**LAPORAN PEMBELAJARAN MESIN**

**Pengembangan model regresi linear yang lebih akurat untuk memprediksi nilai kalori**

**A blue logo with black background

Description automatically generated**

**Dibuat Oleh : Fitrohananda Ikrom (202131137)**

**TEKNIK INFORMATIKA**

**INSTITUT TEKNOLOGI PLN**

**JAKARTA**

**2023/2024**

# DAFTAR ISI

[DAFTAR ISI i](#_Toc157800043)

[BAB 1 PENDAHLUAN 1](#_Toc157800044)

[1.1 LATAR BELAKANG 1](#_Toc157800045)

[BAB 2 TEORI 2](#_Toc157800046)

[2.1 TINJAUAN PUSTAKA 2](#_Toc157800047)

[2.1.1 Theory and Implementation of linear regression 2](#_Toc157800048)

[2.1.2 Real Estate Value Prediction Using Linear Regression 2](#_Toc157800049)

[2.1.3 A Linear Regression Approach to Modeling Software Characteristics for Classifying Similar Software 2](#_Toc157800050)

[2.1.4 An improvement to linear regression classification for face recognition 3](#_Toc157800051)

[2.1.5 A Review on Linear Regression Comprehensive in Machine Learning 3](#_Toc157800052)

[BAB 3 HASIL 3](#_Toc157800053)

[3.1 PROSES PROGRAM REGRESI LINEAR 4](#_Toc157800054)

[3.1.1 Sumber Dataset 4](#_Toc157800055)

[3.1.2 Analisis Univariat 4](#_Toc157800056)

[3.1.3 Analisis Bivariate 4](#_Toc157800057)

[3.1.4 Mengidentifikasi Outliers 5](#_Toc157800058)

[3.1.5 Modeling dan Evaluasi Model 6](#_Toc157800059)

[3.1.6 Hasil Model 6](#_Toc157800060)

[BAB 4 PENUTUP 7](#_Toc157800061)

[4.1 KESIMPULAN 7](#_Toc157800062)

[4.2 DAFTAR PUSTAKA 8](#_Toc157800063)

[4.3 LAMPIRAN 8](#_Toc157800064)

# BAB 1 PENDAHULUAN

## 1.1 LATAR BELAKANG

Di zaman dimana kepedulian terhadap kesehatan dan gizi semakin meningkat, pemahaman yang akurat tentang nilai kalori makanan sangat penting untuk mendukung gaya hidup sehat dan pola makan yang efektif. Meskipun terdapat berbagai model regresi linier yang digunakan untuk memprediksi nilai kalori, masih terdapat tantangan untuk meningkatkan akurasi prediksi, terutama dalam konteks variabel institusional yang kompleks.

Model – model regresi linier saat ini sering kali mengalami keterbatasan dalam menangani variasi nutrisi yang kompleks dan perbedaan metabolisme individu. Oleh karena itu, diperlukan upaya untuk mengembangkan model yang lebih kompleks dan adaptif untuk memberikan prediksi nilai kalor yang lebih akurat.

Dengan kemajuan teknologi dan analisis data, terdapat peluang untuk mengembangkan model regresi linier yang lebih akurat. Menerapkan algoritme pembelajaran mesin tingkat lanjut dan mengintegrasikan fitur-fitur baru dapat membuka jalan bagi model yang lebih mudah beradaptasi dan andal.

Penelitian ini bertujuan untuk mengatasi keterbatasan model regresi linier yang ada dengan mengembangkan model yang lebih akurat dalam memprediksi nilai kalori. Dengan demikian, diharapkan dapat memberikan kontribusi positif terhadap pemahaman gizi, mendukung pola makan yang lebih efektif, dan meningkatkan kualitas informasi gizi yang diberikan kepada masyarakat.

Ada banyak potensi dampak positif dari pengembangan model ini, termasuk memberikan informasi yang lebih baik kepada konsumen, mendukung penelitian di bidang nutrisi secara lebih efektif, dan berkontribusi dalam upaya global untuk meningkatkan kesadaran tentang nutrisi dan pentingnya makan sehat. Dengan meningkatkan akurasi prediksi nilai kalori, diharapkan dapat membuka jalan menuju masyarakat yang lebih sehat dan terinformasi nutrisinya.

