Naufal Alif Anargya-2311110041-SD0401

```
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
```

Kode ini menggunakan dataset diabetes dari `sklearn` untuk memprediksi kondisi pasien berdasarkan data kesehatan mereka. Dalam prosesnya, tiga model regresi diterapkan: **Linear Regression** sebagai model dasar, **Lasso Regression** yang menambahkan regularisasi L1 untuk menghilangkan fitur kurang relevan, dan **Ridge Regression** yang menggunakan regularisasi L2 untuk mengecilkan pengaruh fitur yang kurang penting. Dataset dibagi menjadi bagian pelatihan dan pengujian, sehingga performa model dapat dievaluasi menggunakan **Mean Squared Error (MSE)** untuk mengukur rata-rata kesalahan prediksi (semakin kecil, semakin baik) dan **R^2 Score** untuk melihat seberapa baik model menjelaskan variasi data (nilai mendekati 1 menunjukkan performa yang lebih baik). Melalui pendekatan ini, kode bertujuan membandingkan ketiga model regresi dalam memprediksi kondisi pasien diabetes.

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

Kode `x, y = load_diabetes(return_X_y=True)` berfungsi untuk memuat dataset diabetes dari sklearn dan langsung memisahkan data input (fitur) dan target.

- x = nilai fitur
- y = nilai target

```
lr = LinearRegression()
```

Kode `lr = LinearRegression()` membuat sebuah instance dari model Linear Regression dari library sklearn.

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

```
10
```

Code tersebut berguna untuk menghitung jumlah fitur dalam dataset yang dimana setelah kita lihat pada output menunjukkan bahwa dataset memiliki 10 fitur atau variabel yang digunakan sebagai input untuk model prediksi.

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

Kode tersebut digunakan untuk melatih model regresi linier dan memprediksi nilai target. `lr.fit(x, y)` melatih model menggunakan fitur `x` dan target `y`, sementara `y_pred = lr.predict(x)` menghasilkan prediksi berdasarkan fitur yang sama. Dengan cara ini, model belajar dari data untuk memahami hubungan antara variabel independen dan dependen, sehingga bisa digunakan untuk membuat prediksi di masa mendatang.

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

```
0.5177484222203498
```

R^2 Score: 0.5177, yang berarti model menjelaskan sekitar 51,77% dari variasi data target. Ini menunjukkan hubungan moderat, tetapi tidak terlalu kuat.

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

```
2859.6963475867506
```

Mean Squared Error (MSE): 2859.6963, yang mengindikasikan rata-rata kuadrat kesalahan prediksi. MSE ini relatif tinggi, menunjukkan bahwa terdapat deviasi yang cukup besar antara nilai prediksi dan aktual.

```
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)
```

Kode tersebut digunakan untuk membagi dataset menjadi data pelatihan dan pengujian. `train_test_split(x, y, test_size=0.2, random_state=42)` membagi 80% data untuk pelatihan (`x_train`, `y_train`) dan 20% untuk pengujian (`x_test`, `y_test`). Dengan cara ini, model dapat dilatih pada satu subset dan diuji pada subset lainnya untuk mengevaluasi kinerjanya. Parameter `random_state` memastikan bahwa pembagian data dapat direproduksi.

```
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
```

- R^2 Score: 0.4526, yang berarti model hanya menjelaskan sekitar 45,26% dari variasi data pengujian. Nilai R^2 yang lebih rendah dibandingkan pada blok pertama dapat mengindikasikan bahwa model mungkin mengalami penurunan performa ketika diterapkan pada data yang belum dilihat sebelumnya.
- Mean Squared Error (MSE): 2900.1936, sedikit lebih tinggi dibandingkan dengan MSE pada blok pertama. Ini mengindikasikan bahwa prediksi model pada data pengujian cenderung kurang akurat.

Hasil perbandingan ini menunjukkan bahwa model Linear Regression memiliki performa yang lebih buruk ketika dievaluasi pada data pengujian (dibandingkan dengan seluruh dataset). Nilai R2 yang lebih rendah pada data pengujian (0.4526) menunjukkan bahwa model mungkin tidak menangkap semua pola dalam data. Sementara itu, MSE yang sedikit lebih tinggi (2900.1936) menunjukkan bahwa ada deviasi prediksi dari nilai aktual. Secara keseluruhan, model linear regression ini menunjukkan kinerja yang sedang dalam memprediksi data diabetes, dengan variasi yang cukup besar antara hasil prediksi dan nilai sebenarnya.