Prediksi Risiko Atherosclerotic Heart Disease Menggunakan Logistic Regression: Pendekatan Machine Learning untuk Deteksi Dini dan Pencegahan Penyakit Jantung

Penyakit Jantung Aterosklerotik (Atherosclerotic Heart Disease, AHD) merupakan salah satu penyakit kardiovaskular yang paling umum dan sering kali menjadi penyebab utama kematian di berbagai negara di dunia. AHD terjadi ketika arteri koroner yang memasok darah ke jantung mengalami penyempitan atau penyumbatan akibat penumpukan plak aterosklerotik. Plak ini terdiri dari lemak, kolesterol, dan zat-zat lain yang dapat menyebabkan pembekuan darah dan berpotensi menyebabkan serangan jantung atau angina.

Faktor-faktor risiko yang berkontribusi terhadap pengembangan Atherosclerotic Heart Disease sangat bervariasi dan kompleks. Termasuk di antaranya adalah hipertensi, kolesterol tinggi, merokok, obesitas, kurangnya aktivitas fisik, dan pola makan tidak sehat. Di samping itu, faktor-faktor genetik dan keturunan juga dapat mempengaruhi kerentanan seseorang terhadap penyakit ini.

Dalam konteks ini, penting untuk dikembangkan metode yang dapat membantu mengidentifikasi individu yang berisiko tinggi untuk mengembangkan AHD secara dini. Prediksi risiko menggunakan model logistic regression dapat menjadi alat yang efektif untuk mencapai tujuan ini. Dengan menggunakan data klinis dan demografis yang relevan, model tersebut dapat membantu mengidentifikasi individu yang berisiko tinggi dan memungkinkan intervensi pencegahan yang tepat waktu.

Business Understanding

Pengembangan penelitian ini memberikan sejumlah manfaat yang signifikan dalam konteks pencegahan dan manajemen penyakit jantung aterosklerotik (AHD). Seperti di antaranya model prediktif yang dikembangkan dapat digunakan oleh praktisi kesehatan untuk memantau risiko kesehatan pasien mereka secara lebih efektif. Dengan memperhitungkan faktor-faktor risiko yang terkait dengan AHD, mereka dapat memberikan perawatan yang lebih individual dan tepat waktu kepada pasien yang berisiko.

Problem statement

1. Bagaimana mengembangkan model prediktif menggunakan machine learning untuk memprediksi kemungkinan terjadinya Atherosclerotic Heart Disease (AHD) berdasarkan faktor risiko yang telah diketahui?
2. Bagaimana mengelola dan memproses data klinis dan demografis yang kompleks untuk membangun model prediktif yang akurat dalam memperkirakan risiko individual terhadap Atherosclerotic Heart Disease (AHD)?
3. Bagaimana melakukan evaluasi dan pembandingan kinerja berbagai model logistic regression dalam memprediksi risiko Atherosclerotic Heart Disease (AHD) menggunakan metrik evaluasi yang sesua

Goals

1. Memperoleh pemahaman dalam mengembangkan model prediktif yang dapat memprediksi kemungkinan terjadinya Atherosclerotic Heart Disease (AHD) dengan mempertimbangkan berbagai faktor yang relevan.

2. Berusaha memahami serta mengelola data yang diperlukan sebagai dasar untuk membangun model prediktif yang dapat digunakan dalam memperkirakan risiko Atherosclerotic Heart Disease (AHD), dengan memperhatikan berbagai aspek penting dalam pengolahan dan analisis data.

3. Menyelidiki teknik-teknik evaluasi model yang ada dan berupaya untuk meningkatkan performanya agar menjadi lebih baik, sehingga memungkinkan penggunaan model tersebut dalam memprediksi risiko Atherosclerotic Heart Disease (AHD) dengan tingkat akurasi yang lebih tinggi.

