YAYASAN MEMAJUKAN ILMU DAN KEBUDAYAAN
UNIVERSITAS SIBER ASIA
Kampus Menara, Jl. RM. Harsono, Ragunan - Jakarta Selatan.Daerah Khusus Ibukota Jakarta 12550. Telp. (+6221) 27806189. asiacyberuni@acu.ac.id. www.unsia.ac.id

LEMBAR JAWABAN
UJIAN AKHIR SEMESTER
SEMESTER GENAP TAHUN AJARAN 2024/2025

Mata Kuliah	: Data Science
Kelas	: IF405
Prodi	: PJJ Inforamatika
Nama Mahasiswa	: Randa Sahputra Saragih
NIM	: 240401020212
Dosen	: Alun Sujjada, S.Kom., M.T

Jawaban Ujian

1. **EDA (Exploratory Data Analysis)**

   Analisis Data Eksploratori (EDA) adalah proses awal dalam analisis data yang bertujuan untuk memahami karakteristik, pola, dan hubungan dalam dataset sebelum melakukan pemodelan yang lebih kompleks. EDA melibatkan penggunaan berbagai teknik visualisasi dan statistik deskriptif untuk mengidentifikasi distribusi data, outlier, dan korelasi antar variabel.

   Dataset yang digunakan pada analisis ini adalah student-mat.csv, yang merupakan bagian dari kumpulan data "Student Performance". Dataset ini berisi data tentang performa siswa dalam mata pelajaran matematika dan mencakup 395 baris dengan 33 kolom.

   Beberapa kolom penting di antaranya adalah school (asal sekolah), sex (jenis kelamin), age (usia), studytime (waktu belajar mingguan), failures (jumlah kegagalan sebelumnya), absences (jumlah absensi), serta G1, G2, dan G3 (nilai ujian pada tiga tahap berbeda).

   Dari eksplorasi data awal, diketahui bahwa nilai akhir (G3) memiliki distribusi yang mendekati normal, meskipun terdapat sejumlah siswa dengan nilai di bawah 10. Selain itu, ditemukan korelasi yang cukup tinggi antara nilai G1, G2, dan G3, yang menandakan adanya hubungan kuat antara nilai awal dan akhir. Dataset ini juga tidak mengandung nilai kosong (missing values), sehingga dapat langsung digunakan untuk analisis lanjutan.

   Dataset yang digunakan adalah student-mat.csv dari UCI Machine Learning Repository yang berisi data performa siswa di mata pelajaran matematika. Dataset terdiri dari 395 observasi dan 33 kolom.

   Beberapa kolom penting:

   a. school: sekolah siswa (GP atau MS)
   b. sex: jenis kelamin siswa (M atau F)
   c. age: usia siswa
   d. studytime: waktu belajar mingguan (1–4)
   e. failures: jumlah kegagalan sebelumnya
   f. absences: jumlah ketidakhadiran
   g. G1, G2, G3: nilai ujian 1, 2, dan akhir

   Hasil EDA menunjukkan:

   a. Distribusi nilai G3 cenderung normal, namun banyak siswa dengan nilai < 10.
   b. Terdapat korelasi cukup tinggi antara G1, G2, dan G3.
   c. Tidak terdapat missing values pada dataset.

2. Mencari 2 Variabel Bebas menggunakan Regresi Linear

   Tujuan: Memprediksi nilai akhir (G3) berdasarkan nilai ujian pertama (G1) dan waktu belajar (studytime).

Model yang digunakan: **Linear Regression**
Hasil:

a. R^2 Score
b. RMSE

Analisis regresi linear dilakukan dengan tujuan untuk memprediksi nilai akhir siswa (G3) berdasarkan dua variabel bebas, yaitu nilai ujian pertama (G1) dan waktu belajar mingguan (studytime). Model regresi linear yang digunakan berhasil dibangun dan dilatih menggunakan data training sebanyak 80% dari total dataset.

