I

UB

Big Data Assignment #3

Logistic Regression: Heart Attack Posibility

9 April 2021

DIANNA AGUSTINA S. | 1202912003

Logistic Regression

Logistic Regression atau Regresi logistik (kadang disebut model logistik atau model logit), dalam statistika digunakan untuk prediksi probabilitas kejadian suatu peristiwa dengan mencocokkan data pada fungsi logit kurva logistik. Metode ini merupakan model linier umum yang digunakan untuk regresi binomial. Seperti analisis regresi pada umumnya, metode ini menggunakan beberapa variabel prediktor, baik numerik maupun kategori.

III

Heart Attack Dataset

Intro

Dataset berisi sekumpulan data tentang kondisi seseorang dimana kondisi tersebut dapat berpengaruh terhadap timbulnya serangan jantung.

Info

Set data diambil untuk tujuan pembelajaran. Sumber data: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Sumber Dataset :

https://www.kaggle.com/nareshbhat/health-care-data-set-on-heart-attack-possibility

Business Understanding

Ruang Lingkup Bisnis : Kesehatan

Dalam dunia kesehatan, serangan jantung dapat terjadi sewaktu-waktu dan pada siapapun. Terdapat beberapa parameter yang dapat mempengaruhi kemungkinan terjadinya serangan jantung, diantaranya tekanan darah, kolesterol, gula darah, detak jantung maksimal, dan lain sebagainya. Untuk itu, perlu dilakukan penelitian seberapa besar kemungkinan seseorang dapat mengalami serangan jantung dengan kondisi yang ada pada saat ini.

Data Understanding

Field Name	Description	Data Type
age	*Usia*	int64
sex	*Jenis kelamin*	int64
cp	*Jenis nyeri dada (4 jenis)*	int64
trestbps	*Tekanan darah istirahat*	int64
chol	*Kolesterol dalam mg/dl*	int64
fbs	*Gula darah puasa > 120 mg/dl*	int64
restecg	*Hasil elektrokardiografi istirahat (3 nilai)*	int64
thalach	*Detak jantung maksimal tercapai*	int64
exang	*Latihan angina yang diinduksi*	int64
oldpeak	*Depresi yang disebabkan oleh olahraga*	float64

V

Data Understanding

Field Name	Description	Data Type
slope	*Kemiringan segmen latihan puncak*	int64
ca	*Jumlah pembuluh utama*	int64
thal	*Kecacatan (3 jenis)*	int64
target	*Kemungkinan serangan jantung*	int64

Data Preparation

Melakukan pengecekan data

Data dilakukan pengecekan terhadap informasi seperti jumlah data (baris) dan kolom, tipe data dari setiap kolom, data yang memuat nilai null.

Mengecek ringkasan statistik dan hubungan antar variabel

Pengecekan ringkasan data statistik dengan menggunakan fungsi describe() pada DataFrame. Ringkasan data statistik yang ditampilkan adalah jumlah data, nilai min, max, standar deviasi, median, mean, kuartil bawah, kuartil tengah, dan kuartil atas.

Mengecek variabel signifikan

Variabel signifikan dapat dilihat dengan menggunakan library statsmodel.api. Dari hasil perhitungan, terdapat variabel insignifikan dengan nilai p-value dibawah 0.05 yaitu pada kolom age, trestbps, chol, fbs, restecg, dan slope. Maka, kolom dimaksud dihapus dalam DataFrame.

X

Modeling (Training)

Menentukan data training dan data testing

Data dibagi menjadi data training dan data testing dengan komposisi 50:50. Pembagian data menggunakan fungsi train_test_split pada library sklearn.

Melakukan Normalisasi

Normalisasi data dilakukan dengan menggunakan fungsi StandardScaler() terhadap variabel independen data training dan data testing. Fungsi StandarScaler() mengubah data mean menjadi 0 dan varian 1.