Solution statement

1. Mengaplikasikan model logistic regression sebagai pendekatan utama untuk memprediksi kemungkinan terjadinya Atherosclerotic Heart Disease (AHD), yang dapat memberikan kerangka kerja yang kuat dan dapat diinterpretasikan dengan baik.

2. Melakukan persiapan data dengan mengaplikasikan teknik *encoding* untuk mengubah data kategorikal menjadi format numerik, *cleaning* data dari entri yang tidak valid atau tidak lengkap secara per baris, dan proses modifikasi data untuk mengubah data kategorial menjadi data numerik.

3. Melakukan standardisasi penyerataan range data yang besar dan berpotensi overpowering dengan *StandardScaler* dan menghitung koefisien model untuk memastikan bahwa variabel-variabel yang digunakan memiliki skala yang seragam dan tidak memiliki pengaruh yang tidak seimbang terhadap hasil prediksi

Data Understanding

Data yang digunakan dalam pembuatan model merupakan hasil dari generasi menggunakan perangkat lunak daring, di mana penentuan atribut-atribut didasarkan pada data yang diambil dari jurnal-jurnal terkait.

URL : https://www.mockaroo.com/

Berikut merupakan detail dari dataset yang digunakan untuk pembuatan model:

- Dataset berupa Excel
- Dataset terdiri dari 3030 records dengan 13 buah fitur yang diukur
- Dataset terdiri dari 2 data kategori dan 11 data numerik
- Dataset memiliki missing value sejumlah 303 records

Variabel-variabel Dataset

- Age : atribut yang mengukur usia pasien
- Sex : jenis kelamin pasien yang direpresentasikan dengan nilai biner, di mana nilai 0 mewakili wanita dan nilai 1 mewakili pria.
- Chest_Pain : jenis nyeri dada yang dialami oleh pasien. Typical Angina adalah nyeri dada yang biasanya dipicu oleh aktivitas fisik atau stres emosional dan cenderung mereda dengan istirahat atau penggunaan nitrogliserin. Atypical Angina yaitu nyeri dada dengan ciri-ciri yang tidak khas atau tidak sesuai dengan angina tipikal, tetapi masih bisa terkait dengan masalah jantung. Non-Anginal Pain adalah nyeri dada yang tidak terkait dengan masalah jantung, seperti nyeri otot atau gangguan pencernaan. Sementara, Asymptomatic atau tanpa gejala adalah pasien yang tidak mengalami nyeri dada atau gejala lain yang terkait dengan masalah jantung.
- Rest_BP : tekanan darah istirahat dari pasien, yaitu tekanan darah saat seseorang beristirahat atau tidak melakukan aktivitas fisik yang signifikan
- Serum_Cholesterol : kadar kolesterol serum dalam darah pasien
- Fasting_Blood_Sugar : kadar glukosa dalam darah pasien setelah puasa selama minimal delapan jam. Nilai 0 menunjukkan bahwa kadar gula darah puasa berada di bawah ambang batas yang ditetapkan sebagai normal. Sedangkan Nilai 1 menunjukkan bahwa kadar gula darah puasa berada di atas ambang normal, yang artinya hiperglikemia atau kadar gula darah yang tinggi.

