I

UB

Big Data Assignment #5 Discriminant Analysis: Employee Resilience

23 April 2021

DIANNA AGUSTINA S. | 1202912003

Discriminant Analysis

Analisis diskriminan adalah teknik untuk menganalisis data ketika kriteria atau variabel dependen bersifat kategorikal dan prediktor atau variabel independen bersifat interval. Terdapat dua jenis analisis diskriminan yaitu *linear discriminant analysis* dan *multiple* discriminant analysis. Analisis diskriminan linear adalah suatu metode yang digunakan dalam ilmu statistika, pengenalan pola dan pembelajaran mesin untuk mencari kombinasi linear fitur yang menjadi ciri atau yang memisahkan dua atau beberapa objek atau peristiwa. Kombinasi yang diperoleh dapat dijadikan pengklasifikasi linear, atau biasanya digunakan untuk proses reduksi dimensionalitas sebelum pengklasifikasian. Analisis diskriminan adalah teknik untuk menganalisis data ketika kriteria atau variabel dependen bersifat kategorikal dan prediktor atau variabel independen bersifat interval.

III

Human Resources (HR) Dataset

Info

Dataset berisi sekumpulan data terkait performa pegawai dalam suatu perusahaan, dimana performa tersebut dapat mempengaruhi ketahanan pegawai pada perusahaan tersebut.

Sumber Dataset :

https://www.kaggle.com/liujiaqi/hr-comma-sepcsv

Business Understanding

Ruang Lingkup Bisnis : Manajemen

Suatu perusahaan memiliki sekumpulan informasi mengenai performa para pegawainya, dimana pada perusahaan tersebut terdapat penurunan jumlah pegawai secara drastis dalam kurun waktu 1 tahun terakhir. Dengan memperhatikan sekumpulan data pencapaian dan performa dari setiap pegawai, perusahaan dimaksud membutuhkan penelitian terkait prediksi ketahanan pegawai. Hal ini diperlukan agar tidak terjadi penurunan pegawai secara drastis terutama disaat banyak program prioritas yang perlu dicapai oleh perusahaan.

V

Data Understanding

Field Name	Description	Data Type
satisfaction_level	*Tingkat kepuasan*	float64
last_evaluation	*Penilaian evaluasi*	float64
number_project	*Jumlah project*	int64
average_monthly_hours	*Jumlah jam kerja rata-rata per bulan*	int64
time_spend_company	*Lama bekerja (tahun)*	int64
work_accident	*Mengalami kecelakaan*	int64
left	*Ketahanan pegawai*	int64
promotion_last_5years	*Promosi dalam kurun waktu 5 tahun terakhir*	int64
sales	*Divisi*	object
salary	*Kategori gaji*	object

V

Data Understanding

Left (Dependent Variable)

0	*Pegawai bertahan*
1	*Pegawai meninggalkan perusahaan*

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,** yaitu kolom sales dan kolom salary. Pengecekan 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)

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**.

Menghitung Nilai Intercept

Menghitung nilai intercept dengan menggunakan fungsi **LinearDiscriminantAnalysis**() terhadap data training. Nilai *intercept* yang didapatkan adalah -1.48904201

Menghitung Nilai Koefisien

Nilai koefisien didapatkan dengan fungsi **coef_** pada pemodelan **LinearDiscriminantAnalysis**(). Fungsi ini menghasilkan nilai koefisien sesuai jumlah variabel independen yang telah ditentukan.

Evaluation (Testing and Accuracy)

Pengetesan (Testing) - Predict

Testing dilakukan dengan menjalankan fungsi **predict()** pada **LinearDiscriminantAnalysis()** yang telah dimodelkan dengan menggunakan variabel testing. Fungsi ini menghasilkan nilai 1 atau 0 pada setiap data.

Pengetesan (Testing) - Probability

Testing dilakukan dengan menjalankan fungsi **predict_proba()** pada **LinearDiscriminantAnalysis()** yang telah dimodelkan dengan menggunakan variabel testing. Fungsi ini menghasilkan nilai *probability* pada setiap data.

Penghitungan Akurasi

Pengecekan akurasi dengan menggunakan fungsi classification_report dan menghasilkan nilai akurasi 0.76. Nilai ini menunjukan bahwa akurasi pemodelan yang telah dibuat sebesar 76%.

