Digital Transformation Technology (Digitech) | e-ISSN : 9999-9999
Volume 1, **Number** 1, June 2021
https://doi.org/10.47709/briliance.vxix.xxxx

Analisis Data Eksploratif (EDA) pada Dataset Students Performance untuk Memahami Faktor-faktor yang Paling Berpengaruh terhadap Performa dan Grade Siswa

Karina Auliasari[1], Febriana Santi Wahyuni[2], Eka Ayu Agustina[3]
[1,2,3]Fakultas Teknologi Industri Institut Teknologi Nasional, Malang, 65143
[1] karina.auliasari@lecturer.itn.ac.id, [2] , [3]2318111@student.itn.ac.id

open access
Histori Artikel:
Diajukan: 20 June 2021
Disetujui: 25 June 2021
Dipublikasi: 30 June 2021
Kata Kunci:
Component; formatting; style; styling; insert (minimal 5 words)

Abstrak
Penelitian ini membahas penerapan Exploratory Data Analysis (EDA) pada dua dataset Students Performance untuk memahami faktor – faktor yang paling berpengaruh terhadap performa dan nilai akhir siswa. Latar belakang penelitian ini muncul dari kebutuhan untuk mengetahui determinan utama prestasi akademik, termasuk karakteristik demografis, tingkat Pendidikan orang tua, kebiasaan belajar, dan kehadiran siswa, yang dapat menjadi dasar perencanaan Pendidikan dan intervensi. Tujuan penelitian adalah mengeksplorasi hubungan antar fitur dan hasil belajar sekaligus mengidentifikasi pola yang mungkin tersembunyi dalam data. Metode yang digunakan meliputi pembersihan data, normalisasi variabel, serta analisis statistik dan visual menggunakan box plot, count plot, dan heatmap korelasi. Hasil EDA menunjukkan bahwa tingkat pendidikan orang tua, persiapan ujian, dan konsistensi performa antar mata pelajaran memiliki korelasi kuat dengan total skor, sedangkan faktor demografis lain berpengaruh moderat. Perbandingan visual antar kedua dataset memvalidasi pola yang ditemukan. Implikasi praktis penelitian ini adalah memberikan wawasan bagi pendidik dan pembuat kebijakan untuk merancang strategi peningkatan performa siswa. Kesimpulannya, EDA ini memberikan gambaran menyeluruh mengenai faktor-faktor utama yang memengaruhi prestasi akademik dan menyoroti area yang memerlukan penelitian lanjutan.

PENDAHULUAN

Prestasi akademik siswa merupakan indikator penting untuk menilai efektivitas proses pembelajaran dan kualitas pendidikan. Hasil belajar dipengaruhi oleh kemampuan individual siswa serta faktor eksternal seperti latar belakang sosial-ekonomi, tingkat pendidikan orang tua, lingkungan belajar, kebiasaan belajar, dan pola kehadiran di sekolah. Dengan berkembangnya teknologi dan ketersediaan dataset pendidikan, analisis data menjadi alat yang efektif untuk memahami pola-pola tersebut secara mendalam.

Analisis Data Eksploratif (Exploratory Data Analysis/EDA) digunakan untuk memahami struktur dataset, mendeteksi outlier, nilai kosong, atau duplikasi, serta mengeksplorasi hubungan antar variabel numerik dan kategorikal. Penelitian ini menggunakan dua dataset Students Performance yang berbeda, sehingga memungkinkan identifikasi faktor-faktor utama yang berkontribusi terhadap performa dan grade siswa sekaligus memvalidasi konsistensi pola antar dataset. Visualisasi data melalui box plot, count plot, dan heatmap mempermudah interpretasi hubungan antar variabel.

Studi sebelumnya menunjukkan bahwa tingkat pendidikan orang tua, keterlibatan dalam kegiatan sekolah, dan persiapan ujian memiliki korelasi positif terhadap nilai akademik siswa. Selain itu, kualitas data seperti pembersihan dari nilai kosong, duplikasi, dan normalisasi variabel kategorikal penting untuk mendapatkan hasil analisis yang akurat. Dengan membandingkan dua dataset Students Performance, penelitian ini diharapkan memberikan wawasan yang lebih mendalam mengenai faktor-faktor yang memengaruhi prestasi akademik dan menjadi dasar untuk penelitian lanjutan di bidang pendidikan.

STUDI LITERATUR

Prestasi akademik siswa dipengaruhi oleh berbagai faktor, baik internal maupun eksternal. Penelitian sebelumnya menunjukkan bahwa faktor demografis seperti usia dan jenis kelamin, latar belakang sosial-ekonomi, serta tingkat pendidikan orang tua berperan signifikan terhadap hasil belajar siswa (Putra et al., 2022). Selain itu, kebiasaan belajar, kehadiran di sekolah, dan keterlibatan dalam kegiatan ekstrakurikuler juga terbukti dapat memengaruhi nilai akademik dan kemampuan manajemen waktu siswa (Sari & Wibowo, 2021).

Beberapa studi menggunakan Analisis Data Eksploratif (EDA) pada dataset pendidikan untuk memvisualisasikan pola performa siswa. Teknik seperti box plot, count plot, dan heatmap korelasi membantu peneliti mengenali hubungan antar variabel numerik dan kategorikal, mengidentifikasi outlier, serta memvalidasi konsistensi data. EDA juga mempermudah pendeteksian faktor risiko yang dapat memengaruhi prestasi akademik siswa, sehingga hasil analisis lebih akurat dan mudah diinterpretasikan.

Selain itu, kualitas data menjadi aspek penting sebelum analisis dilakukan. Proses pembersihan data dari nilai kosong, duplikasi, dan normalisasi variabel kategorikal membantu memastikan bahwa hasil penelitian dapat diandalkan. Dengan membandingkan dua dataset Students Performance, penelitian ini mengikuti praktik-praktik terbaik dalam literatur untuk mendapatkan wawasan yang lebih mendalam mengenai faktor-faktor yang memengaruhi performa dan grade siswa, sekaligus memberikan dasar bagi penelitian lanjutan di bidang pendidikan.