- Rest_ECG : aktivitas listrik jantung saat pasien dalam keadaan istirahat. Nilai 0 menunjukkan hasil EKG normal atau tidak adanya abnormalitas. Nilai 1 menunjukkan adanya beberapa jenis abnormalitas pada elektrokardiogram istirahat, tetapi ini biasanya menunjukkan adanya beberapa tanda-tanda yang tidak spesifik atau ringan dari masalah jantung. Nilai 2 menunjukkan adanya abnormalitas yang lebih jelas atau signifikan pada elektrokardiogram istirahat
- Max_Heart_Rate : denyut jantung maksimum (maximum heart rate) yang dicapai oleh pasien selama tes olahraga atau aktivitas fisik yang dilakukan dalam studi tersebut
- Exercise_Induced_Angina : nyeri dada atau ketidaknyamanan yang terjadi karena aliran darah yang terbatas ke otot jantung yang dipicu oleh aktivitas fisik atau Latihan. Nilai 0 menunjukkan bahwa pasien tidak mengalami angina yang dipicu oleh latihan. Ini berarti bahwa aktivitas fisik atau latihan tidak menyebabkan nyeri dada atau ketidaknyamanan pada pasien. Nilai mengindikasikan bahwa pasien mengalami angina yang dipicu oleh latihan. Ini berarti bahwa aktivitas fisik atau latihan menyebabkan timbulnya nyeri dada atau ketidaknyamanan pada pasien.
- Oldpeak : depresi segmen ST (ST segment depression) yang terjadi selama tes latihan atau aktivitas fisik yang diukur dalam elektrokardiogram (ECG)
- Slope : kemiringan (slope) dari segmen ST selama tes latihan atau aktivitas fisik yang diukur dalam elektrokardiogram (ECG). Nilai 1 artinya kemiringan positif dari segmen ST selama aktivitas fisik atau tes. Artinya respons jantung normal terhadap stres fisik dan tidak menunjukkan adanya iskemia miokard yang signifikan. Nilai 2 mengindikasikan segmen ST yang datar selama aktivitas fisik atau tes latihan. Ini bisa menjadi tanda-tanda awal iskemia miokard atau menunjukkan adanya beberapa perubahan yang tidak spesifik dalam aktivitas listrik jantung. Nilai 3 menunjukkan kemiringan negatif dari segmen ST selama aktivitas fisik atau tes yang artinya terindikasi kuat dari iskemia miokard yang signifikan atau masalah jantung lainnya yang serius.
- Coronary_Arteries : jumlah pembuluh darah koroner yang terpengaruh atau tersumbat pada pasien
- Thalassemia : kondisi medis yang disebabkan oleh kelainan genetik yang mengganggu produksi hemoglobin dalam darah. Normal artinya pasien memiliki hemoglobin yang normal atau hanya memiliki kelainan genetik minor yang tidak menyebabkan anemia yang signifikan. Reversible artinya seseorang mengalami anemia karena thalassemia, tetapi anemia tersebut dapat diperbaiki atau dikoreksi dengan pengobatan atau terapi tertentu. Sedangkan fixed merupakan bentuk thalassemia yang lebih parah atau tidak dapat diperbaiki, yang menghasilkan anemia kronis dan memerlukan manajemen medis yang terus menerus.
- Atherosclerotic_Heart_Disease : keberadaan atau tidaknya penyakit jantung aterosklerotik pada pasien

Analisis Data

Analisis awal data dimulai dengan menganalisis data secara visual menggunakan teknik univariat dan bivariat. Analisis univariat adalah proses memeriksa satu variabel pada suatu waktu, sedangkan analisis bivariat melibatkan pembandingan antara dua variabel atau lebih. Dalam konteks ini, analisis bivariat dilakukan dengan membandingkan setiap faktor yang terlibat dengan hasil kejadian terkena AHD. Hasil dari kedua jenis analisis ini kemudian dipresentasikan melalui bentuk visualisasi. Contohnya adalah gambar-gambar dalam Exploratory Data Analysis (EDA), di mana Gambar 1 menggambarkan EDA Analisis Univariat dan Gambar 2 menggambarkan EDA Analisis Bivariat. Melalui visualisasi ini, kita dapat memperoleh pemahaman yang lebih baik tentang karakteristik data serta hubungan antara variabel-variabel yang ada. Berikut adalah gambar 1, yaitu kumpulan data visualisasi bivariat.

