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

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 Alur Final Project

1. **Data Colletion**

   Mendapatkan data BMI yang relavan dan bisa diteliti lebih lanjut.

2. **EDA dan Visualisasi Data**

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

3. **Analisis Korelasi**

   Mengidentifikasi variabel mana yang memiliki pengaruh paling signifikan terhadap nilai BMI.

4. **Regresi Linear Berganda**

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

2. 2 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 (tipe data: *float64*).

2. 3 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.

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**

      Menampilkan 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**

      Menampilkan 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 mengenai bagaimana nilai BMI cenderung berubah seiring bertambahnya usia.

   4. **Heatmap**

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

2. **Eksplorasi dan Pembersihan Data (EDA & Cleaning)**

   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**

   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**

   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**

   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

Tahapan awal dilakukan dengan mengimpor dan membersihkan data dari file *healthcare_dataset.csv*. Proses pembersihan mencakup pemilihan kolom-kolom penting seperti Age, Gender, Height_cm, Weight_kg, dan BMI, serta mengisi nilai kosong (missing values) dan format data yang tidak konsisten sehingga tercipta dataseet bersih *healthcare_cleaned.csv* 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.

- MAE (1.93) menunjukkan rata-rata selisih absolut antara nilai prediksi dan nilai aktual.
- MSE (6.21) menunjukkan 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.
- R^2 (0.9655 atau 96.55%) 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

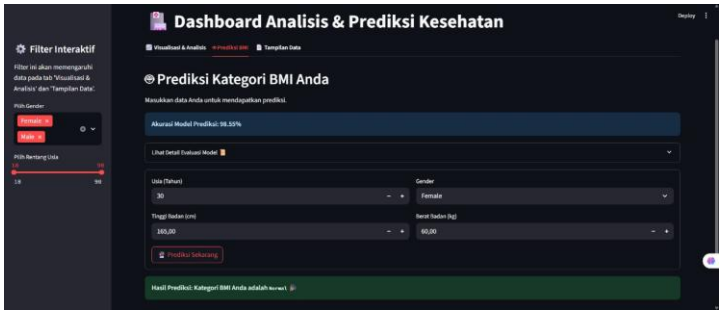
Dataseet

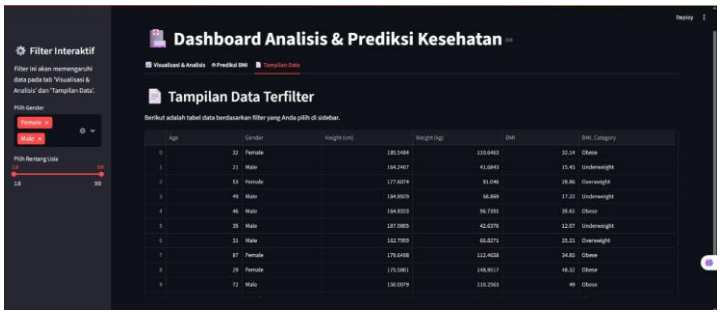
https://www.kaggle.com/datasets/prasad22/healthcare-dataset/data

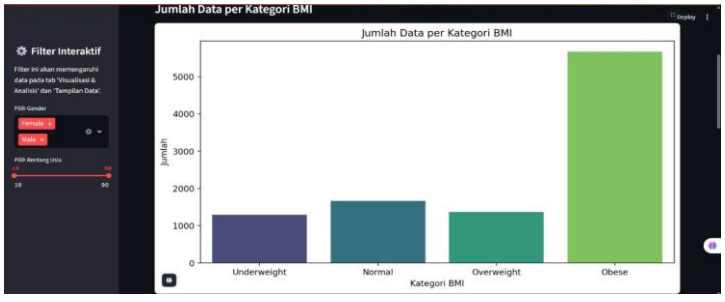
Google Colab

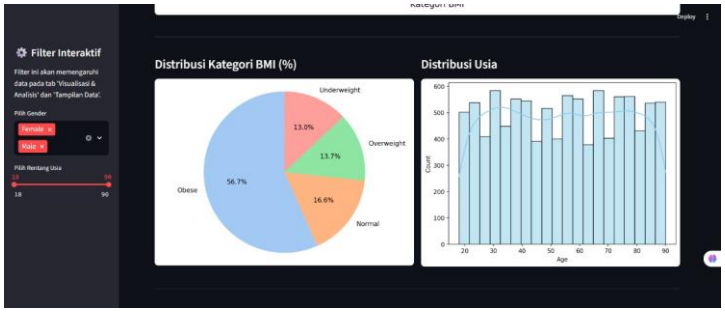
https://drive.google.com/file/d/1CdRqWN_TnY0MWb3xYzOQ0JgRUh1RLulo/view?usp=sharing

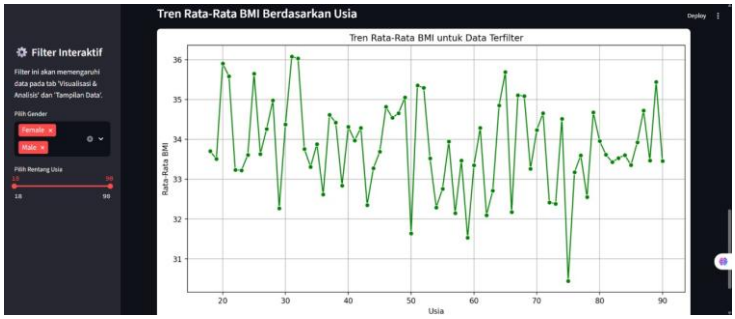
Dashboard





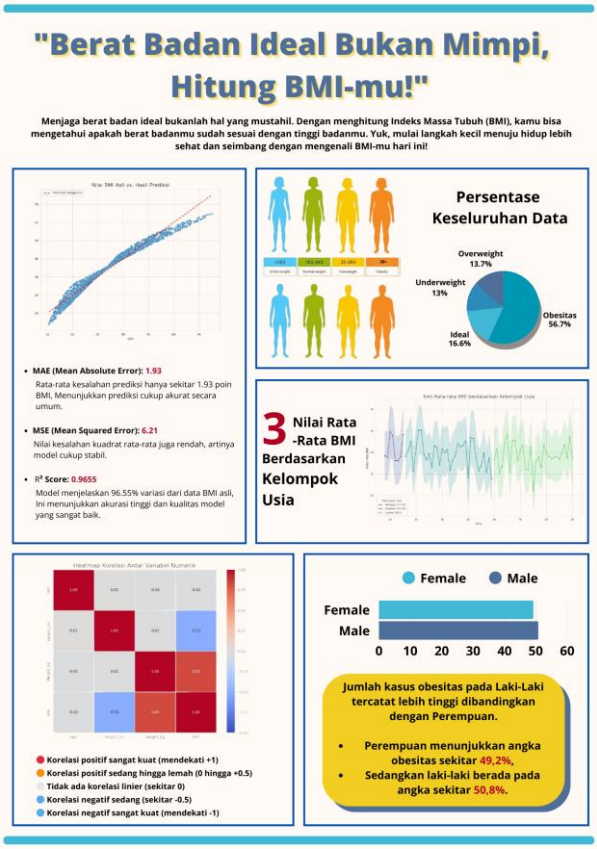






https://github.com/setyok/Final-Projek-Big-Data

Poster



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