# Calculating Estimated Obesity Levels Based on Eating Habits and Physical Condition Using Classification Techniques

# 

**Mata Kuliah Big Data Processing - COMP6579001**

2602128283 - Yervant Denzel Budyawan

2602127721 - Bryan Abraham Sunaryo

2602128863 - Dimas Raihan Al Ghifari

2602127646 - Ryan Tristan Wijaya

# Bab 1: Latar belakang

Obesitas merupakan salah satu masalah kesehatan yang paling marak di abad ke-21. Frekuensi terjadinya obesitas terus meningkat secara global, hal ini memicu peningkatan risiko penyakit kronis seperti diabetes tipe 2, penyakit jantung, dan berbagai bentuk kanker. Obesitas tidak hanya membebani sistem kesehatan tetapi juga mempengaruhi kualitas hidup individu yang mengalaminya.

Dalam beberapa dekade terakhir, kemajuan teknologi dalam analisis data telah membuka peluang baru untuk memahami dan mengatasi masalah obesitas. Analisis data yang menyeluruh dapat membantu mengidentifikasi pola dan faktor risiko yang menyebabkan terjadinya obesitas. Salah satu metode yang efektif dalam analisis prediktif adalah logistic regression / regresi logistik. Logistic regression dapat digunakan untuk memprediksi tingkat obesitas berdasarkan kebiasaan makan dan kondisi fisik seseorang..

Eksperimen ini bertujuan untuk mengembangkan model prediksi tingkat obesitas menggunakan algoritma machine learning, khususnya logistic regression. Model ini diharapkan dapat memberikan prediksi yang akurat mengenai tingkat obesitas seseorang berdasarkan kebiasaan makan dan kondisi fisik mereka. Dengan menggunakan pendekatan ini, diharapkan dapat memberikan wawasan yang lebih mendalam tentang faktor-faktor yang mempengaruhi obesitas.

Melalui Eksperimen ini, kami berharap dapat berkontribusi dalam upaya pencegahan dan pengelolaan obesitas secara lebih efektif dengan memanfaatkan analisis data dan machine learning. Model prediksi yang dihasilkan dapat digunakan oleh profesional kesehatan untuk melakukan deteksi dini dan intervensi yang tepat, sehingga dapat mengurangi frekuensi terjadi obesitas dan meningkatkan kesehatan masyarakat secara keseluruhan.

# 

# 

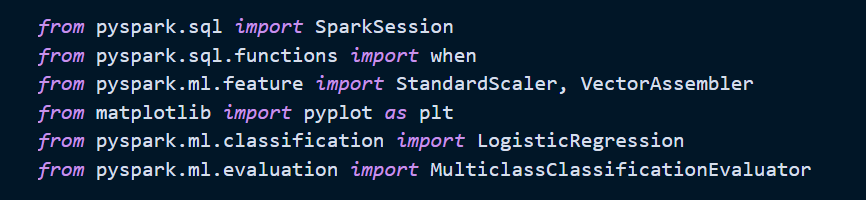
# Bab 2: Metodologi

Pada bab ini, kami akan menjelaskan mengenai langkah-langkah/alur kerja yang dilakukan untuk mengolah dataset mengenai estimasi tingkat obesitas berdasarkan kebiasaan makan dan kondisi fisik.

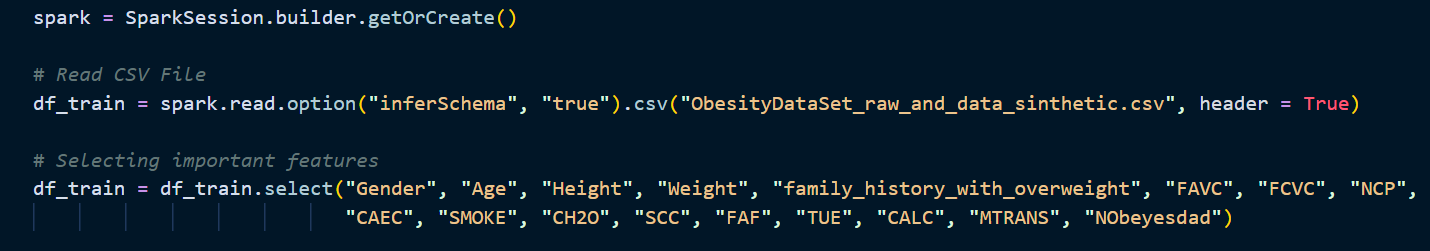
1. Data Preparation

Melakukan pemanggilan modul yang akan digunakan untuk mengolah dataset([Estimation of Obesity Levels Based On Eating Habits and Physical Condition - UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition)).

