BIG DATA & PREDICTIVE ANALYTICS FINAL PROJECT
PEMODELAN PREDIKTIF BMI BERDASARKAN PARAMETER FISIK MENGGUNAKAN METODE REGRESI

Dosen Pengampu : Mulia Sulistiyono, M.Kom

Disusun Oleh :

1.	Nasiha Assakinah	23.11.5395
2.	Arya Nurhikam	23.11.5420
3.	Wahyu Setyo Dwicahyo	23.11.5434
4.	Y Putra Perdana	23.11.5436

PROGRAM STUDI INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA

2025

BAB I

LATAR BELAKANG

Body Mass Index (BMI) merupakan indikator yang digunakan secara luas untuk menilai status gizi seseorang berdasarkan perbandingan antara berat badan dan tinggi badan. Dalam dunia kesehatan, BMI sering dijadikan acuan untuk mengklasifikasikan kondisi berat badan seseorang ke dalam kategori seperti kurus, normal, gemuk, dan obesitas. Klasifikasi ini sangat penting karena berkaitan erat dengan risiko munculnya berbagai penyakit kronis, antara lain hipertensi, diabetes, serta gangguan jantung.

Seiring dengan perkembangan teknologi, pendekatan berbasis data dan machine learning menjadi alternatif yang semakin relevan untuk memprediksi nilai BMI secara lebih cepat dan akurat. Dengan memanfaatkan data kesehatan yang memuat informasi seperti usia, jenis kelamin, tinggi badan, dan berat badan, pemodelan statistik—khususnya metode regresi linier berganda—dapat digunakan untuk membangun model prediktif yang handal. Tahapan ini diawali dengan analisis deskriptif untuk memahami karakteristik data, dilanjutkan dengan visualisasi hubungan antar variabel, serta pemilihan fitur-fitur yang relevan untuk menghasilkan model prediksi yang optimal.

Penelitian ini bertujuan untuk mengembangkan sebuah model prediktif BMI yang mampu mengidentifikasi faktor-faktor paling berpengaruh terhadap kondisi berat badan seseorang. Hasil dari pemodelan ini diharapkan tidak hanya memberikan informasi numerik mengenai nilai BMI, tetapi juga dapat digunakan sebagai acuan dalam upaya pencegahan berbagai penyakit yang berkaitan dengan ketidakseimbangan berat badan.

BAB II

METODE PENELITIAN

2. 1 Metode

1. **Metode Analisis Deskriptif**

   Metode analisis deskriptif digunakan sebagai tahap awal untuk memahami karakteristik data serta hubungan antar variabel yang memengaruhi nilai Body Mass Index (BMI). Hasil dari analisis ini menjadi dasar dalam menentukan variabel-variabel yang relevan untuk dimasukkan ke dalam model prediksi.

2. **Metode Regresi Linier Berganda**

   Metode regresi linier berganda digunakan sebagai pendekatan utama dalam memodelkan dan memprediksi nilai BMI berdasarkan beberapa variabel input, seperti berat badan, tinggi badan, usia, dan jenis kelamin. Metode ini dipilih karena mampu menjelaskan hubungan linier antara variabel bebas dengan variabel terikat (BMI), serta memberikan hasil prediksi yang cukup akurat.

2. 2 Alur Final Project

1. **Data Colletion**

   Bertujuan untuk mendapatkan data BMI yang relavan dan bisa diteliti lebih lanjut.

   Proses :

   - Mengumpulkan data BMI dari berbagai sumber seperti data berat badan,tinggi badan seseorang.
   - Memastikan bahwa data yang diperoleh relevan dan sesuai dengan kebutuhan analisis.

2. **EDA dan Visualisasi Data**

   Dilakukan untuk memahami struktur data, mendeteksi pola, mengetahui distribusi nilai, serta mengidentifikasi hubungan antar variabel sebelum dilakukan pemodelan.

   Proses :

   - Mengecek jumlah data, tipe variabel, nilai yang hilang (missing values), dan data tidak konsisten.