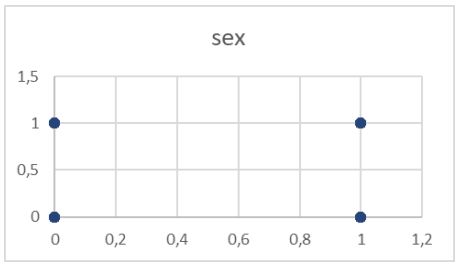
Gambar 1. Visualisasi Data Univariat

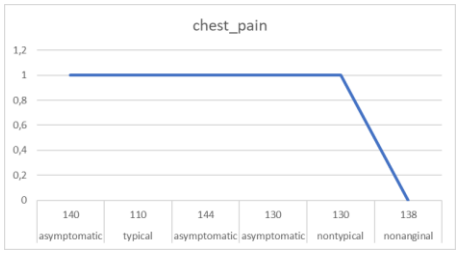
Berdasarkan hasil analisis yang dilakukan, beberapa temuan dapat diidentifikasi sebagai berikut:

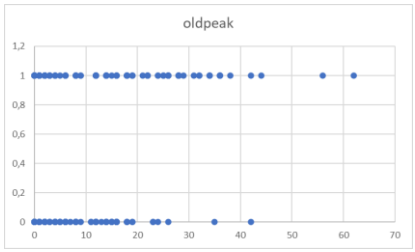
- Rentang usia individu dalam dataset bervariasi antara 29 hingga 77 tahun, sementara mayoritas subjek adalah laki-laki dengan kode biner 1 untuk variabel sex.

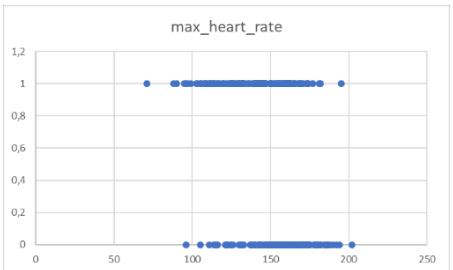
- Nilai maksimum denyut jantung (max heart rate) berkisar antara 71 hingga 202, sementara kolesterol serum (serum cholesterol) berkisar dari 126 hingga 564, dan tekanan darah saat istirahat (rest bp) berkisar dari 94 hingga 200. Nilai oldpeak berkisar antara 0 hingga 62.
- Mayoritas kasus nyeri dada (chest pain) teridentifikasi sebagai tipe asymptomatic, dengan jumlah kasus tipe typical cenderung lebih sedikit.
- Atribut thalassemia didominasi oleh tiga kondisi secara berurutan, yaitu normal, reversable, dan fixed.
- Mayoritas individu memiliki nilai fasting blood sugar yang bernilai 0, menunjukkan bahwa mayoritas memiliki kadar gula darah yang normal.
- Hasil uji elektrokardiogram saat istirahat (rest ECG) menunjukkan mayoritas subjek tidak menunjukkan tanda-tanda abnormalitas, meskipun beberapa menunjukkan tanda-tanda yang jelas. Namun, terdapat beberapa subjek yang menunjukkan tanda-tanda yang samar.
- Dari semua atribut, mayoritas subjek memiliki nilai slope yang normal, sementara beberapa menunjukkan indikasi awal.
- Mayoritas subjek tidak mengalami angina yang dipicu oleh latihan (exercise induced angina), yang menunjukkan bahwa mayoritas tidak merasakan nyeri akibat aktivitas fisik.
- Mayoritas individu dalam dataset ini tidak terkena Atherosclerotic Heart Disease (AHD), yang menunjukkan dominasi subjek yang tidak terkena kondisi tersebut.