* SparkSession berfungsi untuk membuat dataframe
* When berfungsi untuk melakukan transformasi data dari tipe kategorikal ke dalam bentuk numerikal
* StandardScaler berfungsi untuk melakukan normalisasi agar skala perhitungan sebuah feature tidak memiliki rentang yang terlalu jauh
* VectorAssembler berfungsi untuk menggabungkan beberapa kolom feature menjadi satu vector
* Pyplot berfungsi untuk melakukan visualisasi data
* Logistic Regression merupakan sebuah algoritma untuk melakukan klasifikasi
* MultilayerPerceptronClassifier berfungsi untuk mengevaluasi hasil dan akurasi dari model yang telah dibuat

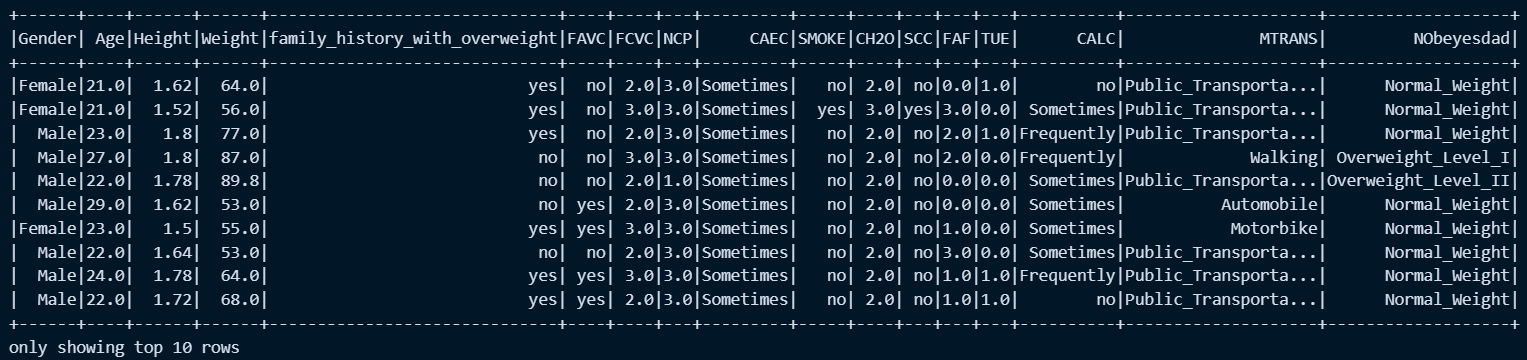


* Memanggil SparkSession untuk membuat dataframe
* Melakukan pembacaan .csv file
* Memilih feature-feature yang akan digunakan dan diproses



Penjelasan feature-feature yang digunakan:

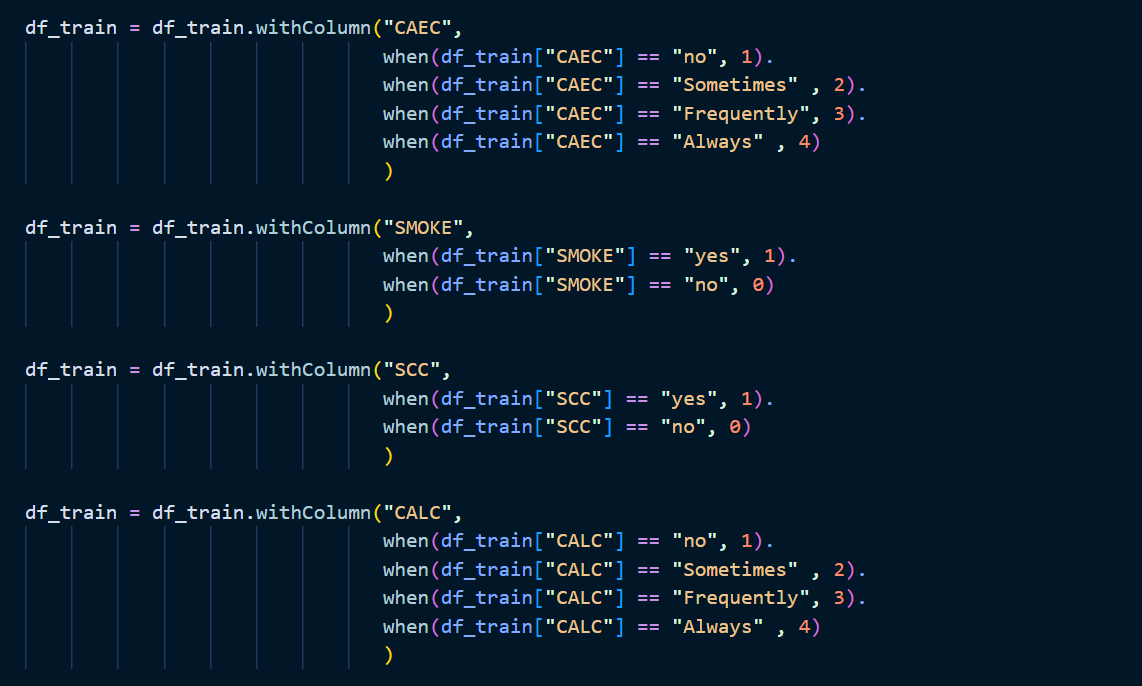
* Family\_history\_with\_overweight, apakah memiliki keluarga yang menderita obesitas
* FAVC, seberapa sering mengonsumsi kalori
* FCVC, seberapa sering kita mengkonsumsi sayuran
* NCP, seberapa banyak makan-makanan berat dalam sehari
* CAEC, seberapa sering kamu ngemil sebelum makan-makanan berat
* SMOKE, merokok atau tidak
* CH2O, seberapa banyak kamu minum air dalam sehari
* SCC, seberapa sering kamu mengawasi kalori yang masuk ke dalam tubuh
* FAF, seberapa sering melakukan aktivitas fisik
* TUE, seberapa sering kamu menggunakan teknologi digital seperti smartphone, komputer, televisi, dll
* CALC, seberapa sering minum alkohol
* MTRANS, jenis transportasi yang biasa digunakan
* NObeyesdad, tingkat obesitas