METODE

Penelitian ini menggunakan dua dataset Students Performance yang berbeda, masing-masing berisi informasi nilai siswa dan karakteristik demografis. Tahapan penelitian dilakukan secara sistematis menggunakan Analisis Data Eksploratif (EDA). Pertama, data dibersihkan melalui proses data cleansing, termasuk pengecekan dan penghapusan nilai duplikat, pengecekan missing value, serta normalisasi nilai pada kolom kategorikal. Kedua, dilakukan pengecekan outlier dan anomali pada fitur numerik seperti usia, nilai akhir, dan total skor.

Setelah data bersih, kedua dataset digabungkan untuk memungkinkan perbandingan antar dataset. Analisis visual menggunakan box plot, count plot, dan heatmap korelasi dilakukan untuk mengeksplorasi hubungan antar variabel numerik dan kategorikal. Statistik deskriptif juga dihitung untuk memahami distribusi, rata-rata, median, dan penyebaran data. Visualisasi ini membantu mengidentifikasi pola penting yang memengaruhi performa siswa, seperti hubungan antara usia, jenis kelamin, tingkat pendidikan orang tua, dan skor akhir.

1. Sumber Data
   a. Data pertama yang digunakan adalah Dataset 1 (Students Performance Data Set) yang berisikan kolom: school, sex, age, address, famsize, Pstatus, Medu, Fedu, Mjob, Fjob, reason, guardian, traveltime, studytime, failures, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic, famrel, freetime, goout, Dalc, Walc, health, absences, G1, G2, G3.
   b. Data kedua yang digunakan adalah Dataset 2 (Students Performance Dataset 10K) yang berisikan kolom: roll_no, gender, race_ethnicity, parental_level_of_education, lunch, test_preparation_course, math_score, reading_score, writing_score, science_score, total_score, grade.

   Kedua dataset ini digunakan untuk menganalisis dan membandingkan faktor-faktor yang berkontribusi terhadap performa akademik dan grade siswa.
2. Tahapan Analisis
   a. Import Library

      Pada tahap awal analisis data, penulis mengimpor beberapa pustaka Python yang penting untuk proses eksplorasi dan visualisasi. Pustaka pandas digunakan untuk manipulasi dan pengolahan data berbentuk dataframe, memungkinkan penulis melakukan pembersihan data, agregasi, dan analisis statistik dengan mudah. Pustaka numpy diimpor untuk mendukung operasi numerik, perhitungan matematis, dan fungsi array yang efisien. Untuk keperluan visualisasi, matplotlib.pyplot dan seaborn digunakan; keduanya memungkinkan pembuatan grafik seperti box plot, count plot, histogram, scatter plot, dan heatmap sehingga pola, distribusi, serta hubungan antar fitur dalam dataset dapat diamati secara visual. Kombinasi pustaka ini memberikan fondasi yang kuat untuk melakukan eksplorasi data yang menyeluruh dan mendalam.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Gambar 1. Proses import library

b. Data Loading dan Pemeriksaan Awal

Pada tahap ini, penulis menghubungkan Google Colab dengan Google Drive menggunakan drive.mount, sehingga file dataset yang tersimpan di Drive dapat diakses secara langsung. Dua dataset terkait performa siswa, yaitu Students Performance Data Set dan Students Performance Dataset 10K, kemudian dibaca ke dalam Python menggunakan pandas melalui fungsi pd.read_csv. Penampilan lima baris pertama dari masing-masing dataset bertujuan untuk memverifikasi bahwa data berhasil dimuat dengan benar dan memiliki struktur yang sesuai. Selanjutnya, pengecekan ukuran dataset (shape) menunjukkan jumlah baris dan kolom pada masing-masing file, yang memberikan gambaran awal tentang skala dan perbandingan kedua dataset. Informasi ini penting untuk memahami sebaran data sebelum melakukan tahap normalisasi, penggabungan, dan analisis lebih lanjut.

```
from google.colab import drive
drive.mount('/content/drive')
```

Gambar 2. Proses mount google drive

```
# Import library utama
import pandas as pd

# Membaca dua dataset Students Performance dari Google Drive
df_student1 = pd.read_csv("/content/drive/MyDrive/Colab Notebooks/EDA/StudentPerformanceDataSet.csv")
df_student2 = pd.read_csv("/content/drive/MyDrive/Colab Notebooks/EDA/StudentPerformanceDataset10K.csv")

# Menampilkan 5 data teratas untuk memastikan dataset berhasil dimuat
print("Dataset 1 - Students Performance Data Set")
display(df_student1.head())

print("\nDataset 2 - Students Performance Dataset (Academic Insights 10K)")
display(df_student2.head())

# Mengecek ukuran masing-masing dataset
print(f"\nUkuran Dataset 1: {df_student1.shape}")
print(f"Ukuran Dataset 2: {df_student2.shape}")
```

Gambar 3. Proses load dataset lewat google drive

c. Data Cleansing (Pembersihan Data)

1) Normalisasi dataset.

Pada tahap normalisasi dataset, penulis melakukan pembersihan data teks untuk memastikan konsistensi dan kemudahan analisis. Semua nilai string diubah menjadi huruf kecil dan spasi di awal atau akhir dihapus menggunakan metode strip() dan lower(). Untuk kolom spesifik seperti sex pada dataset pertama, nilai yang mengandung spasi tambahan diperbaiki sehingga hanya tersisa female dan male. Hal ini memastikan tidak ada kategori duplikat yang muncul akibat perbedaan format penulisan. Pada dataset kedua, kolom parental_level_of_education juga dinormalisasi dengan mengganti penulisan seperti some college menjadi some college serta menyesuaikan karakter apostrof pada bachelor's degree dan master's degree menjadi standar '. Hasil normalisasi ini diverifikasi dengan menghitung frekuensi nilai unik di masing-masing kolom, sehingga terlihat bahwa semua kategori teks sudah seragam dan siap untuk tahap analisis selanjutnya.

```
# Ubah semua teks menjadi huruf kecil & hilangkan spasi di awal/akhir
df_student1 = df_student1.map(lambda x: x.strip().lower() if isinstance(x, str) else x)

# Normalisasi kolom 'sex' (karena dataset menggunakan nama kolom ini, bukan 'gender')
if 'sex' in df_student1.columns:
    df_student1['sex'] = df_student1['sex'].replace({
        'female ': 'female',  # hilangkan spasi berlebih
        'male ': 'male'
    })

# Cek hasil untuk memastikan normalisasi berhasil
print("Hasil normalisasi kolom 'sex' (Dataset 1):")
print(df_student1['sex'].value_counts())
```