- Menghitung nilai-nilai seperti rata-rata, median, standar deviasi, minimum, dan maksimum untuk memahami sebaran dan karakteristik setiap variabel dalam dataset.
- Menggunakan grafik seperti histogram, scatter plot, boxplot, dan heatmap untuk melihat distribusi data, hubungan antar variabel.

3. Analisis Korelasi

Analisis ini bertujuan untuk mengetahui variabel mana yang memiliki pengaruh paling signifikan terhadap nilai BMI.

Proses :

- Mengambil variabel numerik dari dataset yang akan dianalisis, seperti berat badan, tinggi badan, usia, dan BMI.
- Menggunakan metode statistik seperti Pearson Correlation untuk menghitung nilai korelasi antar pasangan variabel,
- Menampilkan hasil korelasi dalam bentuk heatmap agar hubungan antar variabel terlihat lebih jelas dan mudah dianalisis secara visual.

4. Regresi Linear Berganda

Bertujuan untuk membangun model prediktif yang mampu mengukur seberapa besar pengaruh masing-masing variabel input terhadap nilai BMI.

Proses :

- Memilih variabel independen (fitur) yang relevan dan memiliki korelasi signifikan terhadap BMI, lalu memisahkan data menjadi data yg sudah dilatih.
- Menguji performa model menggunakan metrik evaluasi seperti R-squared (R^2), Mean Squared Error (MSE), dan Root Mean Squared Error (RMSE) untuk menilai seberapa baik model memprediksi nilai BMI pada data uji.

2. 3 Dataseet

Dataset yang digunakan dapat diakses melalui platform Kaggle https://www.kaggle.com/datasets/prasad22/healthcare-dataset. Dataset ini berisi data kesehatan individu yang digunakan untuk memprediksi nilai BMI berdasarkan beberapa variabel seperti usia, jenis kelamin, tinggi badan, dan berat badan. jumlah Baris dan Kolom Dataset terdiri dari 10004 baris dan 5 kolom :

- Age : Usia individu (tipe data: *int64*).
- Gender : Jenis kelamin individu (Male/Female) (tipe data: *object*).
- Height_cm : Tinggi badan dalam sentimeter (tipe data: *float64*).

- Weight_kg : Berat badan dalam kilogram (tipe data: *float64*).
  BMI : Body Mass Index atau Indeks Massa Tubuh hasil perhitungan berdasarkan berat dan tinggi (tipe data: *float64*).

```
Nama Kolom Final: ['Age', 'Gender', 'Height_cm', 'Weight_kg', 'BMI']

5 Baris Pertama Data Bersih:
   Age  Gender   Height_cm   Weight_kg    BMI
0   32  Female  185.548351  110.646301  32.14
1   21    Male  164.240726   41.684285  15.45
2   53  Female  177.607370   91.046022  28.86
3   49    Male  184.892880   58.869005  17.22
4   46    Male  164.832291   96.739075  35.61
```

2. 4 Proses EDA

1. Pemeriksaan Data Awal

Menampilkan beberapa baris awal dari dataset untuk memastikan format data sudah benar. Pemeriksaan dilakukan terhadap jumlah baris dan kolom serta tipe data dari masing-masing kolom. Pada tahap ini juga dilakukan identifikasi nilai-nilai yang hilang (missing values) dan mengatasinya.

2. Statistik Deskriptif

Menggunakan fungsi statistik seperti .describe() untuk mengetahui ringkasan numerik (min, max, mean, std) dari variabel seperti Age, Height_cm, Weight_kg, dan BMI. Untuk kolom kategorikal seperti Gender, digunakan .value_counts() untuk melihat distribusinya.

3. Visualisasi dan Analisis Korelasi

Visualisasi data dilakukan dengan berbagai jenis grafik untuk memahami distribusi data, mengidentifikasi pola, serta menganalisis hubungan antar variabel. Beberapa visualisasi yang digunakan antara lain :

1. Bar Chart

Digunakan untuk menampilkan distribusi persentase jenis kelamin dari individu dengan BMI di atas 30 (kategori obesitas). Grafik ini membantu memahami proporsi obesitas berdasarkan gender secara visual.

2. Pie Chart

Digunakan untuk memperlihatkan distribusi kategori BMI (Underweight, Normal, Overweight, Obese) dalam bentuk persentase.

