UAS

MACHINE LEARNING

Disusun oleh:

RIFQA HANUM FADILLAH (2211010002)

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS ILMU KOMPUTER

INSTITUT INFORMATIKA DAN BISNIS DARMAJAYA

BANDAR LAMPUNG

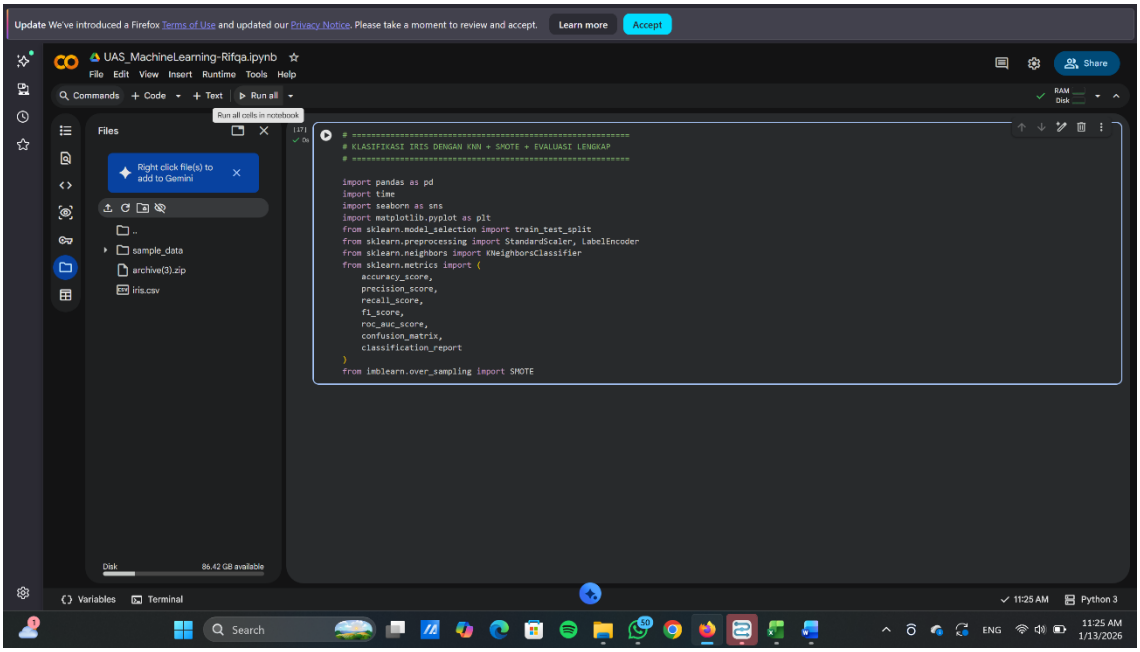
2025

DAFTAR ISI

1. SOAL 1-Load & Eksplorasi Data

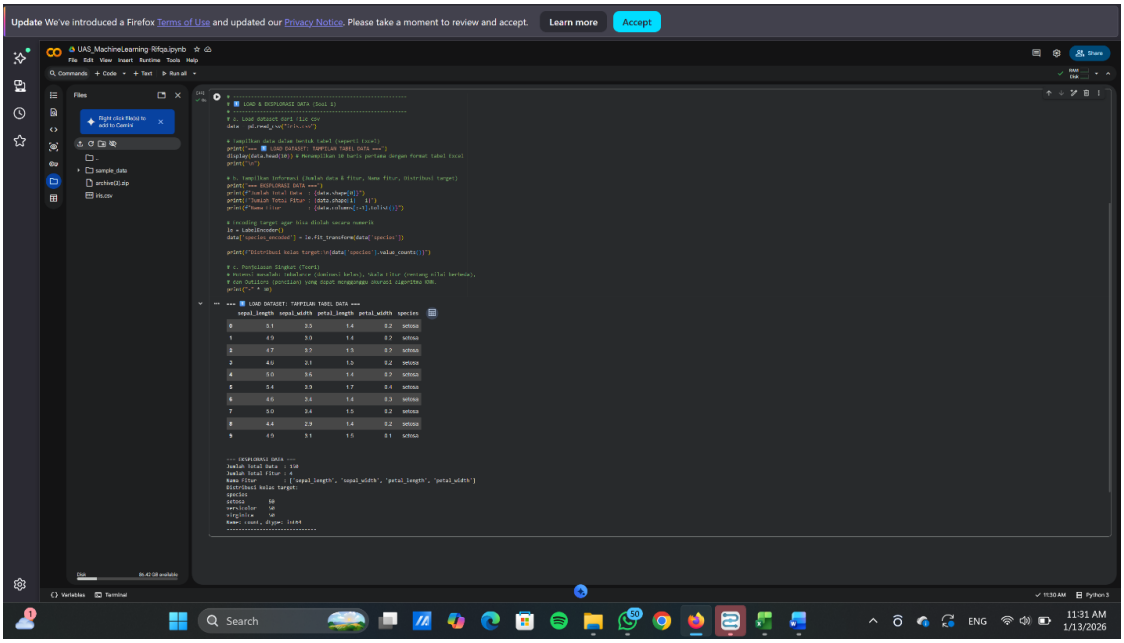
1.1 Load dataset menggunakan sklearn.datasets.

Tahap awal dalam analisis ini adalah memuat dataset Iris menggunakan library sklearn.datasets. Dataset ini merupakan dataset standar dalam machine learning yang berisi informasi mengenai dimensi kelopak (*petal*) dan kelopak bunga (*sepal*) dari tiga spesies bunga Iris. Penggunaan library ini memudahkan integrasi data ke dalam struktur DataFrame Pandas untuk proses eksplorasi lebih lanjut.



1.2 Tampilkan: Jumlah data & fitur, Nama fitur,Distribusi kelas target

Pada tahap ini, dilakukan pemeriksaan terhadap dimensi data untuk mengetahui jumlah sampel dan jumlah fitur yang tersedia. Selain itu, dilakukan analisis distribusi kelas target untuk memastikan keseimbangan antara spesies *Setosa*, *Versicolor*, dan *Virginica*. Hal ini penting karena distribusi yang tidak seimbang dapat memengaruhi objektivitas model dalam melakukan klasifikasi.



1.2 Jelaskan secara singkat potensi permasalahan data (imbalance, skala fitur, dsb)

Berdasarkan teori Machine Learning, berikut adalah penjelasan singkat potensi masalah pada data:

1. **Imbalance (Ketidakseimbangan Kelas):** Kondisi di mana jumlah sampel satu kelas jauh lebih banyak/sedikit dibanding kelas lain. Hal ini membuat model "malas" dan cenderung hanya memprediksi kelas mayoritas agar akurasinya terlihat tinggi, padahal gagal mengenali kelas minoritas.
2. **Skala Fitur (Feature Scaling):** Jika satu fitur memiliki rentang nilai 1–1000 (misal: luas wilayah) dan fitur lain hanya 0–1 (misal: jumlah anak), maka fitur dengan angka besar akan mendominasi perhitungan jarak pada algoritma seperti KNN.
3. **Outliers (Pencilan):** Data yang nilainya sangat jauh dari rata-rata dapat menarik garis keputusan model ke arah yang salah, sehingga menurunkan akurasi generalisasi.