1. Pre-processing

* Menghilangkan NULL value dalam data
* Melakukan data transformasi untuk mengubah feature dengan tipe kategorik menjadi numerik







1. Generate Model and Evaluation

* Memilih features yang akan dimasukan ke dalam machine learning
* remove(“NObeyesdad”) karena tidak digunakan untuk melatih model dan feature tersebut merupakan hasil klasifikasi
* Menggunakan VectorAssembler untuk menggabungkan semua feature yang digunakan ke dalam satu nilai vector
* Hasil dari VectorAssembler kemudian dinormalisasi dengan StandardScaler untuk memperkecil rentang nilai antara para feature
* Melakukan data splitting untuk mendapatkan data yang akan digunakan untuk melatih dan mengetes model
* Menggunakan membuat model machine learning dengan LogisticRegression
* Membuat prediksi dengan menggunakan data testing
* Setelah model dilatih, hasil dievaluasi menggunakan MultiClassClassificationEvaluator untuk mendapatkan akurasi

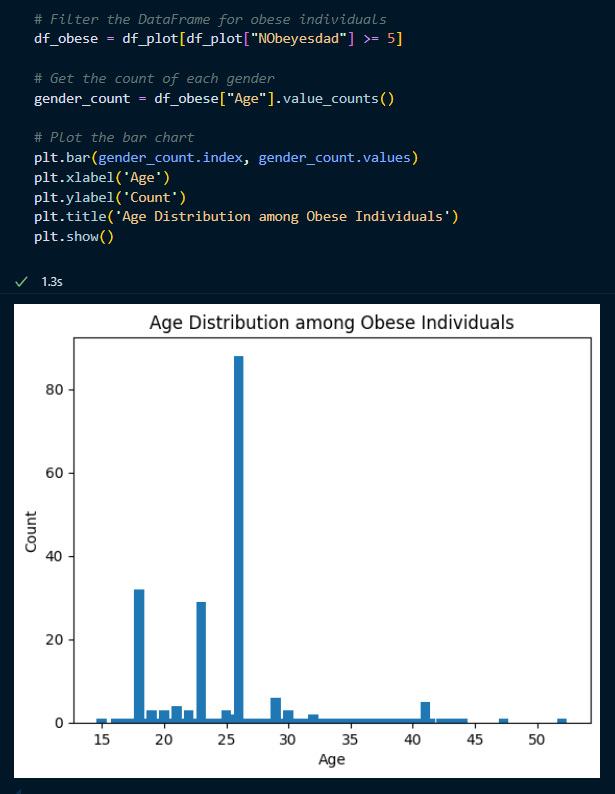


1. Visualization

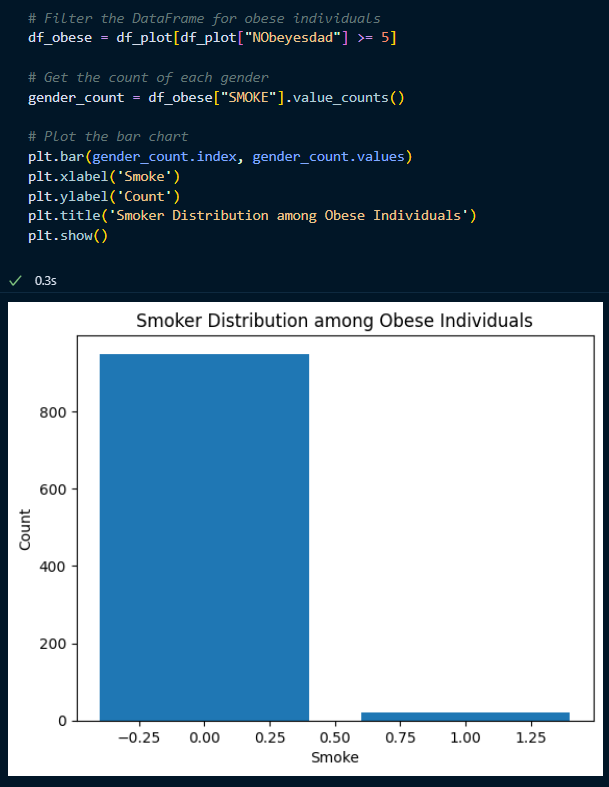
* Gender distribution



* Age Distribution



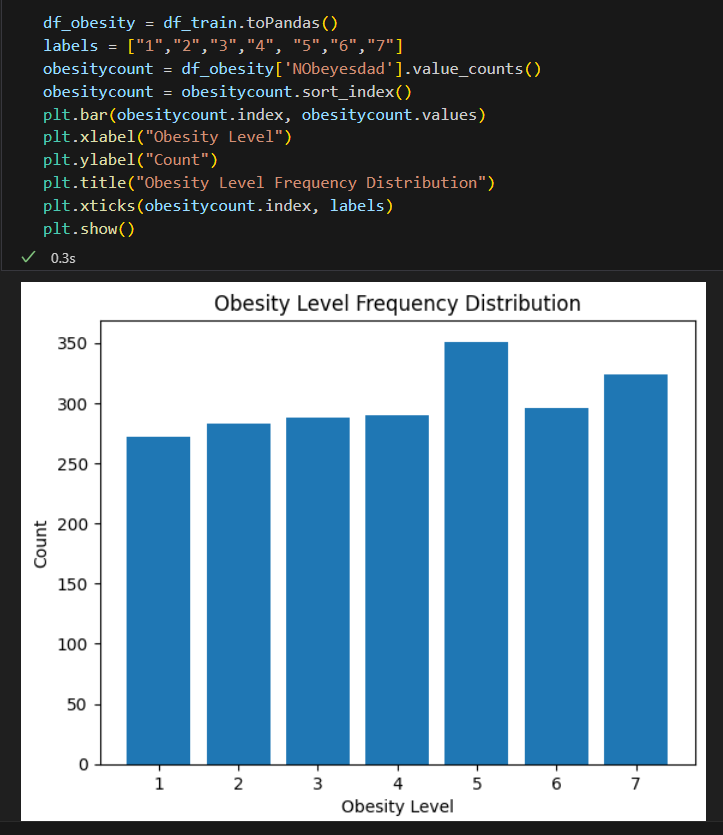
* Smoker distribution



* Exercise distribution



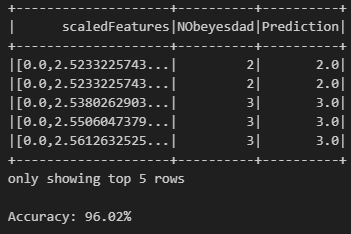
* Obesity Level Frequency with Levels



# Bab 3: Evaluasi dan Detail

Setelah kita melakukan Logistics Regression dengan 100 iterasi menggunakan splitting 7:3, dimana 70% dataset digunakan untuk training model dan 30% digunakan untuk membandingkan dengan modelnya. Hasil akurasi yang kita dapatkan dalam model tersebut adalah 96.02%.