# BAB 2 TEORI

## 2.1 TINJAUAN PUSTAKA

### 2.1.1 Theory and Implementation of linear regression

Pada paper ini, mereka menjelaskan teori dibalik regresi linear dan mengilustrasikan teknik ini dengan dataset dunia nyata. Data ini menghubungkan pendapatan sebuah truk makanan dan ukuran populasi kota tempat truk makanan menjual makanannya. Keimpulannya, Regresi linier adalah teknik yang banyak digunakan di berbagai bidang ilmu pengetahuan dan teknologi. Topik ini merupakan topik utama dalam Pembelajaran Mesin dan Ilmu Data, dua bidang yang sangat populer dan memiliki beragam penerapan. Artikel ini adalah deskripsi mandiri tentang regresi linier, aljabar linier yang diperlukan, dan Python yang diperlukan untuk implementasi. [1]

### 2.1.2 Real Estate Value Prediction Using Linear Regression

Pasar real estate berfokus pada hal harga yang terus berfluktuasi, yang merupakan salah satu bidang utama untuk menerapkan ide pembelajaran mesin tentang cara meningkatkan dan memprediksi biaya dengan akurasi tinggi. Tujuan paper ini adalah untuk memprediksi harga rumah bagi konsumen di kota Mumbai menggunakan algoritma regresi linier dengan mempertimbangkan rencana dan kebutuhan keuangan mereka. Dengan menggambarkan tren pasar di masa lalu dan rentang nilai serta kemajuan di masa depan, sehingga biaya di masa depan dapat diprediksi. Paper ini akan membantu pelanggan menginvestasikan sumber daya dalam surat wasiat tanpa melalui perantara. Hasil penelitian ini menunjukkan bahwa regresi linier memberikan kesalahan prediksi minimal sebesar 0,3713. Kesimpulannya, Di pasar real estate, kesulitan dalam menyimpan dan mengelola informasi dalam jumlah besar dapat diatasi dengan kerangka kerja yang menggunakan algoritma regresi linier. Pendekatan ini meningkatkan keakuratan keputusan pelanggan dan mengurangi risiko alokasi sumber daya melalui penghitungan langsung dan berulang. Perluasan di masa depan melibatkan penambahan database rumah di lebih banyak wilayah perkotaan, memungkinkan pelanggan menjelajahi lebih banyak wilayah dan membuat pilihan yang lebih tepat. Faktor-faktor tambahan seperti penurunan harga tanah mempengaruhi biaya rumah dan elemen-elemen halus dari setiap properti perlu dimasukkan untuk memberikan kesan menarik dari suatu wilayah. Upaya-upaya ini akan membantu menjaga efektivitas kerangka kerja pada tingkat yang lebih tinggi. [2]

### 2.1.3 A Linear Regression Approach to Modeling Software Characteristics for Classifying Similar Software

Regresi linier merupakan suatu pendekatan di antara berbagai metode pembelajaran mesin yang diterapkan untuk memecahkan suatu masalah dengan membuat model dari pelatihan dengan data yang telah diketahui sebelumnya. Pada paper ini, mereka menyajikan metode regresi linier untuk memecahkan masalah di bidang analisis perangkat lunak untuk mengidentifikasi perangkat lunak serupa. Pada hasil eksperimen, mereka menyajikan hasil evaluasi penerapan regresi linier pada klasifikasi perangkat lunak serupa. Dengan mengevaluasi model yang dibangun, mereka membuat dataset uji dari ORO Jakarta dengan total 2.500 dataset. Kesimpulannya, model pembelajaran mesin yang dirancang dengan baik dapat digunakan secara efektif dalam analisis perangkat lunak. Penerapan pembelajaran mesin pada analisis perangkat lunak juga diharapkan dapat meningkatkan pemahaman tentang karakteristik perangkat lunak. [3]

### 2.1.4 An improvement to linear regression classification for face recognition

Klasifikasi regresi linier (LRC) telah menarik banyak perhatian karena kinerjanya yang menjanjikan dalam pengenalan wajah. Masalahnya kinerja model akan menurun jika sampel pelatihan yang terbatas. Dalam jurnal ini, sebuah metode baru berbasis LRC diusulkan untuk mengatasi masalah ini. Hasil percobaan pada database wajah ORL, FERET, Libor94 dan CMU-PIE menunjukkan bahwa metode yang diusulkan mendapatkan tingkat pengenalan yang lebih tinggi daripada beberapa metode pengenalan wajah yang canggih. Kesimpulannya, Metode klasifikasi Regresi Linier Konvensional (LRC) sering mengabaikan kesalahan dalam sampel pelatihan, yang menyebabkan kegagalan dalam tugas klasifikasi dunia nyata ketika jumlah sampel pelatihan per kelas tidak mencukupi. Dalam makalah ini, metode klasifikasi berbasis representasi diusulkan untuk secara bersamaan mempertimbangkan kesalahan dalam sampel pengujian dan pelatihan. Metode ini mengatasi kelemahan LRC dengan menghilangkan dampak negatif dari sampel pelatihan yang tidak mencukupi, sehingga mencapai akurasi yang lebih tinggi dibandingkan LRC. Evaluasi eksperimental menunjukkan efektivitas dan keunggulan metode ini. [4]

