I

UB

Big Data Assignment #4 Multinomial Logistic Regression: Student Grade

16 April 2021

DIANNA AGUSTINA S. | 1202912003

II

UB

Multinomial Logistic Regression

Regresi logistik merupakan salah satu metode yang dapat digunakan untuk mencari hubungan variabel respon yang bersifat dichotomous (berskala nominal atau ordinal dengan dua kategori) atau polychotomous (mempunyai skala nominal atau ordinal dengan lebih dari dua kategori) dengan satu atau lebih variabel prediktor dan variabel respon bersifat kontinyu atau kategorik. Regresi logistik multinomial atau disebut juga model logit politomus adalah model regresi yang digunakan untuk menyelesaikan kasus regresi dengan variabel dependen berupa data kualitatif berbentuk multinomial (lebih dari dua kategori) dengan satu atau lebih variabel independen.

III

Student Grade Dataset

Info

Dataset berisi sekumpulan data terkait kebiasaan dari siswa sekolah menengah atas dengan mata pelajaran Matematika dan hasil nilai akhir mata pelajaran dimaksud.

Sumber Dataset :

https://www.kaggle.com/uciml/student-alcohol-consumption

Business Understanding

IV

UB

Ruang Lingkup Bisnis : Education

Suatu sekolah menengah atas akan melakukan perhitungan atau prediksi terkait kelulusan nilai yang akan didapatkan oleh siswanya. Dengan memperhatikan variabel-variabel yang dapat mempengaruhi nilai kelulusan, terutama pengaruh penggunaan alkohol di kehidupan sehari-hari. Output dari perhitungan ini akan menghasilkan formula untuk dapat memprediksi kelulusan dari siswa pada sekolah dimaksud.

Data Understanding

Field Name	Description	Data Type
school	*Nama sekolah*	object
sex	*Jenis kelamin*	object
age	*Usia*	int64
address	*Alamat siswa*	object
famsize	*Jumlah saudara*	object
Pstatus	*Hubungan kedua orang tua*	object
Medu	*Pendidikan ibu*	int64
Fedu	*Pendidikan ayah*	int64
Mjob	*Pekerjaan ibu*	object
Fjob	*Pekerjaan ayah*	object

Data Understanding

Field Name	Description	Data Type
reason	*Alasan memilih sekolah*	object
guardian	*Wali siswa*	object
traveltime	*Waktu tempuh dari rumah ke sekolah*	int64
studytime	*Waktu belajar mingguan*	int64
failures	*Jumlah mengalami kegagalan*	int64
schoolsup	*Support pendidikan tambahan*	object
famsup	*Support keluarga*	object
paid	*Biaya ekstra*	object
activities	*Mengikuti ekstrakulikuler*	object
nursery	*Mengikuti sekolah keperawatan*	object

Data Understanding

Field Name	Description	Data Type
higher	*Keinginan untuk mengambil pendidikan tinggi*	object
internet	*Akses internet dirumah*	object
romantic	*Memiliki hubungan romantis*	int64
famrel	*Kualitas hubungan keluarga*	int64
freetime	*Waktu luang setelah sekolah*	int64
goout	*Menghabiskan waktu bersama teman*	int64
Dalc	*Konsumsi alkohol harian*	int64
Walc	*Konsumsi alkohol mingguan*	int64
health	*Kondisi kesehatan*	int64
absences	*Jumlah absen*	int64

Data Understanding

Field Name	Description	Data Type
G1	*Nilai grade 1*	int64
G2	*Nilai grade 2*	int64
G3	*Nilai grade 3 - Kelulusan*	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.

◆ Menghapus kolom dan mengecek deskripsi data

Penghapusan kolom dilakukan terhadao kolom dengan tipe data **object. Dari total 32 kolom variabel independen, menjadi hanya 8 kolom variabel independen.** Pengecekan deskripsi data dilakukan dengan menggunakan fungsi describe(). Hasil yang ditampilkan jumlah data, nilai min, max, standar deviasi, median, mean, kuartil bawah, kuartil tengah, dan kuartil atas.

◆ Melakukan normalisasi

Normalisasi dilakukan dengan menggunakan fungsi **StandardScaler(),** dimana nilai *mean* dijadikan 0 dan *variance* 1. Normalisasi dilakukan terhadap seluruh data independen variabel.