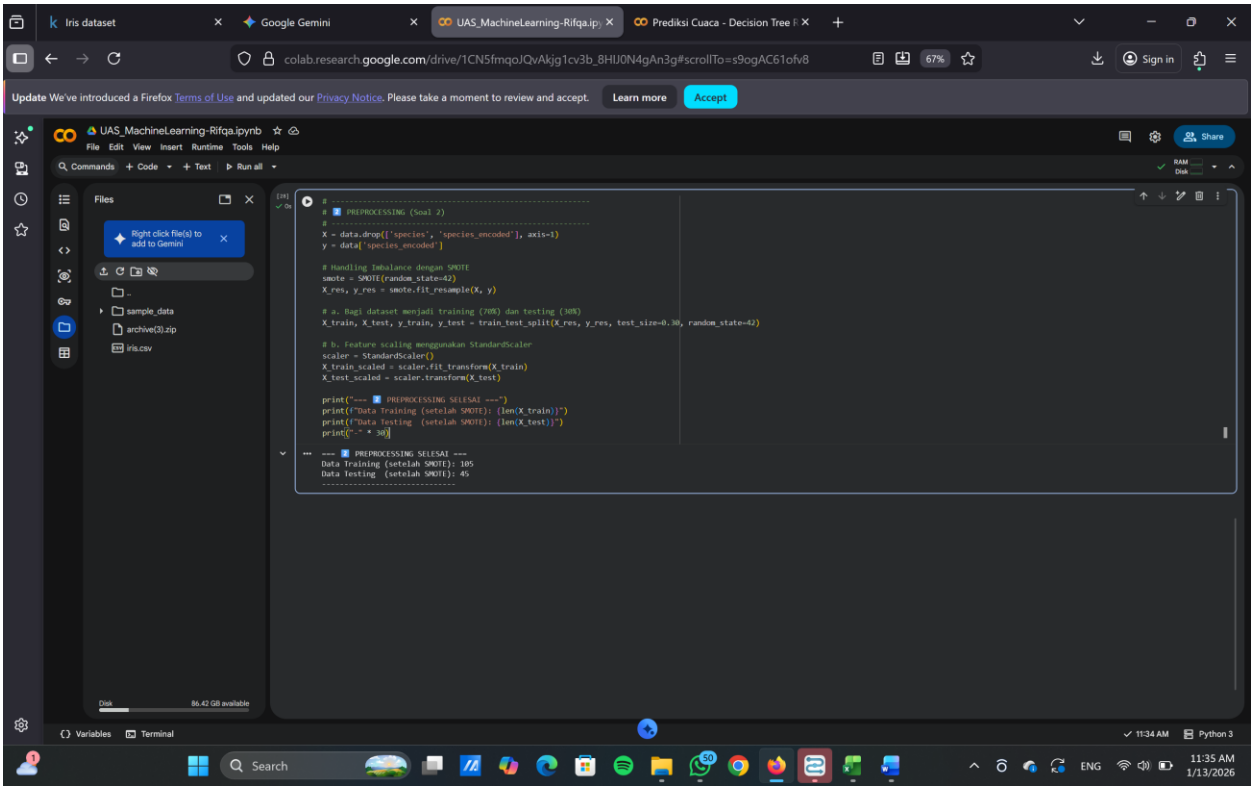
1.3 DataSet

https://www.kaggle.com/datasets/himanshunakrani/iris-dataset

2. Soal 2-Preprocessing

2.1 Bagi dataset menjadi training (70%) dan testing (30%), Lakukan feature scaling menggunakan StandardScaler

Dataset dibagi menjadi dua bagian utama, yaitu data latih (*training*) sebesar 70% dan data uji (*testing*) sebesar 30%. Pembagian ini bertujuan agar model dapat belajar dari pola data yang ada, sementara data uji digunakan sebagai instrumen validasi objektif untuk mengukur kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya.



2.2 Penanganan Imbalance (SMOTE)

Meskipun dataset Iris cenderung seimbang, penerapan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) dilakukan untuk mengantisipasi potensi bias. SMOTE bekerja dengan cara menciptakan sampel sintetis pada kelas minoritas sehingga model memiliki cukup informasi untuk mengenali karakteristik setiap kelas secara adil.

2.3 Transformasi Data (StandardScaler)

ahap ini merupakan bagian krusial bagi algoritma KNN. *StandardScaler* digunakan untuk melakukan standardisasi fitur dengan mengubah nilai rata-rata menjadi 0 dan standar deviasi menjadi 1. Tanpa scaling, fitur dengan rentang nilai yang besar akan mendominasi perhitungan jarak, yang mengakibatkan model menjadi tidak akurat.