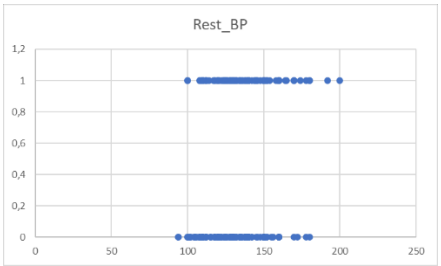
Selanjutnya, dilakukan analisis data bivariat, yang membandingkan dua data dari dua variabel yang berbeda, yaitu untuk setiap nilai x dengan nilai y. Analisis bivariat penting karena memungkinkan kita untuk memahami hubungan antara dua variabel dan bagaimana keduanya berinteraksi. Dengan membandingkan nilai variabel yang berbeda, kita dapat melihat pola atau tren yang mungkin terjadi dalam data dan menarik kesimpulan yang lebih dalam tentang hubungan antara variabel tersebut. Berikut adalah gambar 2, yaitu kumpulan data visualisasi bivariat.

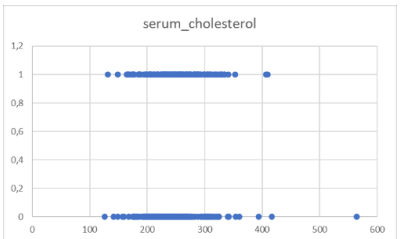


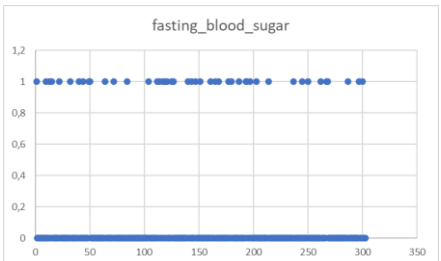


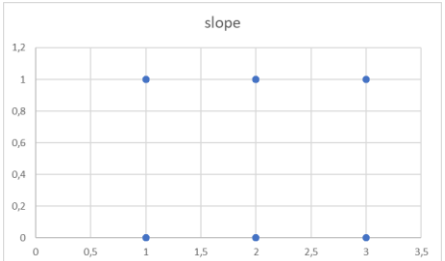


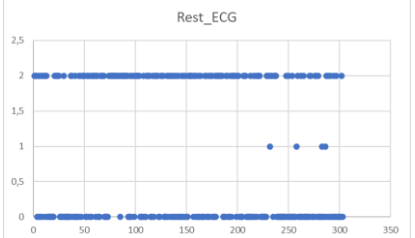


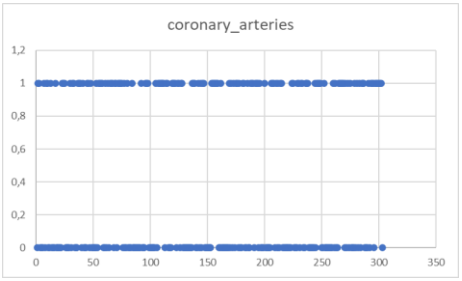












Gambar 2. Visulisasi Data Bivariat

Berdasarkan hasil analisis yang dilakukan, beberapa temuan dapat diidentifikasi sebagai berikut:

- Pasien yang mengalami chest pain memiliki risiko lebih tinggi terkena Atherosclerotic Heart Disease (AHD), kecuali pada jenis non-anginal. Ini menandakan bahwa gejala chest pain, terutama pada jenis tertentu, dapat menjadi indikator penting dalam diagnosis AHD.
- Ditemukan bahwa jumlah pasien yang terkena AHD cenderung lebih sedikit yang memiliki riwayat fasting blood sugar yang tinggi. Mayoritas pasien dengan kadar fasting blood sugar normal atau rendah, tidak terkena AHD. Hal ini menunjukkan bahwa kadar fasting blood sugar dapat menjadi faktor penentu dalam risiko terkena AHD.
- Atribut-atribut sisanya dalam dataset menunjukkan distribusi data yang cukup merata antara pasien yang terkena AHD dan yang tidak terkena. Hal ini menandakan bahwa variasi atribut-atribut tersebut tidak secara signifikan terkait dengan keberadaan atau ketidakterkenaan AHD, dan mungkin memerlukan analisis lebih lanjut untuk memahami hubungannya dengan kondisi tersebut