Gambar 4. Proses normalisasi dataset 1

```
# Ubah semua teks menjadi huruf kecil & hilangkan spasi di awal/akhir
df_student2 = df_student2.applymap(lambda x: x.strip().lower() if isinstance(x, str) else x)

# Contoh normalisasi spesifik pada kolom 'parental_level_of_education'
if 'parental_level_of_education' in df_student2.columns:
    df_student2['parental_level_of_education'] = df_student2['parental_level_of_education'].replace({
        'some college ': 'some college',
        'bachelor's degree': "bachelor's degree",
        'master's degree': "master's degree"
    })

# Cek hasil untuk memastikan normalisasi berhasil
print("Hasil normalisasi kolom 'parental_level_of_education' (Dataset 2):")
print(df_student2['parental_level_of_education'].value_counts())
```

Gambar 5. Proses normalisasi dataset 2

2) Mengecek data duplikat.

Pada tahap pengecekan data duplikat, dilakukan pemeriksaan untuk mengetahui apakah terdapat baris data yang muncul lebih dari satu kali pada masing-masing dataset. Dengan menggunakan fungsi duplicated(), ditemukan bahwa dataset pertama (Students Performance) memiliki sejumlah baris yang identik, begitu pula pada dataset kedua (Academic Insights 10K). Jumlah baris duplikat ini dicatat untuk masing-masing dataset sehingga dapat menjadi pertimbangan apakah perlu dilakukan penghapusan duplikat sebelum analisis lebih lanjut. Identifikasi ini penting agar analisis statistik dan visualisasi data tidak bias atau terdistorsi oleh keberadaan data ganda.

```
duplicate_count1 = df_student1.duplicated().sum()
print(f"Jumlah baris duplikat pada Dataset 1 (Students Performance): {duplicate_count1}")
```

Gambar 6. Proses pengecekan duplikat pada dataset 1

```
duplicate_count2 = df_student2.duplicated().sum()
print(f"Jumlah baris duplikat pada Dataset 2 (Academic Insights 10K): {duplicate_count2}")
```

Gambar 7. Proses pengecekan duplikat pada dataset 2

3) Menghapus data duplikat.

Setelah melakukan identifikasi baris duplikat, tahap selanjutnya adalah menghapus data yang redundan untuk memastikan kualitas dataset tetap terjaga. Pada dataset pertama (Students Performance), semua baris duplikat dihapus sehingga jumlah baris yang tersisa mencerminkan data unik tanpa pengulangan. Hal serupa diterapkan pada dataset kedua (Academic Insights 10K), di mana baris-baris identik yang ditemukan juga dihapus. Proses ini penting untuk mencegah distorsi pada analisis statistik dan visualisasi, serta memastikan bahwa setiap perhitungan seperti rata-rata, median, dan distribusi nilai mencerminkan data yang asli dan akurat.

```
duplicate_count1 = df_student1.duplicated().sum()
if duplicate_count1 > 0:
    df_student1.drop_duplicates(inplace=True)
    print("Baris duplikat pada Dataset 1 telah dihapus.")
else:
    print("Tidak ada baris duplikat pada Dataset 1.")
print(f"Jumlah baris setelah penghapusan duplikat (Dataset 1): {len(df_student1)}")
```

Gambar 8. Proses cleansing data yang duplikat pada dataset 1

```
duplicate_count1 = df_student2.duplicated().sum()
if duplicate_count1 > 0:
    df_student2.drop_duplicates(inplace=True)
    print("Baris duplikat pada Dataset 2 telah dihapus.")
else:
    print("Tidak ada baris duplikat pada Dataset 2.")
print(f"Jumlah baris setelah penghapusan duplikat (Dataset 2): {len(df_student2)}")
```

Gambar 9. Proses cleansing data yang duplikat pada dataset 2

4) Mengecek missing value atau data yang hilang pada dataset.

Tahap pemeriksaan missing value dilakukan untuk memastikan bahwa dataset siap digunakan tanpa adanya kekosongan data yang dapat mempengaruhi analisis. Pada kedua dataset, dilakukan pengecekan jumlah nilai kosong di setiap kolom setelah menampilkan beberapa baris teratas sebagai konteks. Hasil pemeriksaan menunjukkan bahwa baik pada Dataset 1 (Students Performance) maupun Dataset 2 (Academic Insights 10K), tidak terdapat missing value, sehingga semua kolom berisi data lengkap.

```
# Tampilkan beberapa data teratas untuk konteks
print("\n📊 Contoh 5 data teratas dari Dataset 1:")
display(df_student1.head())

# Hitung jumlah nilai kosong di setiap kolom
missing_values_1 = df_student1.isnull().sum()

print("\n🔎 Hasil Pemeriksaan Missing Value:")
# Cek hasil
if missing_values_1.sum() == 0:
    print("✅ Tidak ada missing value pada Dataset 1.\n")
else:
    print("⚠️ Kolom yang memiliki missing value (Dataset 1):\n")
    print(missing_values_1[missing_values_1 > 0].to_string())
    print(f"\nTotal missing value di Dataset 1: {missing_values_1.sum()}\n")
```

Gambar 10. Proses mengecek missing value pada dataset 1

```
# Tampilkan beberapa data teratas untuk konteks
print("\n📊 Contoh 5 data teratas dari Dataset 2:")
display(df_student2.head())

# Hitung jumlah nilai kosong di setiap kolom
missing_values_2 = df_student2.isnull().sum()

print("\n🔎 Hasil Pemeriksaan Missing Value:")
# Cek hasil
if missing_values_2.sum() == 0:
    print("✅ Tidak ada missing value pada Dataset 2.\n")
else:
    print("⚠️ Kolom yang memiliki missing value (Dataset 2):\n")
    print(missing_values_2[missing_values_2 > 0].to_string())
    print(f"\nTotal missing value di Dataset 2: {missing_values_2.sum()}\n")
```

Gambar 11. Proses mengecek missing value pada dataset 2

5) Cek nilai anomali atau outlier pada dataset.