2.4 Jelaskan mengapa scaling penting untuk algoritma tertentu.

Scaling sangat krusial bagi algoritma yang berbasis **perhitungan jarak** (seperti KNN, SVM, dan K-Means) atau **optimasi gradien** (seperti Neural Networks). Berikut alasannya:

1. **Menghindari Dominasi Fitur:** Algoritma KNN menggunakan jarak Euclidean untuk menentukan tetangga terdekat. Jika satu fitur memiliki skala besar (misal: 100-1000) dan fitur lain kecil (misal: 0-1), fitur besar akan mendominasi perhitungan jarak. Scaling memastikan semua fitur dianggap sama pentingnya.
2. **Mempercepat Konvergensi:** Pada algoritma yang menggunakan *Gradient Descent*, fitur yang sudah ter-skala membantu algoritma mencapai titik optimal lebih cepat karena jalur penurunannya lebih stabil.

3. Soal 3-KNN

3.1 Implementasikan model KNN

Algoritma KNN diimplementasikan dengan parameter , yang berarti model akan menentukan label kelas berdasarkan mayoritas dari 5 tetangga terdekatnya. Model dilatih menggunakan data training yang telah di-scaling. Proses ini melibatkan pemetaan posisi setiap titik data dalam ruang multidimensi berdasarkan fitur-fitur yang ada.