Berikut adalah hasil prediksi:



“NObeyesdad” merupakan data yang digunakan sebagai testing dan “Prediction” dapat menunjukkan hasil yang kita dapat dari training model Logistics Regression.

# 

# 

# 

# Bab 4: Kesimpulan

Berdasarkan hasil kegiatan yang telah dilakukan, penelitian ini berhasil menghasilkan model mengenai prediksi tingkat obesitas dengan menggunakan teknik machine learning dengan algoritma logistic regression. Proses pengolahan dataset terbagi menjadi beberapa tahapan penting mulai dari persiapan data, pre-processing, pembangunan model, hingga evaluasi hasil.

Dalam penelitian ini, data mengenai kebiasaan makan dan kondisi fisik seseorang diolah sedemikian rupa hingga mampu memprediksi tingkat obesitas. Model logistic regression yang digunakan untuk mengembangkan model machine learning menunjukkan akurasi prediksi sebesar 96.02%. Hal ini menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam mengklasifikasikan tingkat obesitas berdasarkan variabel-variabel yang dipilih.

Dengan demikian, penelitian ini mampu memberikan wawasan mendalam mengenai faktor-faktor yang dapat mempengaruhi obesita. Penelitian ini juga membuktikan bahwa teknologi mengenai analisis data dan implementasi machine learning dapat menjadi alat yang efektif dalam mengatasi permasalahan kesehatan global seperti obesitas.