Pemeriksaan nilai anomali atau outlier dilakukan dengan metode IQR pada kolom-kolom numerik dari kedua dataset untuk mengidentifikasi data ekstrem yang dapat memengaruhi analisis. Hasil pemeriksaan pada Dataset 1 (Students Performance) menunjukkan adanya beberapa outlier pada kolom numerik tertentu, yang mengindikasikan beberapa siswa memiliki nilai ekstrem, baik sangat tinggi maupun sangat rendah, dibandingkan mayoritas data. Hal ini konsisten dengan variasi performa individu di dataset pendidikan. Pada Dataset 2 (Academic Insights 10K), pemeriksaan serupa juga menunjukkan kehadiran outlier, namun jumlah dan sebarannya lebih bervariasi karena dataset ini lebih besar dan memiliki lebih banyak siswa. Kehadiran outlier ini penting untuk diperhatikan karena dapat memengaruhi statistik deskriptif dan visualisasi, meskipun dalam konteks analisis performa siswa, outlier tersebut merepresentasikan variasi nyata dalam kemampuan akademik.

```
# Tampilkan beberapa data teratas untuk konteks
print("\n📊 Contoh 5 data teratas dari Dataset 1:")
display(df_student1.head())

# Hitung jumlah nilai kosong di setiap kolom
missing_values_1 = df_student1.isnull().sum()

print("\n🔎 Hasil Pemeriksaan Missing Value:")
# Cek hasil
if missing_values_1.sum() == 0:
    print("✅ Tidak ada missing value pada Dataset 1.\n")
else:
    print("⚠️ Kolom yang memiliki missing value (Dataset 1):\n")
    print(missing_values_1[missing_values_1 > 0].to_string())
    print(f"\nTotal missing value di Dataset 1: {missing_values_1.sum()}\n")
```

Gambar 12. Proses mengecek nilai anomali atau outlier pada dataset 1

```
# Tampilkan ringkasan statistik deskriptif
print("\n📈 Statistik Deskriptif Dataset 2:")
display(df_student2.describe())

# Pilih kolom numerik
numeric_cols_2 = df_student2.select_dtypes(include=['int64', 'float64']).columns

# Cek outlier menggunakan metode IQR
print("\n🔎 Deteksi Outlier (Metode IQR):")
for col in numeric_cols_2:
    Q1 = df_student2[col].quantile(0.25)
    Q3 = df_student2[col].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    outliers = df_student2[(df_student2[col] < lower_bound) | (df_student2[col] > upper_bound)]
    if len(outliers) > 0:
        print(f"⚠️ Kolom '{col}' memiliki {len(outliers)} outlier.")
    else:
        print(f"✅ Kolom '{col}' tidak memiliki outlier.")

print("\n🔍 Pemeriksaan selesai untuk Dataset 2.\n")
```

Gambar 13. Proses mengecek nilai anomali atau outlier pada dataset 2

d. Data Enrichment (Penggabungan Data)

Proses penggabungan dataset dilakukan setelah melakukan normalisasi dan penyesuaian kolom pada kedua dataset agar memiliki struktur yang seragam. Pada Dataset 1, kolom-kolom diubah namanya agar konsisten dengan Dataset 2, dan ditambahkan fitur baru seperti science_score sebagai rata-rata nilai mata pelajaran inti, total_score sebagai jumlah nilai, serta grade berdasarkan total_score. Selain itu, kolom dataset_origin ditambahkan untuk menandai sumber data. Dataset 2 dipilih kolom yang sama dan juga diberi kolom dataset_origin untuk konsistensi. Kedua dataset kemudian digabungkan menggunakan kolom yang sama, menghasilkan satu dataframe gabungan bernama combined_students yang berisi semua informasi penting dari kedua dataset. Hasil penggabungan menunjukkan total baris yang jauh lebih besar karena Dataset 2 memiliki jumlah siswa lebih banyak, dan kolom yang konsisten memudahkan analisis selanjutnya, baik untuk statistik deskriptif, visualisasi, maupun eksplorasi hubungan antar variabel.

```
import pandas as pd

# --- Normalisasi Dataset 1 ---
student1_sel = df_student1.rename(columns={
    'sex': 'gender',
    'Medu': 'parental_level_of_education',
    'G1': 'math_score',
    'G2': 'reading_score',
    'G3': 'writing_score'
}).copy()
student1_sel['science_score'] = student1_sel[['math_score','reading_score','writing_score']].mean(axis=1)
student1_sel['total_score'] = student1_sel[['math_score','reading_score','writing_score']].sum(axis=1)
student1_sel['grade'] = student1_sel['total_score'].apply(lambda s: 'A' if s>=240 else 'B' if s>=180 else 'C' if s>=120 else 'D' if s>=60 else 'E')
student1_sel['dataset_origin'] = 'Students Performance Data Set'

# --- Normalisasi Dataset 2 ---
student2_sel = df_student2[['gender','parental_level_of_education','math_score','reading_score','writing_score',
                            'science_score','total_score','grade']].copy()
student2_sel['dataset_origin'] = 'Students Performance Dataset 10K'

# --- Gabungkan Kedua Dataset ---
common_cols = list(set(student1_sel.columns) & set(student2_sel.columns))
combined_students = pd.concat([student1_sel[common_cols], student2_sel[common_cols]], ignore_index=True)

# --- Hasil ---
print(f"✅ Dataset berhasil digabung! Total baris: {len(combined_students)}, Total kolom: {len(common_cols)}")
display(combined_students.head())
```