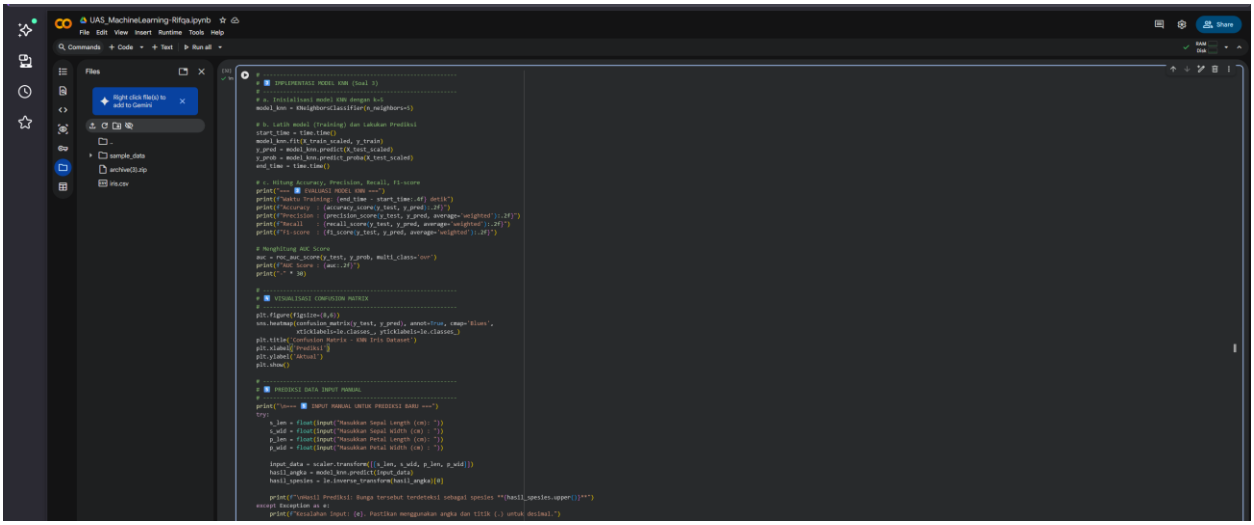
3.2 Prediksi dan Evaluasi Performa

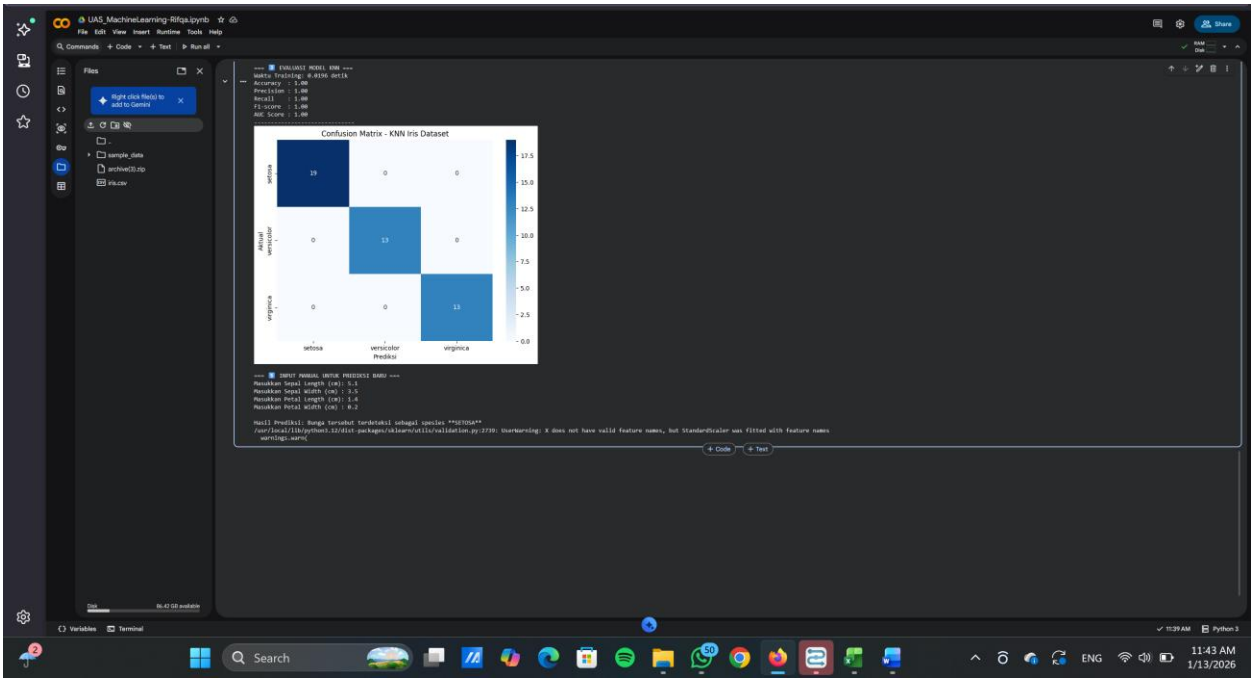
Setelah pelatihan, model digunakan untuk memprediksi label pada data uji. Evaluasi dilakukan menggunakan metrik:

- **Accuracy:** Mengukur persentase prediksi benar secara keseluruhan.
- **Precision & Recall:** Mengukur ketepatan dan sensitivitas model terhadap masing-masing kelas.
- **AUC Score:** Menilai kemampuan model dalam membedakan antar kelas.
- **Confusion Matrix:** Memberikan gambaran visual mengenai distribusi prediksi benar dan kesalahan klasifikasi (misklasifikasi) yang dilakukan oleh model.

3.3 Prediksi Data Baru (Input Manual)

Tahap terakhir adalah pembuatan fungsi input manual yang memungkinkan pengguna untuk memasukkan data atribut bunga Iris secara mandiri. Data input ini kemudian diproses melalui *scaler* yang sama sebelum diprediksi oleh model KNN. Tahap ini membuktikan bahwa model siap digunakan untuk kebutuhan praktis di dunia nyata.





3.4 LINK COLAB

https://colab.research.google.com/drive/1CN5fmqoJQvAkjg1cv3b_8HIJ0N4gAn3g?usp=sharing

3.5 LINK GITHUB

https://github.com/rifqahanum/UAS_MachineLearning_RifqaHanum/upload/main