Pie chart memberikan gambaran menyeluruh tentang komposisi status gizi dalam dataset.

3. Line Chart

Menampilkan tren nilai rata-rata BMI terhadap usia, yang dibagi dalam kelompok usia remaja, dewasa, dan lansia. Visualisasi ini memberikan informasi mengenai bagaimana nilai BMI cenderung berubah seiring bertambahnya usia.

4. Heatmap

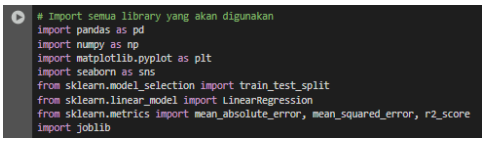
Selain visualisasi kategori, dilakukan juga analisis korelasi numerik antar variabel menggunakan fungsi .corr(), yang kemudian divisualisasikan dalam bentuk heatmap. Heatmap ini berguna untuk melihat kekuatan hubungan antara variabel-variabel seperti berat badan, tinggi badan, usia, dan BMI.

BAB III

EKSPERIMEN

3. 1 Proses Eksperimen

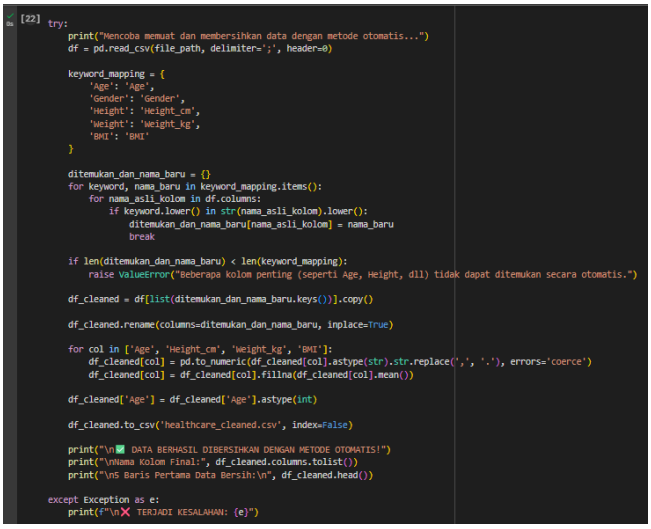
1. Import Library yang Dibutuhkan



```
# Import semua library yang akan digunakan
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import joblib
```

Import library seperti pandas, numpy, matplotlib, seaborn, dan sklearn untuk pengolahan data, visualisasi, serta pembangunan dan evaluasi model.

2. Eksplorasi dan Pembersihan Data (EDA & Cleaning)



```
[22] try:
        print("Mencoba memuat dan membersihkan data dengan metode otomatis...")
        df = pd.read_csv(file_path, delimiter=';', header=0)

        keyword_mapping = {
            'Age': 'Age',
            'Gender': 'Gender',
            'Height': 'Height_cm',
            'Weight': 'Weight_kg',
            'BMI': 'BMI'
        }

        ditemukan_dan_nama_baru = {}
        for keyword, nama_baru in keyword_mapping.items():
            for nama_asli_kolom in df.columns:
                if keyword.lower() in str(nama_asli_kolom).lower():
                    ditemukan_dan_nama_baru[nama_asli_kolom] = nama_baru
                    break

        if len(ditemukan_dan_nama_baru) < len(keyword_mapping):
            raise ValueError("Beberapa kolom penting (seperti Age, Height, dll) tidak dapat ditemukan secara otomatis.")

        df_cleaned = df[list(ditemukan_dan_nama_baru.keys())].copy()

        df_cleaned.rename(columns=ditemukan_dan_nama_baru, inplace=True)

        for col in ['Age', 'Height_cm', 'Weight_kg', 'BMI']:
            df_cleaned[col] = pd.to_numeric(df_cleaned[col].astype(str).str.replace(',', '.'), errors='coerce')
            df_cleaned[col] = df_cleaned[col].fillna(df_cleaned[col].mean())

        df_cleaned['Age'] = df_cleaned['Age'].astype(int)

        df_cleaned.to_csv('healthcare_cleaned.csv', index=False)

        print("\n✅ DATA BERHASIL DIBERSIHKAN DENGAN METODE OTOMATIS!")
        print("\nNama Kolom Final:", df_cleaned.columns.tolist())
        print("\n5 Baris Pertama Data Bersih:\n", df_cleaned.head())

    except Exception as e:
        print(f"\n❌ TERJADI KESALAHAN: {e}")
```