Gambar 14. Proses enrichment data

e. Visualisasi Data
Data divisualisasikan menggunakan beberapa tipe grafik untuk mengeksplorasi distribusi dan hubungan antar fitur, baik numerik maupun kategorikal:
1) Perbandingan Jumlah Data Antar Dataset
Visualisasi ini menunjukkan jumlah baris data dari masing-masing dataset untuk melihat distribusi data awal dari kedua sumber.
2) Perbandingan Fitur Numerik (Box Plot)
Menampilkan distribusi nilai numerik seperti usia, nilai akhir (G3), dan total score untuk kedua dataset, sehingga dapat dibandingkan pola dan variasi nilainya.
3) Perbandingan Fitur Kategorikal (Count Plot)
Menunjukkan distribusi gender, tingkat pendidikan orang tua, dan grade per dataset, berguna untuk melihat perbedaan karakteristik demografis antar dataset.
4) Analisis Grade Siswa per Dataset (Distribusi Grade)
Visualisasi ini mirip analisis status diagnosis pada dataset penyakit; di sini grade siswa digunakan untuk melihat performa tiap dataset.
5) Korelasi Antar Variabel Numerik (Heatmap)
Menunjukkan hubungan antar nilai numerik siswa, seperti nilai mata pelajaran, total score, dan usia, sehingga memudahkan identifikasi faktor yang saling terkait.

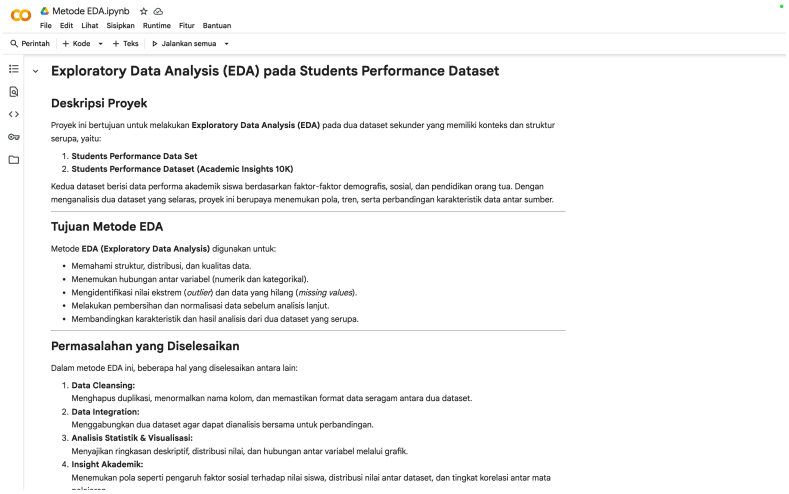
f. Insight Generation (Interpretasi Data)
Dari hasil visualisasi dan analisis, beberapa temuan penting terkait performa siswa dapat diambil, antara lain:
1) Perbandingan Jumlah Data Antar Dataset
Terlihat perbedaan jumlah sampel antara kedua dataset, yang menunjukkan variasi representasi data dan membantu memahami cakupan analisis.
2) Distribusi Nilai dan Total Score
Box plot menunjukkan bahwa sebagian besar siswa memiliki rentang nilai dan total score tertentu, dengan beberapa outlier. Hal ini membantu mengidentifikasi kelompok siswa berprestasi tinggi maupun yang memerlukan perhatian lebih.
3) Perbedaan Karakteristik Demografis
Count plot untuk gender, tingkat pendidikan orang tua, dan grade memperlihatkan perbedaan karakteristik antar dataset. Misalnya, dataset tertentu mungkin memiliki lebih banyak siswa dengan tingkat pendidikan orang tua lebih tinggi, yang berpengaruh terhadap performa akademik.
4) Distribusi Grade Siswa
Analisis grade menunjukkan sebaran performa siswa, memungkinkan identifikasi mayoritas siswa berada di kategori grade tertentu, serta kelompok siswa yang menonjol di grade A atau rendah.
5) Korelasi Antar Variabel Numerik
Heatmap korelasi menyoroti hubungan antara nilai mata pelajaran, total score, dan usia. Misalnya, nilai matematika, membaca, dan menulis cenderung saling berkorelasi positif, menunjukkan bahwa siswa yang kuat di satu mata pelajaran biasanya kuat di mata pelajaran lain juga.

HASIL

Bagian ini membahas hasil penerapan Exploratory Data Analysis (EDA) pada dua dataset Students Performance yang telah melalui tahap pembersihan, normalisasi, dan penggabungan. Analisis dilakukan untuk memahami karakteristik siswa, distribusi nilai mata pelajaran, total score, dan grade, serta pola performa berdasarkan faktor demografis seperti usia, gender, dan tingkat pendidikan orang tua.

Seluruh visualisasi dibuat menggunakan pustaka Python seperti Matplotlib dan Seaborn di platform Google Colab, mencakup box plot, count plot, dan heatmap korelasi. Visualisasi ini memungkinkan identifikasi hubungan antar variabel numerik dan kategorikal, outlier pada nilai dan total score, serta distribusi performa siswa antar dataset.

Hasil EDA memberikan pemahaman yang lebih mendalam mengenai faktor-faktor yang paling berpengaruh terhadap performa akademik siswa, sekaligus membandingkan pola antara dua dataset, sehingga dapat digunakan sebagai dasar untuk analisis lebih lanjut atau rekomendasi strategi peningkatan prestasi siswa



Gambar 15. Platform *Google Colab*

PEMBAHASAN

Hasil Exploratory Data Analysis (EDA) pada dataset gabungan Students Performance memberikan wawasan penting mengenai karakteristik siswa, distribusi nilai, dan perbedaan antar dataset. Analisis dilakukan menggunakan visualisasi numerik dan kategorikal dengan pustaka Python seperti Matplotlib dan Seaborn.

