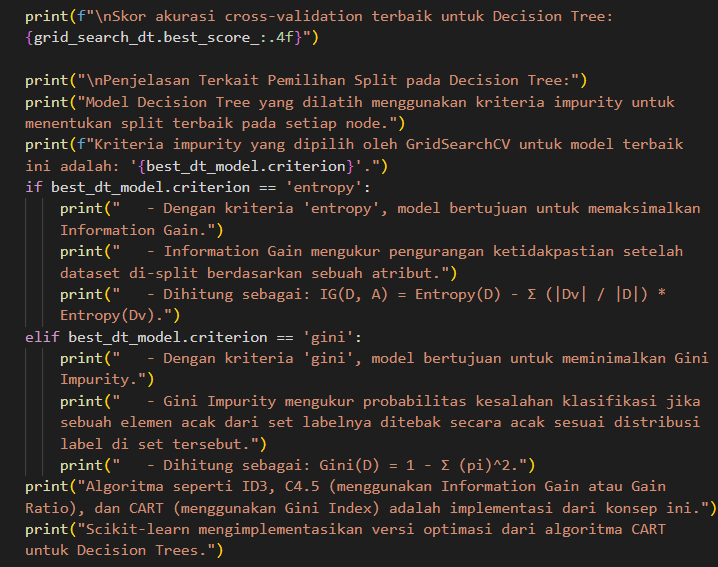
1. **Tuliskan/screenshotkan codingnya**

**Jawab:**

Pelatihan Model dengan Hyperparameter Tuning

**A computer screen shot of a program code

AI-generated content may be incorrect.**

****

1. **Tuliskan/screenshotkan hasil**

**Jawab:**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**5. Tampilkan hasil akurasi dan tabel confusion matrix nya**

Setelah model dilatih, performanya akan dievaluasi menggunakan data testing.

**Jawab:**

**A screen shot of a computer program

AI-generated content may be incorrect.**

**A screenshot of a computer

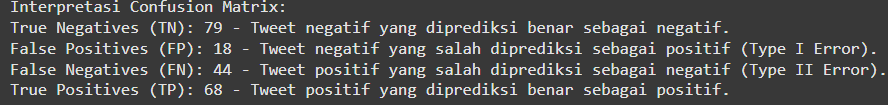
AI-generated content may be incorrect.**

**A screen shot of a computer code

AI-generated content may be incorrect.**

A blue squares with white text

AI-generated content may be incorrect.

****

**6. Tampilkan pohon keputusannya**

**Jawab:**

**A screen shot of a computer program

AI-generated content may be incorrect.**

A diagram of a diagram

AI-generated content may be incorrect.

**Visualisasi dengan graphviz (Untuk Potensi Export):**

**A screen shot of a computer screen

AI-generated content may be incorrect.**

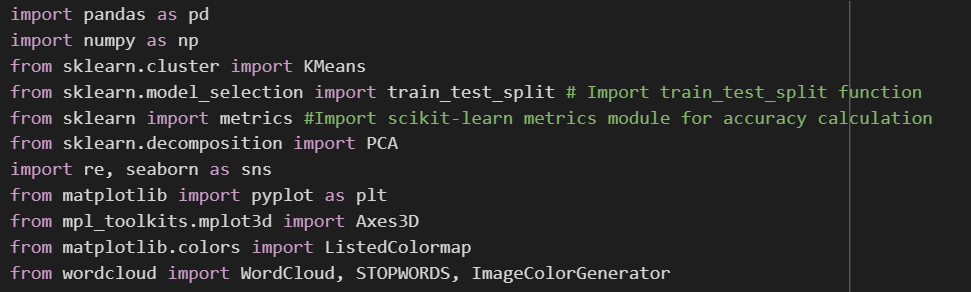
A diagram of a algorithm

AI-generated content may be incorrect.