### 2.1.5 A Review on Linear Regression Comprehensive in Machine Learning

Paper ini membahas berbagai karya peneliti berbeda tentang regresi linier dan regresi polinomial dan membandingkan kinerjanya dengan metode terbaik untuk mengoptimalkan prediksi dan akurasi. Pada sebagian besar artikel yang dianalisis untuk menentukan efektivitas model, model harus dikorelasikan dengan nilai aktual yang diperoleh untuk variabel penjelas. Hasil makalah ini menggambarkan bahwa hampir semua penelitian memperkirakan model menggunakan kumpulan data, keakuratan model yang diukur dan kemampuan prediksi metode ini sangat penting, untuk mengukur efektivitas metode regresi, dilakukan studi perbandingan. antara nilai prediksi dan nilai sampel. [5]

# BAB 3 HASIL

## 3.1 PROSES PROGRAM REGRESI LINEAR

### 3.1.1 Sumber Dataset

Dataset yang digunakan pada laporan ini bersumber dari [Indonesian Food and Drink Nutrition Dataset (kaggle.com)](https://www.kaggle.com/datasets/anasfikrihanif/indonesian-food-and-drink-nutrition-dataset) yang memiliki 1346 baris, yang terdiri dari atribut id, calories, proteins, fat, carbohydrate, name, dan image.

### 3.1.2 Analisis Univariat

Analisis Univarite dilakukan untuk memahami karakter masing masing fitur, serta menjadi dasar untuk analisis bivariate. Dalam program ini, analisis Univariate membantu memahami rentang nilai kalori, protein, lemak, dan karbohidrat. Serta dapat membandingkan kandungan nutrisi pada berbagai jenis makanan serta mengidentifikasi makanan yang berpotensi kurang atau kelebihan nutrsi tertentu.

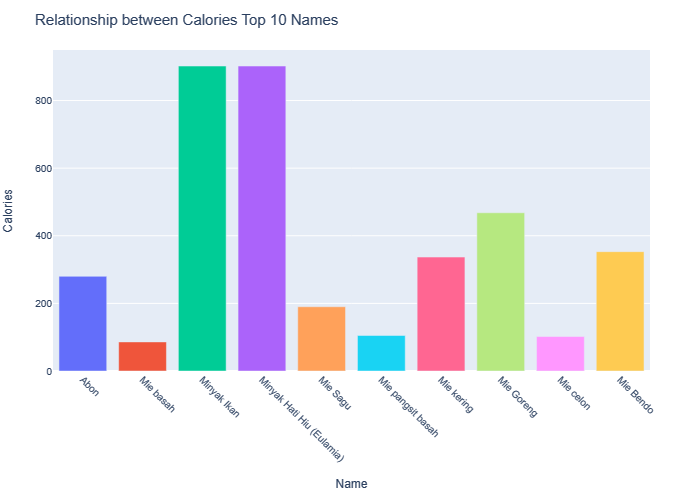
A graph of calories

Description automatically generated

Pada gambar distribusi kalori diatas, garis X menunjukkan nilai kalori dalam dataset, garis Y menunjukkan berapa banyak data yang terdapat pada setiap nilai kalori. Gambar tersebut mununjukkan bahwa distribusi data kalori tidak merata karena ada lebih banyak data pada nilai rendah (400 – lebih dari 800) dibanding dengan nilai kalori tinggi (0 – 400).

### 3.1.3 Analisis Bivariate

Analisis Bivariate dilakukan untuk mengidentifikasi pola hubungan antara dua fitur, memahami perbedaan kandungan lemak, protein, kalori, dan karbohidrat pada makanan yang berbeda. Pada tahap ini model mempelajari hubungan antara fitur sehingga dapat membantu dalam membuat prediksi, misalnya memperkirakan kandungan kalori suatu makanan berdasarkan kandungan lemak dan karbohidratnya.



Pada BarChart diatas merupakan hasil dari analisis Bivariate. Garis X menunjukkan 10 nama makanan terpopuler, diurutkan berdasarkan jumlah kemunculan dalam dataset. Nilai Y menunjukkan jumlah kalori pada setiap makananya. Dapat kita ketahui bahwa Minyak Ikan dan Minyak Hati Hiu (Eulamia) merupakan Outlier, karena memiliki nilai kalori jauh diatas rata-rata, sedangkan makanan lain kandungan kalorinya dibawah 450. Kesimpulannya Mie goreng memiliki kandungan kalori paling tinggi di antara 10 nama makanan terpopuler. Mie Celon dan Mie basah memiliki kandungan kalori yang relatif rendah. Minyak Ikan dan Minyak Hati Hiu (Eulamia) merupakan outlier dengan kandungan kalori yang jauh di atas rata-rata.