Data preparation

Selama proses Persiapan Data, beberapa langkah yang penting termasuk sebagai berikut:

a) Pengumpulan Data

Data diimpor ke dalam format yang dapat dibaca dengan mudah menggunakan dataframe Pandas.

b) Evaluasi Data

Tahap ini melibatkan penggunaan Python untuk mengevaluasi jumlah data yang tidak valid atau memiliki nilai NaN. Data tersebut kemudian dibersihkan dengan menggunakan metode penghapusan untuk menghilangkan data yang hilang.

c) Transformasi Data

Langkah terakhir melibatkan transformasi data di mana variabel kategorikal diubah menjadi bentuk numerik. Proses ini dilakukan menggunakan dua metode, yaitu:

- Metode astype: Digunakan dalam Python untuk mengubah tipe data dari suatu objek menjadi tipe data yang diinginkan.

- Metode cat.codes: Digunakan untuk mengambil representasi numerik dari nilai-nilai dalam sebuah Series pada Pandas DataFrame setelah proses encoding kategori menggunakan cat.codes. Metode ini membantu dalam mengubah data kategori menjadi data numerik untuk analisis lebih lanjut.

Modeling

Pemodelan penelitian ini menggunakan metode regresi, yaitu metode statistik yang digunakan untuk memahami dan memodelkan hubungan antara satu atau lebih variabel independen (X) dengan

variabel dependen (Y). Tujuan utama regresi adalah untuk memprediksi atau menjelaskan nilai variabel dependen berdasarkan nilai-nilai variabel independen yang diberikan. Regresi sering digunakan dalam analisis data untuk memahami dan memprediksi hubungan antara variabel-variabel yang terkait.

Dalam regresi, variabel independen (X) digunakan untuk memprediksi atau menjelaskan variabel dependen (Y) dengan memanfaatkan hubungan matematis di antara keduanya. Misalnya, dalam regresi linier sederhana, hubungan antara dua variabel dapat dinyatakan sebagai garis lurus:

$$Y = \beta 0 + \beta 1 X$$

Di mana :

Y adalah variabel dependen.

X adalah variabel independen.

$\beta 0$ adalah intercept (perpotongan garis regresi dengan sumbu Y)

$\beta 1$ adalah koefisien regresi (gradien garis regresi).

Kelebihan Regresi:

- Memberikan pemahaman tentang hubungan antara variabel-variabel
- Memungkinkan untuk memprediksi nilai variabel dependen berdasarkan nilai variabel independen
- Sederhana dan mudah diinterpretasikan

Kekurangan Regresi:

- Asumsi tentang hubungan linier antara variabel mungkin tidak selalu terpenuhi
- Rentan terhadap pengaruh oleh outlier atau anomali dalam data
- Memerlukan perhatian terhadap asumsi-asumsi dasar seperti homoskedastisitas dan independensi dari kesalahan

Secara lebih spesfik, pemodelan ini menggunakan logistic regression yang mana merupkana salah satu hasil turunan dari salah satu teknik regresi yaitu regresi linear. Meskipun berbeda dalam bentuk dan tujuan dengan regresi linear, namun secara konseptual, logistic regression dapat dianggap sebagai turunan dari regresi linear. Ini karena dasar matematis logistic regression sebagian besar berasal dari regresi linear. Secara formal, model logistic regression dapat dinyatakan sebagai berikut:

$$p\left(Y = 1 \,\middle|\, X \right) = \frac{1}{1 + e^{-(\beta 0 + \beta 1 X)}}\Big)$$

Di mana :

$p(Y = 1 \mid X)$ adalah probabilitas bahwa Y sama dengan 1 (variabel dependen biner) dengan diberikan nilai X.

$\beta 0$ adalah intercept

$\beta 1$ adalah koefisien regresi

Kelebihan Logistic Regression:

- Cocok untuk kasus di mana variabel dependen adalah biner
- Hasilnya mudah diinterpretasikan dalam bentuk probabilitas
- Tidak memerlukan asumsi tentang distribusi dari variabel dependen.