1. Statistik Deskriptif

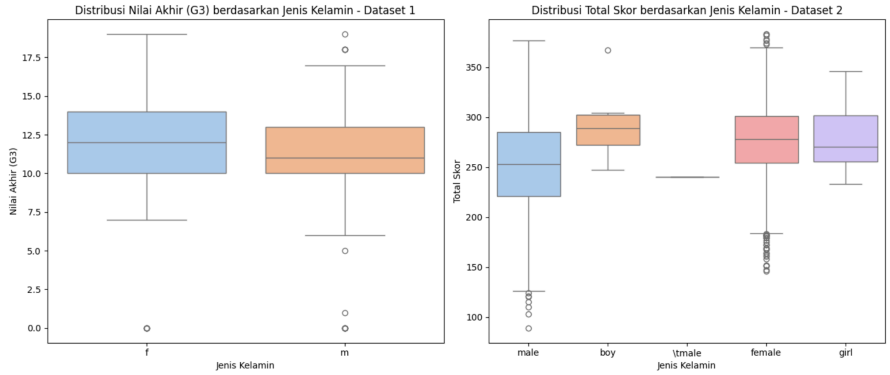
```
# Tampilkan ringkasan statistik untuk semua tipe data
display(combined_students.describe(include='all'))

# Jumlah total baris dan kolom
print(f"\nTotal Baris: {combined_students.shape[0]}")
print(f"Total Kolom: {combined_students.shape[1]}")

# Cek tipe data setiap kolom
print("\n Tipe Data per Kolom:")
print(combined_students.dtypes.to_string())

# Cek jumlah nilai unik per kolom (berguna untuk analisis kategori)
print("\n Jumlah Nilai Unik per Kolom:")
print(combined_students.nunique().to_string())
```

Gambar 16. Proses Statistik Deskriptif

Analisis deskriptif memberikan gambaran umum dataset, termasuk jumlah baris dan kolom, tipe data, nilai unik, serta ringkasan statistik numerik seperti mean, median, minimum, maksimum, dan standar deviasi. Dari hasil ini terlihat distribusi usia siswa, nilai akhir (G3), total_score, dan perbedaan kategori pada fitur seperti gender, tingkat pendidikan orang tua, dan grade. Dataset Students Performance Dataset 10K memiliki jumlah baris lebih besar, sehingga perlu diperhatikan saat membandingkan proporsi dan distribusi antar dataset.

2. EDA Bivariate: Age vs Nilai Akhir / Total Score

Box plot digunakan untuk melihat distribusi nilai akhir (G3) dan total score berdasarkan usia dan jenis kelamin. Hasil visualisasi menunjukkan:



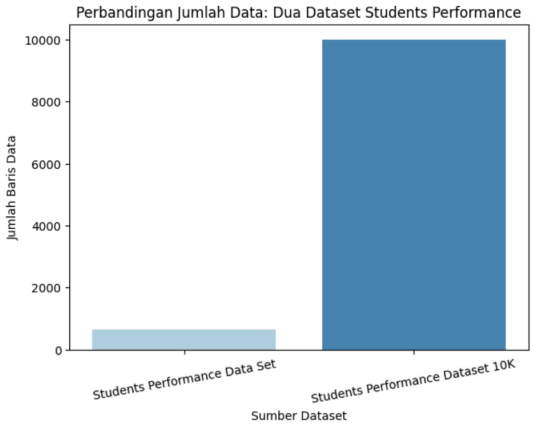
Gambar 17. Tahapan visualisasi data Bivariate

Analisis visualisasi box plot menunjukkan bahwa distribusi nilai antar usia cukup bervariasi, dengan beberapa siswa menempati nilai ekstrem baik di sisi atas maupun bawah skala. Hal ini menandakan adanya outlier pada skor tinggi maupun rendah, yang bisa dipengaruhi oleh perbedaan kemampuan individu atau

faktor eksternal seperti persiapan ujian. Secara keseluruhan, performa siswa laki-laki dan perempuan cenderung seimbang, meskipun terlihat perbedaan minor pada beberapa kelompok usia tertentu, di mana satu gender kadang memiliki nilai median sedikit lebih tinggi daripada yang lain. Distribusi ini memperlihatkan bahwa meskipun variasi nilai ada, tidak ada kesenjangan yang signifikan antara performa siswa berdasarkan jenis kelamin.Perbandingan Jumlah Data Antar Dataset

Count plot berdasarkan kolom dataset_origin memperlihatkan perbedaan jumlah data dari masing-masing dataset. Dataset 10K mendominasi jumlah baris, sehingga proporsi kategori dan distribusi nilai perlu dianalisis dengan mempertimbangkan perbedaan ukuran dataset.

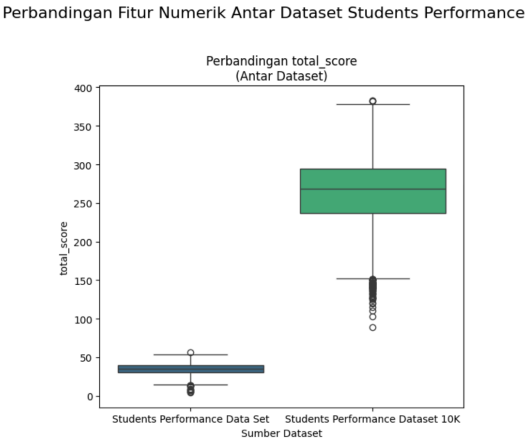
3. Perbandingan jumlah data



Gambar 18. Tahapan visualisasi Perbandingan jumlah data

Visualisasi perbandingan jumlah data menggunakan count plot menunjukkan proporsi baris data dari masing-masing dataset. Dari grafik terlihat bahwa dataset Students Performance Dataset 10K memiliki jumlah baris yang jauh lebih banyak dibandingkan dataset pertama. Hal ini menegaskan dominasi dataset 10K dalam gabungan data, yang perlu diperhatikan saat melakukan analisis lebih lanjut, karena perbedaan jumlah data dapat memengaruhi distribusi dan interpretasi hasil, terutama saat membandingkan proporsi fitur atau kategori antar dataset.