### 3.1.4 Mengidentifikasi Outliers

Mengidentifikasi Outlier (nilai yang sangat menyimpang) pada fitur-fitur yang dipilih dalam dataset dapat mempengaruhi hasil analisis data, sehingga penting untuk dideteksi dan ditangani dengan tepat. Kita dapat menghapus outlier jika dianggap kesalahan data, memodifikasi nilai outlier agar tidak terlalu ekstrem, menggunakan metode analisis yang robust terhadap outlier. Tujuannya untuk Memahami distribusi data dan melihat bentuk distribusi, nilai pusat, dan penyebaran data secara visual.

Outlier Sebelum :

A green rectangular object with black lines

Description automatically generated

Gambar BoxPlot diatas terdapat 4 Outlier dengan nilai kalori diatas 800, terlihat pada bagian kumis BoxPlot 4 titik data terletak diluar kumis (whisker) sebagai outlier. Kotak yang miring kekanan (positif skewed) menunjukkan distribusi data yang tidak normal.

Outlier Sesudah :

A graph with a red rectangle and black lines

Description automatically generated

Gambar BoxPlot diatas menunjukkan BoxPlot yang sudah tidak memiliki Outlier, membuat distribusi data jauh lebih normal setelah Outlier dihapus, sehingga tidak ada lagi data dengan nilai kalori yang jauh di atas rata-rata dan data lebih terpusat disekitar rata-rata.

### 3.1.5 Modeling dan Evaluasi Model

Modeling dilakukan untuk memprediksi jumlah kalori dalam produk dan mempelajari hubungan linier antara fitur dan variabel target. Evalusi model pada program ini untuk menghitung Mean Squared Error (MSE) dan R-Squared (R^2). MSE menunjukkan kesalahan rata-rata dalam satuan kuadrat variabel target. Semakin kecil MSE, semakin baik kinerja model. Sedangkan R^2 menunjukkan proporsi varians dalam variabel target yang dijelaskan oleh model. Nilai R^2 berkisar antara 0 hingga 1, dengan nilai 1 menunjukkan bahwa model menjelaskan semua varians dalam variabel target.

### 3.1.6 Hasil Model

A graph with blue dots

Description automatically generated

A black text on a white background

Description automatically generated

terdapat korelasi positif yang tinggi antara nilai kalori aktual dan nilai kalori prediksi, dapat dilihat pada garis lurus yang medekati diagonal. Kebanyakan titik data terdistribusi dekat dengan garis regresi, menunjukkan kesalahan prediksi umumnya kecil dan tidak ada pola sistematis dalam error prediksi.

Nilai MSE (Mean Squared Error) yang rendah membuktikan bahwa rata-rata error prediksi sangat kecil, dan cukup akurat dalam memprediksi nilai kalori. Hasil nilai R-Squared yang tinggi membuktikan bahwa model cukup baik dalam menamgkap hubungan antara variabel independen dan variabel dependen.

# BAB 4 PENUTUP

## 4.1 KESIMPULAN

Berdasarkan hasil evaluasi, model regresi linear yang dibuat memiliki tingkat akurasi yang cukup baik dalam memprediksi nilai kalori aktual dan nilai kalori prediksi. Hal ini ditunjukkan oleh nilai MSE, dan R-squared yang menunjukkan performa model yang baik. Model regresi linear yang dibuat dapat digunakan untuk memprediksi nilai kalori actual dengan tingkat akurasi yang cukup tinggi. Sangat penting untuk melakukan persiapan data dengan baik sebelum membangun model regresi linear. Pentingnya visualisasi data pada program ini memudahkan menganalisis performa model dan mengidentifikasi potensi masalah.

## 4.2 DAFTAR PUSTAKA

[1] M. Huang, “Theory and Implementation of linear regression,” in *Proceedings - 2020 International Conference on Computer Vision, Image and Deep Learning, CVIDL 2020*, Institute of Electrical and Electronics Engineers Inc., Jul. 2020, pp. 210–217. doi: 10.1109/CVIDL51233.2020.00-99.

[2] *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*.

[3] H. Il Lim, “A linear regression approach to modeling software characteristics for classifying similar software,” in *Proceedings - International Computer Software and Applications Conference*, IEEE Computer Society, Jul. 2019, pp. 942–943. doi: 10.1109/COMPSAC.2019.00152.

[4] Y. Peng, J. Ke, S. Liu, J. Li, and T. Lei, “An improvement to linear regression classification for face recognition,” *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 9, pp. 2229–2243, Sep. 2019, doi: 10.1007/s13042-018-0862-1.

[5] D. Maulud and A. M. Abdulazeez, “A Review on Linear Regression Comprehensive in Machine Learning,” *Journal of Applied Science and Technology Trends*, vol. 1, no. 4, pp. 140–147, Dec. 2020, doi: 10.38094/jastt1457.

## 4.3 LAMPIRAN

Untuk dataset dan source code silahkan akses [GitHub](https://github.com/Faikrom/UAS-PM-D-202131137-FITROHANANDA-IKROM.git)