Kekurangan Logistic Regression:

- Tidak cocok untuk masalah dengan variabel dependen kontinu
- Rentan terhadap overfitting dengan fitur-fitur yang banyak
- Tidak memperhitungkan hubungan antara variabel independen

Oleh karena itu, logistic regression menjadi sebuah alat yang sangat berguna dalam memprediksi probabilitas kejadian biner berdasarkan variabel independen. Teknik ini telah terbukti bermanfaat dalam berbagai bidang, termasuk analisis risiko, kesehatan, dan pemasaran, di mana kebutuhan untuk memahami dan memprediksi hasil berbasis probabilitas sangat penting.

Di samping itu, sebelum melakukan pemodelan pada data yang telah terbagi, penting untuk melakukan standardisasi terlebih dahulu. Hal ini dikarenakan rentang nilai yang sangat besar dalam dataset dapat menyebabkan overpowering, yang dapat menghasilkan bias dalam model. Dengan demikian, standardisasi data diperlukan untuk memastikan keakuratan dan kebersihan dataset yang digunakan dalam pembangunan model, sehingga hasil prediksi yang dihasilkan lebih dapat diandalkan.

Result

Dari hasil modeling, didapatkan hasil sebagai berikut:

- Coefficients: [[-0.02019491 0.56756264 -0.75586373 0.45897328 0.11951092 -0.25701509 0.26884793 -0.43003966 0.44812257 0.3428912 0.440473 1.07625574 0.52813531]]
- Intercept: [-0.22101055]

Dengan demikian, persamaan model untuk masalah ini adalah:

$$\begin{aligned} Y = {} & -0.02019491(age) + 0.56756264(sex) - 0.75586373(\text{chest\_pain}) \\ & + 0.45897328(\text{rest\_bp}) + 0.11951092(\text{serum\_cholesterol}) \\ & - 0.25701509(\text{fasting\_blood\_sugar}) + 0.26884793(\text{rest\_ecg}) \\ & - 0.43003966(Max\_Heart) + 0.44812257(exercise\_induced\_angina) \\ & + 0.3428912(oldpeak) + 0.440473(slope) + 1.07625574(coronary_{arteries}) \\ & + 0.52813531(thalassemia) - 0.22101055 \end{aligned}$$

Evaluation

Setelah membangun model, evaluasi dilakukan untuk memastikan bahwa data yang digunakan cocok atau layak untuk digunakan sebagai dasar pembuatan model prediktif, dan model yang dihasilkan mampu melakukan prediksi dengan tingkat akurasi yang tinggi. Salah satu metode yang digunakan untuk evaluasi ini disebut 'score'.

Metode 'score' dalam logistic regression (dan juga dalam banyak model machine learning lainnya) digunakan untuk mengevaluasi kinerja model pada dataset uji. Khususnya, metode 'score' pada logistic regression menghitung akurasi model pada dataset uji. Dari pembuatan model prediktif di atas, didapatkan nilai untuk data training sebesar 87% dan data testing sebesar 86%, yang menunjukkan hasil prediksi yang baik.

Akurasi adalah salah satu metrik evaluasi yang umum digunakan untuk model klasifikasi, termasuk logistic regression. Metrik ini mengukur persentase prediksi yang benar dibandingkan dengan jumlah total sampel dalam dataset uji.

Penggunaan metode 'score' dalam logistic regression sangat penting karena membantu dalam mengevaluasi seberapa baik model bekerja dalam mengklasifikasikan data baru yang tidak digunakan dalam proses pelatihan model. Semakin tinggi nilai akurasi yang dihasilkan oleh metode 'score', semakin baik kinerja model dalam memprediksi kelas dari sampel-sampel yang tidak dikenal. Dengan demikian, metode 'score' adalah alat penting dalam proses evaluasi dan validasi model logistic regression.