X

Modeling (Training)

Mengecek data setiap kategori

Pengecekan terhadap koefisien dan intercept yang didapatkan dengan menggunakan menggunakan fungsi MNLogit().

Menghitung Nilai Intercept

Menghitung nilai intercept dengan menggunakan fungsi **LogisticRegression**(). Pada proses ini menghasilkan nilai intercept setiap kategori yang akan digunakan dalam perhitungan.

Menghitung Nilai Koefisien

Nilai koefisien didapatkan dengan fungsi **coef_** pada pemodelan **LogisticRegression**() dengan parameter multi_class='multinomial'. Fungsi ini menghasilkan nilai koefisien pada setiap variabel independen pada setiap kategori.

XI

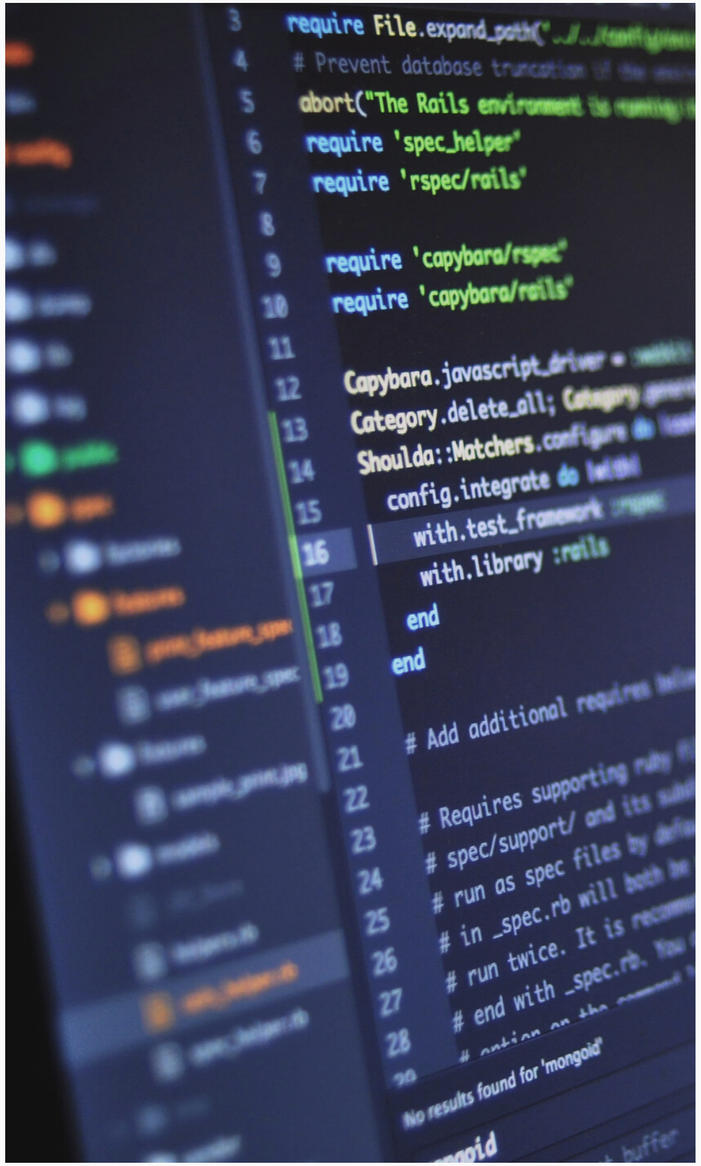
Evaluation (Testing and Accuracy)

Pemodelan Manual

Pada tahap ini, pemodelan dilakukan secara manual dengan menggunakan rumus multinomial logistic regression, dimana dalam perhitungan ini membutuhkan nilai intercept, nilai koefisien variabel independen, dan nilai setiap variabel independen.

Penghitungan Akurasi

Pengecekan akurasi dengan menggunakan fungsi classification_report dan menghasilkan nilai akurasi 0.81. Nilai ini menunjukan bahwa akurasi pemodelan yang telah dibuat sebesar 81%.



Deployment

Tahap ini dilakukan menggunakan bahasa pemrograman Python dengan IDE Kaggle.

Code secara keseluruhan terdapat pada link Github :

https://github.com/dyanaagustina/Learn-BigData/tree/Big-Data/Multinomial%20LR

Thank You