Data dianalisis secara deskriptif untuk memahami distribusi, pola, dan potensi anomali. Selanjutnya dilakukan pembersihan data dengan mengatasi nilai kosong (missing value), data duplikat, atau kesalahan input.

3. Visualisasi Data dan Analisis Awal

```python
# 1. Filter dataset untuk mengambil data dengan BMI > 30
df_obese = df_cleaned[df_cleaned['BMI'] > 30].copy()

print(f"Ditemukan {len(df_obese)} pasien dengan BMI di atas 30.")

# 2. Hitung jumlah berdasarkan gender, lalu ubah menjadi persentase
gender_counts = df_obese['Gender'].value_counts()
gender_percentages = (gender_counts / len(df_obese)) * 100

print("\nPersentase berdasarkan gender untuk kelompok ini:")
print(gender_percentages)

# 3. Buat bar chart dari data persentase
plt.figure(figsize=(8, 6))
ax = sns.barplot(x=gender_percentages.index, y=gender_percentages.values, palette='viridis')
plt.title('Persentase Gender untuk BMI di Atas 30 (Kategori Obesitas)', fontsize=16)
plt.xlabel('Gender', fontsize=12)
plt.ylabel('Persentase (%)', fontsize=12)
plt.ylim(0, 100)

# 4. Tambahkan label persentase di atas setiap bar
for p in ax.patches:
    height = p.get_height()
    ax.text(p.get_x() + p.get_width() / 2., height + 1, f'{height:.1f}%', ha="center")

# Simpan dan tampilkan grafik
plt.savefig('bmi_obese_percentage_barchart.png')
print("\n✅ Bar chart persentase berhasil dibuat.")
plt.show()
```

```python
bins = [0, 18.5, 25, 30, np.inf]
labels = ['Underweight', 'Normal', 'Overweight', 'Obese']
df_cleaned['BMI_Category'] = pd.cut(df_cleaned['BMI'], bins=bins, labels=labels, right=False)

bmi_counts = df_cleaned['BMI_Category'].value_counts()

plt.figure(figsize=(10, 8))
plt.pie(bmi_counts, labels=bmi_counts.index, autopct='%1.1f%%', startangle=140, colors=sns.color_palette('pastel'))
plt.title('Distribusi Kategori BMI', fontsize=16)
plt.savefig('bmi_category_piechart.png')
print("✅ Pie chart dibuat.")
plt.show()
```

```python
try:
    # 1. Buat kolom baru 'Age_Group' untuk mengkategorikan usia
    # Bins (rentang): 17-25, 26-59, 60 ke atas
    bins = [16, 25, 59, float('inf')]
    labels = ['Remaja (17-25)', 'Dewasa (26-59)', 'Lansia (60+)']

    # Menggunakan pd.cut untuk membuat kategori secara efisien
    df_cleaned['Age_Group'] = pd.cut(df_cleaned['Age'], bins=bins, labels=labels, right=True)

    # 2. Buat plot
    plt.figure(figsize=(14, 8))
    sns.set_style("whitegrid")

    # 3. Gunakan 'hue' untuk membuat garis terpisah berdasarkan 'Age_Group'
    sns.lineplot(
        data=df_cleaned,
        x='Age',
        y='BMI',
        hue='Age_Group',
        marker='o',
        palette='viridis'
    )

    # 4. Atur judul dan label
    plt.title('Tren Rata-rata BMI berdasarkan Kelompok Usia', fontsize=18)
    plt.xlabel('Usia', fontsize=12)
    plt.ylabel('Rata-rata BMI', fontsize=12)
    plt.legend(title='Kelompok Usia')
    plt.grid(True)

    # Simpan dan tampilkan grafik
    plt.savefig('bmi_age_groups_linechart.png')
    print("✅ Line chart berdasarkan kelompok usia berhasil dibuat.")
    plt.show()

except NameError:
    print("❌ KESALAHAN: DataFrame 'df_cleaned' tidak ditemukan. Jalankan sel pembersihan data terlebih dahulu.")
except Exception as e:
    print(f"❌ Terjadi kesalahan: {e}")
```

