TUGAS PRAKTIKUM 4 PENGENALAN REGRESI

MATA KULIAH INFRASTRUKTUR DAN PLATFORM UNTUK SAINS DATA

Rabu, 23 Oktober 2024

NAMA: IKA WIDA NURAGUSTIN
NIM: 2311110001
KELAS: S1SD-04-01

Kasih komen penejelasan di code pertemuan 6, interpretasikan hasil R^2 *Score* dan MSE nya!

1. *Code Cell* Pertama

```
from sklearn.datasets import load_diabetes
from sklearn.linear_model import LinearRegression, Lasso, Ridge
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
```

Penjelasan:

Kode diatas digunakan untuk memanggil *library* di python *(import library).* load_diabetes dari sklearn.datasets digunakan untuk mengambil dataset, sedangkan LinearRegression, Lasso, dan Ridge dari sklearn.linear_model digunakan untuk melatih tiga model regresi yang berbeda. Fungsi train_test_split dari sklearn.model_selection membagi dataset menjadi data pelatihan dan pengujian, dan mean_squared_error serta r2_*score* dari sklearn.metrics digunakan untuk mengevaluasi performa model dengan metrik *Mean Squared Error* (MSE) dan R^2 *score.*

2. *Code Cell* Kedua

```
x, y = load_diabetes(return_X_y=True)
```

Penjelasan:

Kode diatas digunakan untuk mengupload data/*load* data diabetes. Kode x, y = load_diabetes(return_X_y=True) memuat dataset diabetes dan membaginya menjadi variabel fitur (x) dan target (y). x berisi data variabel independen seperti usia dan BMI, sedangkan y berisi tingkat progresi diabetes, yang menjadi target prediksi.

3. *Code Cell* Ketiga

```
lr = LinearRegression()
```

Penjelasan:

Kode lr = LinearRegression() membuat objek model regresi linier bernama lr dari pustaka sklearn. Model ini dapat digunakan untuk mempelajari hubungan linier antara fitur dan target dalam data.

4. *Code Cell* Keempat

```
len(load_diabetes()['feature_names'])
```

```
10
```

Penjelasan:

Kode diatas digunakan untuk mengetahui jumlah elemen didalam dataset. Kode len(load_diabetes()['feature_names']) menghitung jumlah fitur dalam dataset diabetes. Fungsi load_diabetes() memuat dataset diabetes, dan ['feature_names'] mengambil daftar nama fitur, kemudian len() digunakan untuk menghitung jumlah elemen dalam daftar tersebut. Dari *output* diketahui bahwa banyaknya elemen dalam daftar adalah 10.

5. *Code Cell* Kelima

```
lr.fit(x, y)
y_pred = lr.predict(x)
```

Penjelasan:

Kode lr.fit(x, y) melatih model regresi linier lr dengan menggunakan data fitur x dan target y, memungkinkan model mempelajari hubungan linier antara keduanya. Selanjutnya, kode y_pred = lr.predict(x) digunakan untuk menghasilkan prediksi nilai target berdasarkan data fitur x yang telah dilatih, menyimpan hasil prediksi tersebut dalam variabel y_pred.

6. *Code Cell* Keenam

```
print(r2_score(y, y_pred))
```

```
0.5177484222203499
```

Penjelasan:

Kode diatas digunakan untuk menghitung dan mencetak nilai R^2 *score* antara nilai aktual y dan nilai prediksi y_pred yang dihasilkan oleh model regresi linier. R^2 *score,* yang berkisar antara 0 hingga 1, menunjukkan seberapa baik model menjelaskan variabilitas dalam data; nilai yang lebih mendekati 1 menunjukkan bahwa model memiliki kemampuan yang baik dalam memprediksi nilai target, sedangkan nilai yang lebih mendekati 0 menunjukkan kinerja yang buruk.

Dari *output* yang dihasilkan didapatkan nilai R^2 *score* yaitu 0.5177484222203499, ini menunjukkan bahwa model regresi linier yang digunakan mampu menjelaskan sekitar 51.77% dari variabilitas dalam data target (y) berdasarkan fitur (x). Ini berarti bahwa lebih dari setengah variabilitas dalam progresi diabetes dapat diprediksi oleh model. Namun, nilai R^2 ini juga mengindikasikan bahwa hampir 48.23% dari variabilitas masih tidak terjelaskan, yang mungkin disebabkan oleh faktor lain yang tidak tercakup dalam model atau oleh adanya *noise* dalam data.

7. *Code Cell* Ketujuh

```
print(mean_squared_error(y, y_pred))
```

```
2859.69634758675
```

Penjelasan:

Kode diatas digunakan untuk menghitung dan mencetak nilai *Mean Squared Error* (MSE) antara nilai aktual y dan nilai prediksi y_pred dari model. Dari *output* yang dihasilkan didapatkan nilai 2859.69634758674, ini menunjukkan bahwa, secara rata-rata, kuadrat kesalahan antara nilai prediksi dan nilai aktual pada dataset diabetes adalah sekitar 2859.70. Nilai ini mengindikasikan bahwa model regresi linier mengalami kesulitan dalam memberikan prediksi yang akurat, dengan kesalahan yang cukup besar.

8. *Code Cell* Kedelapan

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)
```

Penjelasan:

Kode diatas digunakan untuk membagi dataset menjadi data pelatihan dan data pengujian, di mana 20% dari data digunakan untuk pengujian dan 80% untuk pelatihan. Parameter random_state=42 memastikan bahwa pembagian data bersifat acak tetapi konsisten setiap kali kode dijalankan, memungkinkan evaluasi yang dapat direproduksi terhadap kinerja model.

9. Code Cell Kesembilan

```
lr = LinearRegression()
lr.fit(x_train, y_train)
y_pred = lr.predict(x_test)
print(r2_score(y_test, y_pred))
print(mean_squared_error(y_test, y_pred))
```

```
0.4526027629719195
2900.193628493482
```

Penjelasan:

Kode diatas digunakan untuk melatih dan mengevaluasi model regresi linier. Pertama, objek model regresi linier lr dibuat, kemudian dilatih menggunakan data pelatihan x_train dan y_train. Setelah model dilatih, ia digunakan untuk memprediksi nilai target berdasarkan data pengujian x_test, dan hasil prediksi disimpan dalam variabel y_pred. Selanjutnya, kode menghitung dan mencetak nilai R^2 *score* dan *Mean Squared Error (*MSE) antara nilai aktual y_test dan prediksi y_pred, yang menunjukkan seberapa baik model menjelaskan variabilitas data pengujian. Dari *output* yang

dihasilkan didapatkan nilai R^2 *score* yaitu 0.4526027629719195 dan nilai MSE yaitu 2900.193628493482.

Nilai R^2 *score* sebesar 0.4526027629719195 menunjukkan bahwa model regresi linier mampu menjelaskan sekitar 45.26% dari variabilitas dalam data pengujian. Ini berarti bahwa hampir setengah dari variasi dalam nilai target masih tidak terjelaskan oleh model, yang menunjukkan bahwa ada ruang untuk perbaikan.

Sementara itu, nilai *Mean Squared Error* (MSE) sebesar 2900.193628493482 menunjukkan bahwa rata-rata kuadrat kesalahan antara nilai prediksi dan nilai aktual pada dataset pengujian adalah sekitar 2900.19. MSE yang relatif tinggi ini mengindikasikan bahwa model mungkin tidak cukup akurat dalam memprediksi nilai target, dan ada kemungkinan faktor-faktor lain yang perlu dipertimbangkan atau penggunaan model yang lebih kompleks untuk meningkatkan akurasi prediksi.