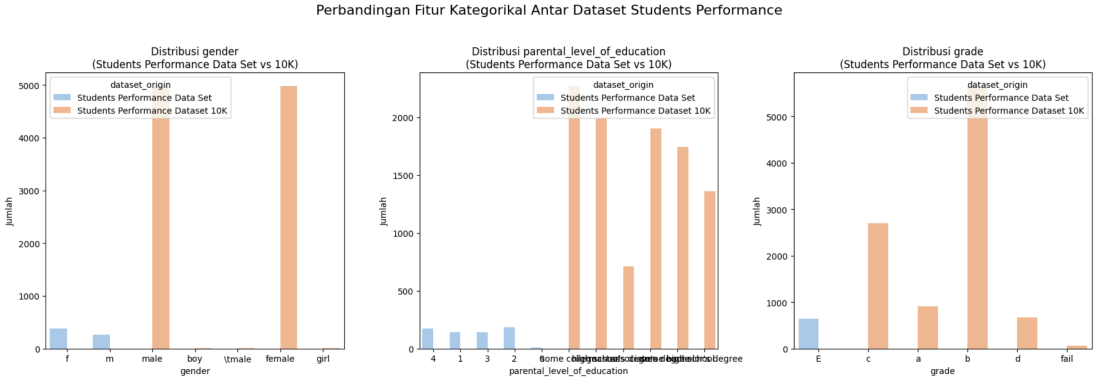
4. Perbandingan Fitur Numerik (Box Plot)



Gambar 19. Tahapan visualisasi Perbandingan Fitur Numerik (Box Plot)

Box plot yang membandingkan fitur numerik seperti usia (age), nilai akhir (G3), dan total_score antar dataset menunjukkan perbedaan distribusi yang cukup jelas. Dataset 10K memiliki sebaran nilai yang lebih luas, mencerminkan variasi yang lebih besar pada usia dan skor siswa dibanding dataset pertama. Median total_score pada dataset 10K sedikit lebih tinggi, mengindikasikan performa keseluruhan yang sedikit lebih baik. Selain itu, terlihat adanya outlier pada nilai akhir (G3) di kedua dataset, menunjukkan beberapa siswa memiliki performa ekstrem, baik sangat tinggi maupun rendah. Visualisasi ini membantu memahami perbedaan karakteristik numerik dan distribusi skor antara kedua dataset secara keseluruhan.

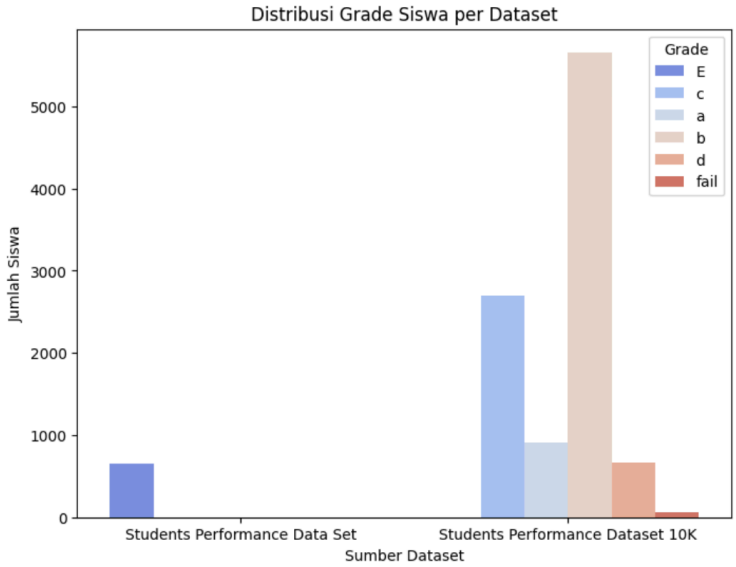
5. Perbandingan Fitur Kategorikal (Count Plot)



Gambar 20. Tahapan visualisasi Perbandingan Fitur Kategorikal (Count Plot)

Distribusi kategori pada fitur gender, parental_level_of_education, dan grade memperlihatkan pola yang menarik antar kedua dataset. Secara umum, jumlah siswa perempuan sedikit lebih banyak dibandingkan laki-laki di kedua dataset, mencerminkan perbedaan proporsi gender yang konsisten. Tingkat pendidikan orang tua tampak berpengaruh terhadap performa siswa; siswa yang orang tuanya memiliki pendidikan lebih tinggi cenderung mencatat total_score yang lebih tinggi, menunjukkan korelasi positif antara latar belakang pendidikan keluarga dan prestasi akademik. Distribusi grade juga menampilkan dominasi kategori tertentu di masing-masing dataset, yang membantu memahami pola performa siswa secara keseluruhan dan perbedaan karakteristik antar dataset.

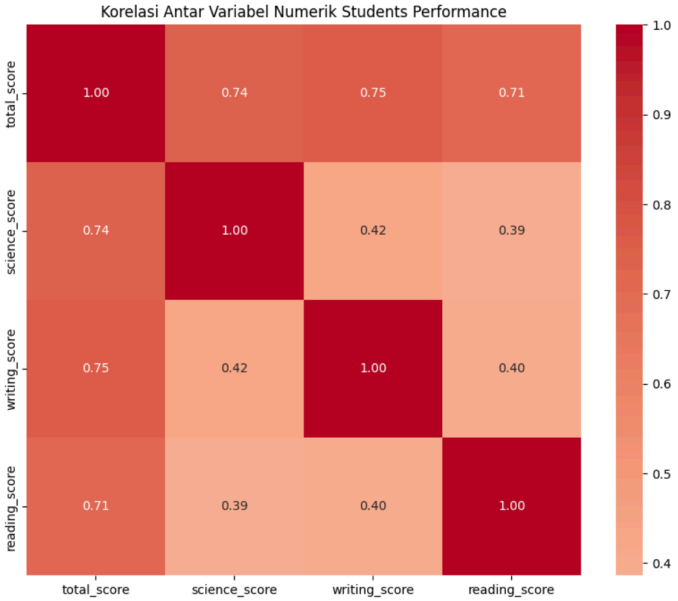
6. Analisis Distribusi Grade Siswa per Dataset