Menghitung Nilai Intercept

Menghitung nilai intercept dengan menggunakan fungsi LogisticRegression() terhadap data training. Nilai intercept yang didapatkan adalah 0.08730542

Evaluation (Testing and Accuracy)

Pengetesan (Testing)

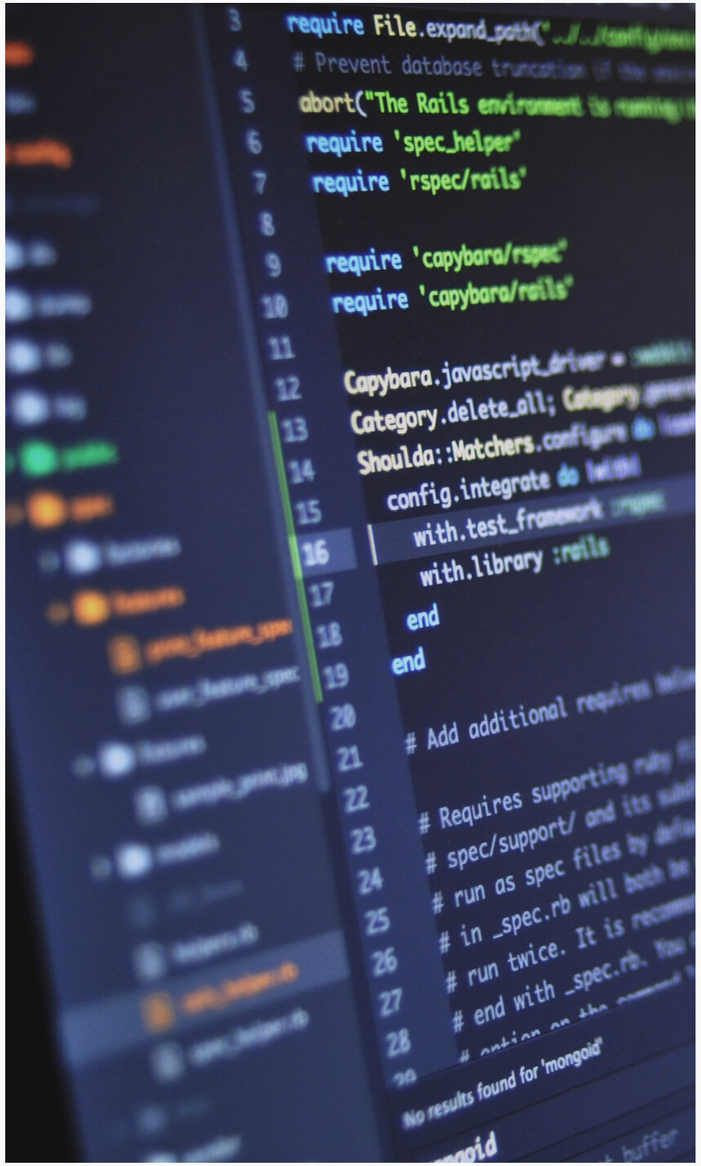
Testing dilakukan dengan menjalankan fungsi LogisticRegression() terhadap data testing. Proses ini menghasilkan data prediksi klasifikasi 0 atau 1 pada setiap data.

Penambahan Kolom Hasil Penghitungan

Penambahan kolom nilai probabilitas dan klasifikasi dilakukan terhadap data testing. Penambahan ini untuk menampilkan nilai probabilitas dan klasifikasi pada setiap data.

Penghitungan Akurasi

Pengecekan akurasi dengan menggunakan fungsi classification_report dan menghasilkan nilai akurasi 0.79. Nilai ini menunjukan bahwa akurasi pemodelan yang telah dibuat sebesar 79%.



Deployment

Tahap ini dilakukan menggunakan bahasa pemrograman Python dengan IDE Visual Studio Code.

Code secara keseluruhan terdapat pada link Github :

https://github.com/dyanaagustina/Learn-BigData/tree/Big-Data/LR

Thank You