Untuk memahami pola dalam data, digunakan beberapa visualisasi, antara lain pie chart untuk menunjukkan distribusi kategori BMI (Underweight hingga Obese), bar chart untuk membandingkan proporsi obesitas berdasarkan jenis kelamin, dan line chart untuk melihat tren rata-rata BMI berdasarkan kelompok usia (Remaja, Dewasa, Lansia).

4. Analisis Korelasi dan Pemilihan Fitur

```
try:
    numerical_cols = ['Age', 'Height_cm', 'Weight_kg', 'BMI']
    correlation_matrix = df_cleaned[numerical_cols].corr()

    plt.figure(figsize=(10, 8))
    sns.heatmap(
        correlation_matrix,
        annot=True,          # Menampilkan angka korelasi
        cmap='coolwarm',     # Palet warna: biru (negatif), merah (positif)
        fmt=".2f",           # Format angka menjadi 2 desimal
        linewidths=.5,
        square=True,
        vmin=-1, vmax=1
    )

    plt.title('Heatmap Korelasi Antar Variabel Numerik', fontsize=16)

    # Simpan dan tampilkan grafik
    plt.savefig('correlation_heatmap_improved.png')
    print("Heatmap berhasil dibuat.")
    plt.show()

except NameError:
    print("❌ KESALAHAN: DataFrame 'df_cleaned' tidak ditemukan. Jalankan sel pembersihan data terlebih dahulu.")
except Exception as e:
    print(f"❌ Terjadi kesalahan: {e}")
```

Mengidentifikasi hubungan antar variabel dan memilih fitur-fitur yang memiliki pengaruh signifikan terhadap variabel target, guna meningkatkan performa model.

5. Pembangunan Model Regresi Linier Berganda & Evaluasi Model

```
features = ['Height_cm', 'Weight_kg', 'Age']
target = 'BMI'
X = df_cleaned[features]
y = df_cleaned[target]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)

joblib.dump(model, 'bmi_prediction_model.pkl')
y_pred = model.predict(X_test)

mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("--- Hasil Evaluasi Model ---")
print(f"Mean Absolute Error (MAE): {mae:.2f}")
print(f"Mean Squared Error (MSE): {mse:.2f}")
print(f"R-squared (R2 Score): {r2:.4f}")

plt.figure(figsize=(10, 8))
sns.scatterplot(x=y_test, y=y_pred, alpha=0.6)
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], color='red', linestyle='--', lw=2, label='Prediksi Sempurna')
plt.title('Nilai BMI Asli vs. Hasil Prediksi', fontsize=16)
plt.legend()
plt.grid(True)
plt.savefig('prediction_vs_actual.png')
print("\n Plot evaluasi")
plt.show()
```

Model dibangun menggunakan pendekatan regresi linier berganda untuk memprediksi variabel target berdasarkan beberapa fitur input. Model kemudian dievaluasi menggunakan metrik statistik seperti R^2, MSE, atau RMSE untuk mengukur tingkat akurasi dan efektivitasnya.

3. 2 Library dan Tools

1. **Tools :**
    - **Python**
      Sebagai bahasa pemrograman utama dalam seluruh proses analisis dan pemodelan.
    - **Google Colab**
      Platform cloud-based yang digunakan untuk menjalankan dan menyimpan notebook Python secara interaktif tanpa perlu instalasi lokal.
2. **Library :**
    - **Pandas**
      Untuk mengolah dan menganalisis data dalam bentuk tabel (DataFrame).
    - **Numpy**
      Untuk komputasi numerik, terutama manipulasi array dan operasi matematika.
    - **Matplotlib**
      Untuk membuat grafik seperti line chart, bar chart, dan scatter plot.
    - **Seaborn**
      Untuk visualisasi data statistik dengan tampilan grafik yang lebih menarik.
    - **Sklearn (scikit-learn)**
      Untuk membangun dan mengevaluasi model machine learning.
    - **Joblib**
      Untuk menyimpan dan memuat model machine learning secara efisien.