**KLASTERING MENGGUNAKAN K-MEANS**

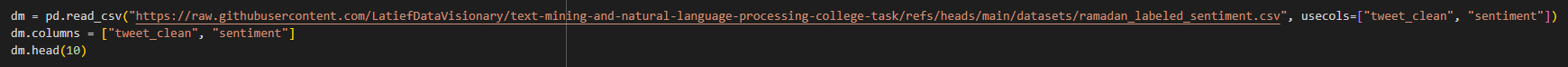
1. **Tulis/Screenshot Import Library yang digunakan**

**Jawab:**

****

1. **Tulis/screenshot import dataset yang digunakan**

**Jawab:**

****

1. **Perhitungan jumlah klaster**
2. **Tulis/screenshot codingnya**

**Jawab:**

**A computer screen shot of a program code

AI-generated content may be incorrect.** **A screen shot of a computer program

AI-generated content may be incorrect.**

1. **Tulis/screenshot hasilnya**

**Jawab:**

A graph with a line

AI-generated content may be incorrect.

1. **Pembuatan grafik dalam bentuk scatter plot berdasarkan centroid**
2. **Tuliskan/screenshotkan codingnya**

**Jawab:**

**A computer screen shot of a computer code

AI-generated content may be incorrect.**

**A computer screen shot of a program

AI-generated content may be incorrect.**

1. **Tuliskan/screenshotkan hasilnya**

**Jawab:**

**A screenshot of a computer screen

AI-generated content may be incorrect.**

A diagram of a graph

AI-generated content may be incorrect.

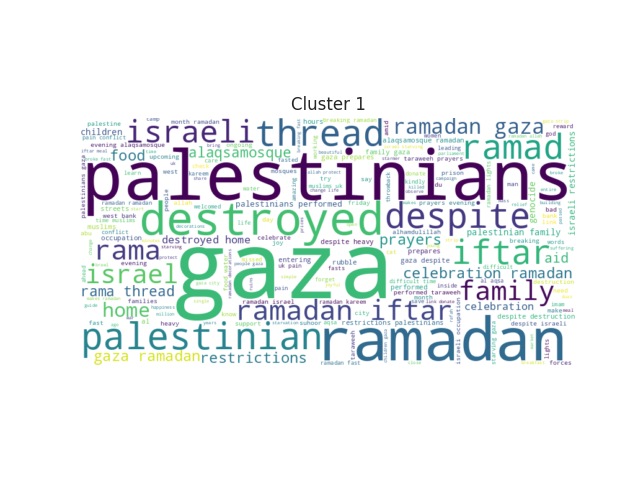
1. **Pembuatan grafik dalam bentuk wordcloud berdasarkan centroid**
2. **Tulis/screenshot codingnya**

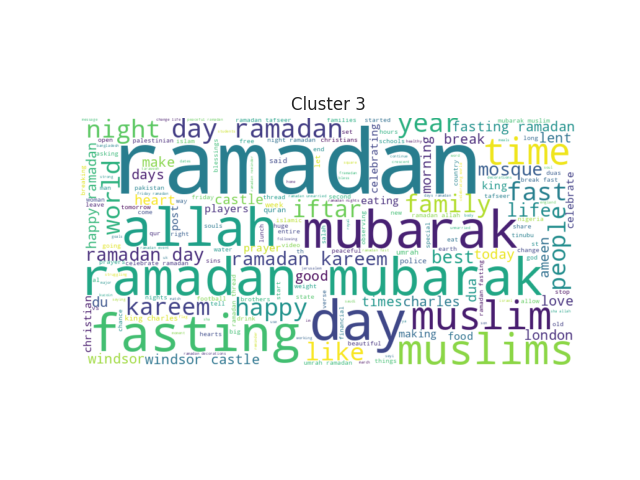
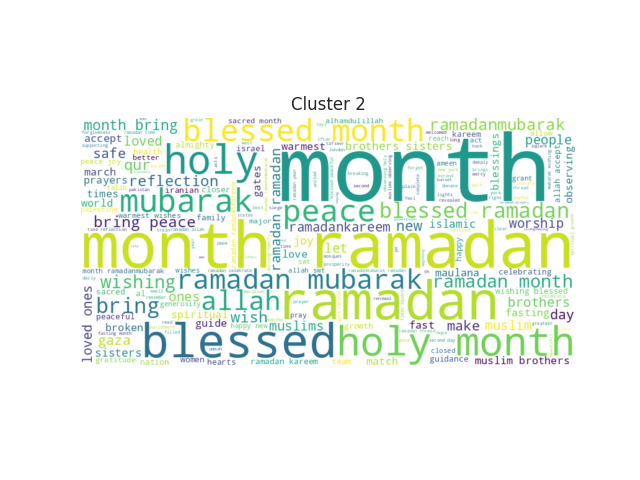
**Jawab:**