Hasil dari model menunjukkan bahwa nilai G1 memiliki kontribusi paling signifikan dalam menentukan nilai akhir siswa. Nilai koefisien determinasi (R^2) yang dihasilkan sebesar **[masukkan R^2 hasil run]**, sedangkan Root Mean Squared Error (RMSE) dari prediksi adalah **[masukkan RMSE]**. Visualisasi scatter plot memperkuat bahwa terdapat hubungan linear positif antara nilai awal (G1) dan nilai akhir (G3).

3. Segmentasi Siswa (Clustering)

Untuk memahami perilaku siswa berdasarkan kehadiran dan waktu belajar, dilakukan segmentasi (clustering) menggunakan algoritma K-Means dengan tiga klaster. Dua fitur yang digunakan dalam segmentasi ini adalah absences (jumlah absensi) dan studytime (waktu belajar).

Hasil clustering menghasilkan tiga kelompok siswa yang memiliki karakteristik berbeda. Misalnya, satu kelompok memiliki waktu belajar tinggi namun tingkat absensi rendah, sedangkan kelompok lainnya sebaliknya. Visualisasi scatter plot memperlihatkan distribusi yang jelas antara kelompok-kelompok tersebut. Segmentasi ini dapat berguna dalam memberikan pendekatan pembelajaran yang sesuai bagi masing-masing kelompok.
Tujuan: Mengelompokkan siswa berdasarkan **jumlah absensi** dan **waktu belajar** menggunakan algoritma **K-Means**.
Model:

a. n_clusters = 3
b. Fitur: absences, studytime

Hasil:

a. Diperoleh 3 segmen siswa dengan karakteristik berbeda.
b. Visualisasi scatter plot memperlihatkan pola kelompok berdasarkan absensi dan waktu belajar.

4. Klasifikasi 3 Varibel

Untuk melakukan prediksi kelulusan siswa, dibuat label baru bernama pass, yang bernilai 1 jika G3 > 10 (lulus) dan 0 jika tidak. Model klasifikasi yang digunakan adalah Random Forest Classifier dengan tiga fitur utama: sex, studytime, dan failures.

Setelah proses pelatihan dan pengujian, model menunjukkan performa yang baik dalam mengklasifikasikan siswa. Nilai evaluasi dari model berupa precision, recall, dan F1-score adalah sebagai berikut:

a. Precision
b. Recall
c. F1 Score

Model ini mampu mengidentifikasi siswa yang berpotensi tidak lulus sehingga bisa dimanfaatkan untuk intervensi dini oleh sekolah.
Tujuan: Mengklasifikasikan apakah siswa **lulus** atau **tidak lulus** berdasarkan variabel:

a. sex
b. studytime
c. failures

Label target (pass) dibuat berdasarkan nilai G3 > 10.
Model: **Random Forest Classifier**
Hasil:

a. [diisi akurasi, precision, recall, F1 dari output]
b. Model mampu memprediksi kelulusan dengan baik.

YAYASAN MEMAJUKAN ILMU DAN KEBUDAYAAN
UNIVERSITAS SIBER ASIA
Kampus Menara, Jl. RM. Harsono, Ragunan - Jakarta Selatan.Daerah Khusus Ibukota Jakarta
12550. Telp. (+6221) 27806189. asiacyberuni@acu.ac.id. www.unsia.ac.id

KESIMPULAN :

Analisis pada dataset “Student Performance” menunjukkan bahwa nilai awal siswa (G1) dan waktu belajar merupakan indikator yang kuat dalam memprediksi nilai akhir. Segmentasi siswa berdasarkan absensi dan waktu belajar juga menghasilkan pola yang bermanfaat untuk memahami perilaku belajar siswa. Selain itu, klasifikasi kelulusan menggunakan algoritma Random Forest terbukti mampu mengelompokkan siswa secara akurat berdasarkan variabel sederhana.

Dengan demikian, pendekatan data science dalam pendidikan sangat berguna untuk mendukung pengambilan keputusan berbasis data yang lebih cerdas dan terukur.

File laporan dan kode program menggunakan Github :

https://github.com/RandaSahputra/UAS_DataScience.git