BAB IV

HASIL DAN EVALUASI

Hasil

Tahapan awal dilakukan dengan mengimpor dan membersihkan data dari file healthcare_dataset.csv. Proses pembersihan mencakup pemilihan kolom-kolom penting seperti usia (Age), jenis kelamin (Gender), tinggi badan (Height_cm), berat badan (Weight_kg), dan indeks massa tubuh (BMI), serta penanganan terhadap nilai kosong (missing values) dan format data yang tidak konsisten.

Kemudian eksplorasi awal untuk memahami karakteristik data. Visualisasi distribusi BMI menunjukkan bahwa sebagian besar individu berada dalam kategori normal dan overweight, dengan proporsi tertentu yang termasuk dalam kategori obesitas. Selain itu, analisis juga memperlihatkan bahwa kategori obesitas lebih banyak ditemukan pada kelompok laki-laki dibandingkan perempuan.

Selanjutnya, dilakukan analisis korelasi antar variabel numerik untuk mengidentifikasi hubungan yang signifikan terhadap BMI. Hasil visualisasi korelasi dalam bentuk heatmap menunjukkan bahwa berat badan memiliki korelasi paling kuat terhadap nilai BMI, diikuti oleh tinggi badan dan usia.

Model regresi linier berganda kemudian dibangun dengan menggunakan tiga variabel input: Age, Height_cm, dan Weight_kg. Model ini dilatih menggunakan data yang telah dibagi secara proporsional (train-test split), dan kemudian dievaluasi untuk mengukur akurasi prediksi terhadap nilai BMI.

Evaluasi

- MAE menunjukkan rata-rata selisih absolut antara nilai prediksi dan nilai aktual. Nilai 1.93 berarti bahwa, secara rata-rata, prediksi BMI oleh model berbeda sekitar 1.93 satuan dari nilai sebenarnya.
- MSE mengukur rata-rata dari kuadrat selisih antara nilai prediksi dan nilai aktual. Karena selisih dikuadratkan, metrik ini memberi penalti lebih besar terhadap kesalahan yang lebih besar. Nilai 6.21 menunjukkan bahwa rata-rata kuadrat kesalahan prediksi model adalah 6.21.
- R^2 atau koefisien determinasi mengukur proporsi variasi dalam data target yang dapat dijelaskan oleh model. Nilai 0.9655 atau 96.55% menunjukkan bahwa model mampu menjelaskan sebagian besar variasi dalam data BMI berdasarkan variabel prediktor (Age, Height_cm, Weight_kg).

BAB V

KESIMPULAN

5. 1 Kesimpulan

Berdasarkan eksperimen yang telah dilakukan, dapat disimpulkan bahwa model Regresi Linier Berganda efektif untuk memprediksi nilai BMI berdasarkan variabel tinggi badan (Height_cm), berat badan (Weight_kg), dan usia (Age). Hasil evaluasi menunjukkan bahwa model memiliki akurasi yang sangat tinggi, dengan nilai R^2 sebesar 0.9655, yang berarti model mampu menjelaskan lebih dari 96% variasi dalam data. Dengan demikian, model ini dapat digunakan sebagai alat prediktif yang akurat dalam konteks analisis kesehatan, khususnya untuk memantau dan mengklasifikasikan risiko obesitas berdasarkan data fisik dasar seseorang.