Evaluation (Testing and Accuracy)

Pemodelan Manual - *Discriminant Score*

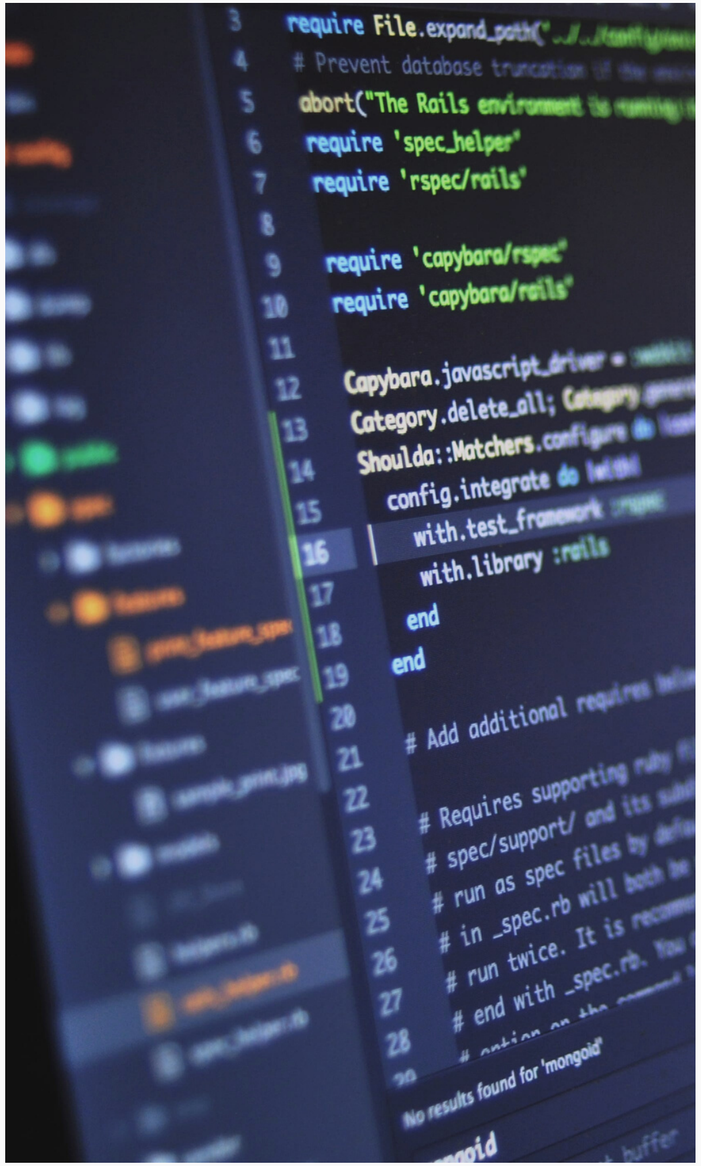
Pada tahap ini, pemodelan dilakukan secara manual dengan menggunakan rumus *discriminant score*, dimana dalam perhitungan ini membutuhkan nilai intercept, nilai koefisien variabel independen, dan nilai setiap variabel independen.

Pemodelan Manual - *Z Score*

Pada tahap ini, pemodelan dilakukan secara manual dengan menggunakan rumus z score, dimana dalam perhitungan ini hanya membutuhkan nilai koefisien variabel independen, dan nilai setiap variabel independen.

Perbandingan Nilai

Hasil dari ketiga perhitungan (Fungsi LinearDiscriminantAnalysis(), *Discriminant Score*, *dan Z Score*) kemudian dibandingkan. Dalam perbandingan hasil tersebut, terdapat perbedaan jumlah pada setiap prediksi.



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/LDA

LinearDiscriminantAnalysis()

13,166	*Pegawai bertahan*
1,833	*Pegawai meninggalkan perusahaan*

Discriminant Score (D-Score)

7,707	*Pegawai bertahan*
7,292	*Pegawai meninggalkan perusahaan*

Z-Score

7,704	*Pegawai bertahan*
7,295	*Pegawai meninggalkan perusahaan*

Thank You