****

1. **Tulis/screenshot hasilnya**

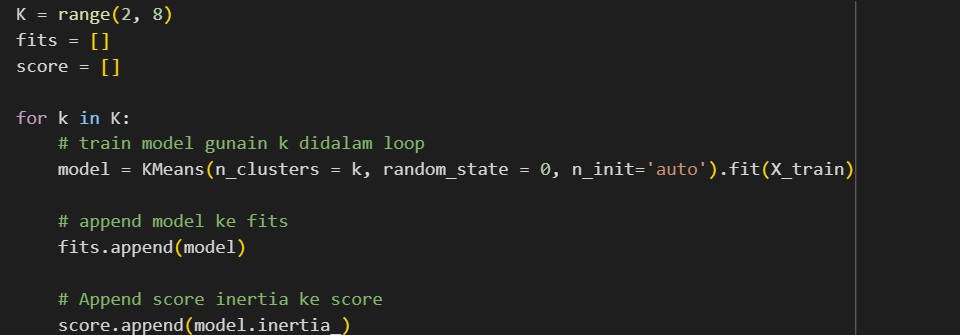
**Jawab:**

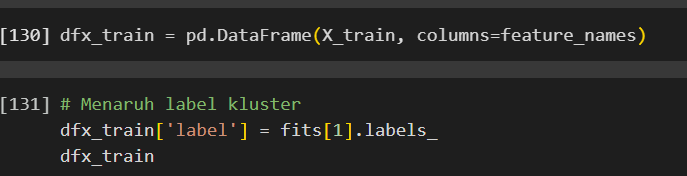
****

****

1. **Hitung klaster masing-masing record data**
2. **Tuliskan/screenshotkan codingnya**

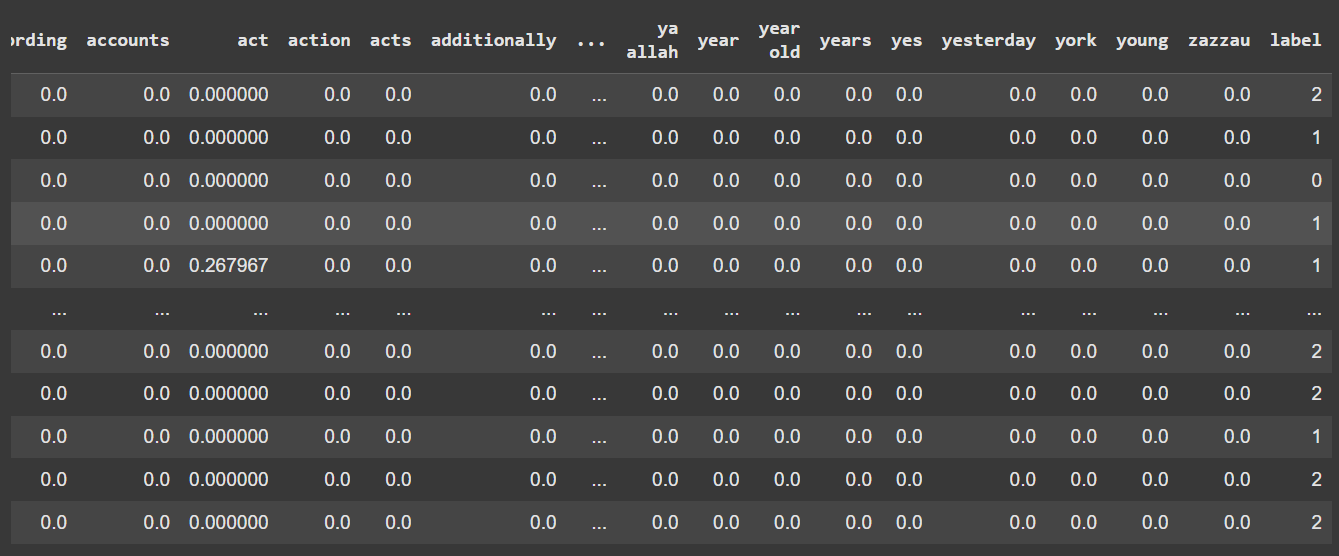
**Jawab:**

****

****

1. **Tuliskan/screenshotkan hasilnya**

**Jawab:**

****