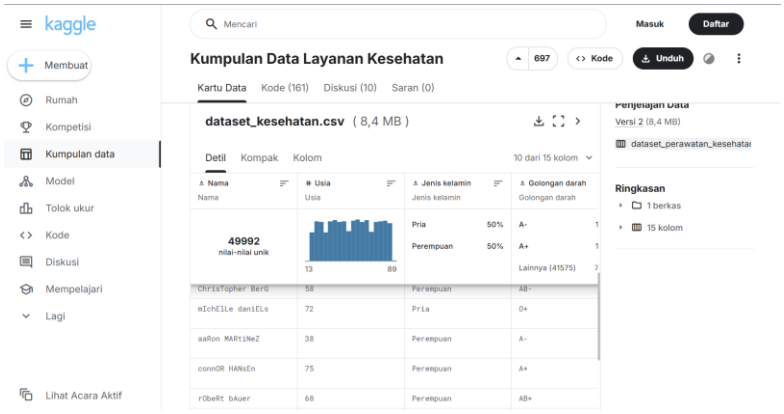
.

5. 2 Kontribusi

NIM	NAMA	KONTRIBUSI
23.11.5395	Nasiha Assakinah	Mencari dataseet, coding eksperimen dataset, membuat laporan
23.11.5420	Arya Nurhikam	Membuat poster, coding eksperimen dataseet
23.11.5434	Wahyu Setyo Dwicahyo	Membuat poster, coding eksperimen dataseet
23.11.5436	Y Putra Perdana	Coding eksperimen dataseet, coding dashboard streamlit

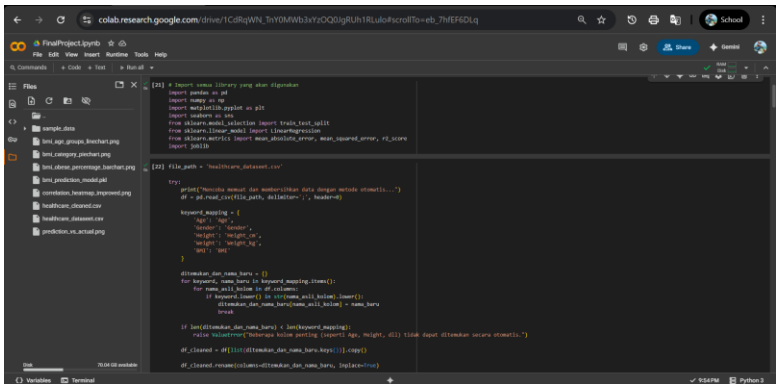
LAMPIRAN

Dataseet



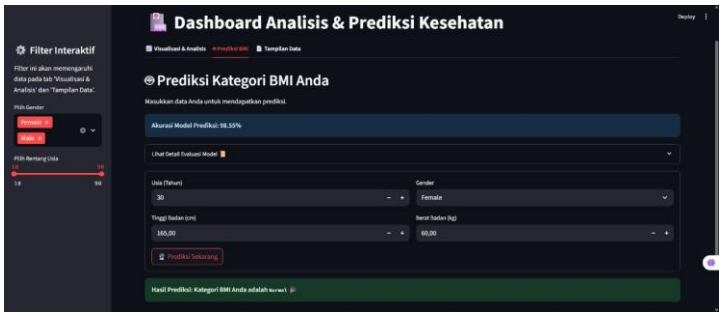
https://www.kaggle.com/datasets/prasad22/healthcare-dataset/data