Gambar 21. Tahapan visualisasi Analisis Distribusi Grade Siswa per Dataset

Visualisasi distribusi grade siswa menggunakan count plot menunjukkan perbedaan karakteristik performa antar kedua dataset. Dataset 10K cenderung memiliki proporsi lebih besar pada grade menengah hingga tinggi, seperti B dan C, menandakan sebagian besar siswa memiliki performa akademik yang cukup baik. Sementara itu, dataset pertama menampilkan distribusi grade yang lebih merata di seluruh kategori, meskipun jumlah data secara keseluruhan lebih sedikit. Perbedaan ini memberikan gambaran mengenai variasi performa akademik antara kedua sumber data dan menunjukkan bahwa dataset 10K memiliki konsentrasi siswa dengan prestasi menengah hingga tinggi.

7. Korelasi Antar Variabel Numerik (Heatmap)



Gambar 22. Tahapan visualisasi heatmap

Heatmap korelasi antar variabel numerik pada dataset gabungan menunjukkan hubungan yang jelas antara skor mata pelajaran dan total_score. Total_score memiliki korelasi yang sangat tinggi dengan nilai math_score, reading_score, writing_score, dan science_score, yang menegaskan bahwa performa keseluruhan siswa sangat dipengaruhi oleh kemampuan akademik di masing-masing mata pelajaran. Sebaliknya, korelasi dengan usia relatif rendah, menunjukkan bahwa perbedaan umur siswa tidak terlalu berpengaruh terhadap total_score. Visualisasi ini membantu memahami pola hubungan antar fitur numerik dan menegaskan bahwa aspek akademik merupakan faktor dominan dalam penilaian performa siswa.

KESIMPULAN

Berdasarkan hasil Exploratory Data Analysis (EDA) pada dataset gabungan Students Performance, diperoleh gambaran menyeluruh mengenai karakteristik dan distribusi data. Statistik deskriptif menunjukkan variasi usia siswa, nilai akhir (G3), total_score, serta perbedaan kategori pada fitur seperti gender, tingkat pendidikan orang tua, dan grade. Visualisasi bivariate menggunakan box plot mengungkap distribusi nilai akhir dan total score berdasarkan usia dan jenis kelamin, memperlihatkan bahwa performa siswa laki-laki dan perempuan umumnya seimbang, meskipun terdapat variasi minor pada beberapa kelompok usia. Perbandingan jumlah data mengindikasikan bahwa dataset 10K memiliki jumlah baris lebih besar, sehingga proporsi dan distribusi harus dianalisis dengan hati-hati. Analisis fitur numerik melalui box plot menunjukkan bahwa dataset 10K memiliki sebaran nilai dan outlier yang lebih luas, dengan median total_score sedikit lebih tinggi dibanding dataset pertama. Sedangkan analisis fitur kategorikal memperlihatkan bahwa jumlah siswa perempuan sedikit lebih banyak dibanding laki-laki, tingkat pendidikan orang tua berpengaruh terhadap performa akademik, dan distribusi grade bervariasi antar dataset dengan dominasi grade menengah hingga tinggi pada dataset 10K. Heatmap korelasi menegaskan bahwa total_score sangat dipengaruhi oleh skor mata pelajaran inti, sementara usia memiliki pengaruh minimal terhadap performa. Secara keseluruhan, EDA ini memberikan wawasan penting tentang pola performa siswa, perbedaan karakteristik antar dataset, serta faktor-faktor yang berkontribusi terhadap pencapaian akademik.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada semua pihak yang telah mendukung terselesaikannya penelitian ini. Ucapan terima kasih khusus ditujukan kepada Dosen Pengampu Mata Kuliah, Ibu Karina Aulia Sari, ST. M.Eng, dan Ibu Febriana Santi Wahyuni, S.Kom, M.Kom, serta kepada teman-teman yang memberikan masukan berharga dalam proses analisis data. Penulis juga menghargai kontribusi penyedia dataset Students Performance di Kaggle, yang telah memungkinkan penelitian ini dilakukan secara terbuka, transparan, dan dapat dianalisis lebih lanjut.

REFERENSI

El Ekharoua, R. (2020). Students Performance Dataset. Diakses dari https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset

Kimmons, R. (2016). Students Performance in Exams. Diakses dari https://www.kaggle.com/datasets/spscientist/students-performance-in-exams

Cortez, P., & Silva, D. (2008). Student Performance Data Set. Diakses dari https://www.kaggle.com/datasets/larsen0966/student-performance-data-set

Shetty, A. B. (2020). Exploratory Data Analysis on Student Performance. Diakses dari https://www.kaggle.com/code/adithyabshetty100/exploratory-data-analysis-on-student-performance

Manne, A. (2017). EDA on Students Performance in Exams. Diakses dari https://www.kaggle.com/code/amulyamanne/eda-on-students-performance-in-exams

Naing, T. T. (2024). Exploratory Data Analysis (EDA) — Analyzing Factors Influencing Student Performance. Medium. Diakses dari https://medium.com/@thitnaing86/exploratory-data-analysis-eda-analyzing-factors-influencing-student-performance-7ff7a576b463

Admasu Yekun, E., & Teklay, A. (2019). Student Performance Prediction with Optimum Multilabel Ensemble Model. Diakses dari https://arxiv.org/abs/1909.07444

Injadat, M., Moubayed, A., Bou Nassif, A., & Shami, A. (2020). Multi-split Optimized Bagging Ensemble Model Selection for Multi-class Educational Data Mining. Diakses dari https://arxiv.org/abs/2006.05031