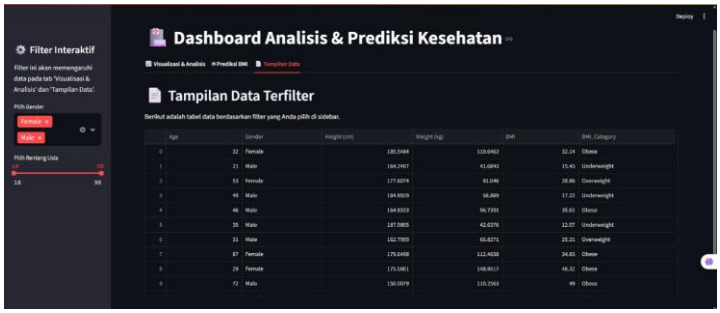
Google Colab

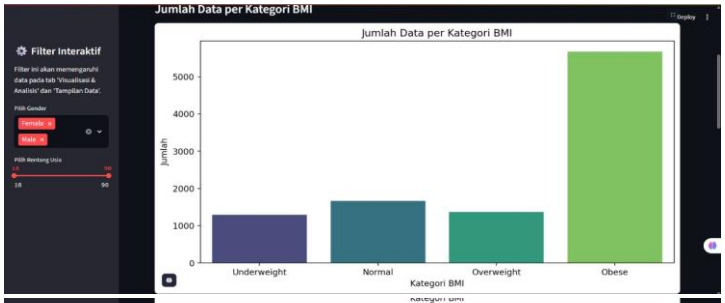


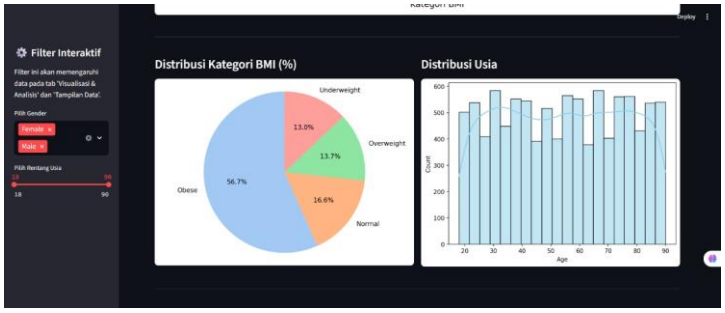
https://drive.google.com/file/d/1CdRqWN_TnY0MWb3xYzOQ0JgRUh1RLulo/view?usp=sharing

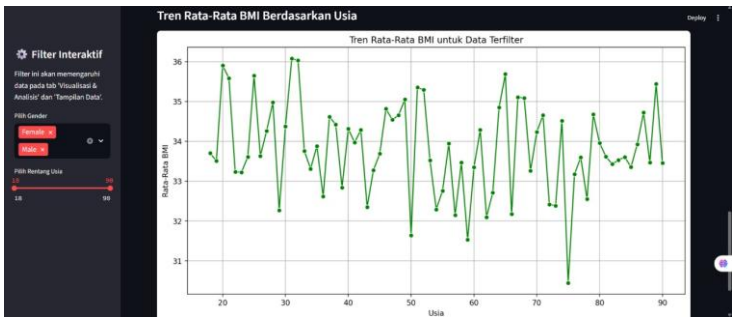
Dashboard





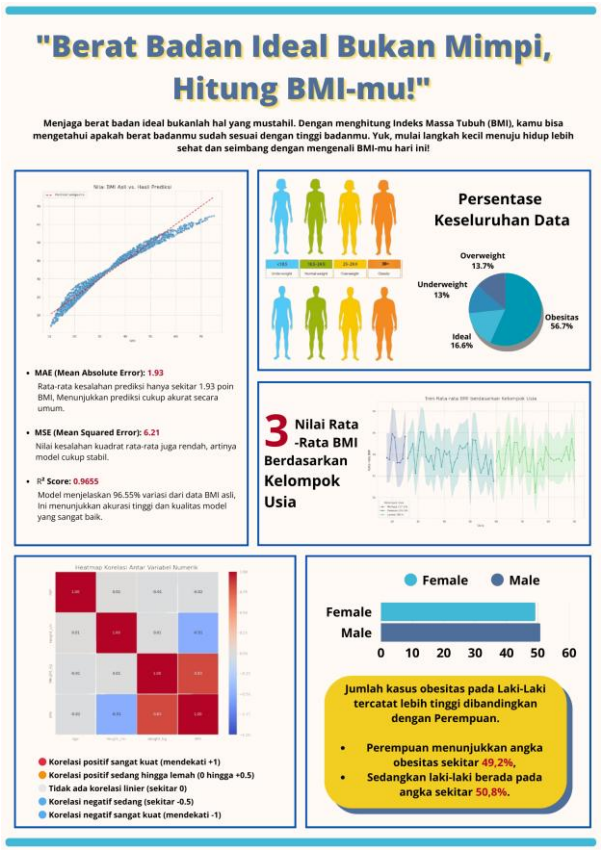






https://github.com/setyok/Final-Projek-Big-Data

Poster



https://www.canva.com/design/DAGtHLebqb0/8BexVqQIlIGz9IT-RcUDBw/edit?utm_content=DAGtHLebqb0&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton