CSGE603130 • Kecerdasan Artifisial dan Sains Data Dasar
Semester Ganjil 2021/2022
Fakultas Ilmu Komputer, Universitas Indonesia

Tugas 3 : Regresi
Tenggat Waktu: 18 November 2021, 23.55 WIB

Ketentuan:

1. Dataset yang digunakan pada tugas ini beserta deskripsinya telah disediakan di SCeLe.
2. Buatlah program Jupyter Notebook yang menjawab pertanyaan sesuai dengan perintah soal yang disediakan.
3. Program Jupyter Notebook yang telah dibuat dikumpulkan dengan format penamaan **Kelas_TugasX_NPM_Nama.ipynb**
   Contoh: F_Tugas3_1706979341_Lulu Ilmaknun Qurotaini.ipynb
4. Kumpulkan dokumen tersebut pada submisi yang telah disediakan di SCeLe sesuai dengan kelas masing-masing sebelum **18 November 2021, 23.55 WIB**. Keterlambatan pengumpulan akan dikenakan pinalti.
5. Tugas ini dirancang sebagai **tugas mandiri**. Plagiarisme tidak diperkenankan dalam bentuk apapun. Adapun kolaborasi berupa diskusi (tanpa menyalin maupun mengambil jawaban orang lain) dan literasi masih diperbolehkan dengan mencantumkan kolaborator dan sumber.
6. Template Tugas 3 dapat diakses pada link berikut: https://colab.research.google.com/drive/1sDW_aITZo0hMvVP-qu-fdUGFMRQqxb-S?usp=sharing

Deskripsi dataset cancer_reg:

Dataset berisi data survey sensus di Amerika. Berikut merupakan deskripsi dari setiap atribut pada dataset:

TARGET_deathRate: Dependent variable. Mean *per capita* (100,000) cancer mortalities(*a*)

TARGET_logistic: Dependent variable. Klasifikasi dari kolom **TARGET_deathRate**, dimana :

1 => **TARGET_deathRate <=180**

2 **=> TARGET_deathRate > 180**

TARGET_softmax: Dependent variable. Klasifikasi dari kolom **TARGET_deathRate**, dimana :

1 => **TARGET_deathRate < 160**

2 => **160 <= TARGET_deathRate < 190**

3 => **TARGET_deathRate >= 190**

- **avgAnnCount:** Mean number of reported cases of cancer diagnosed annually(*a*)
- **avgAnnCount**: Mean number of reported cases of cancer diagnosed annually(a)
- **avgDeathsPerYear**: Mean number of reported mortalities due to cancer(a)
- **incidenceRate**: Mean per capita (100,000) cancer diagoses(a)
- **medianIncome**: Median income per county (b)
- **popEst2015**: Population of county (b)
- **povertyPercent**: Percent of populace in poverty (b)
- **studyPerCap**: Per capita number of cancer-related clinical trials per county (a)
- **binnedInc**: Median income per capita binned by decile (b)
- **MedianAge**: Median age of county residents (b)
- **MedianAgeMale**: Median age of male county residents (b)
- **MedianAgeFemale**: Median age of female county residents (b)
- **Geography**: County name (b)
- **AvgHouseholdSize**: Mean household size of county (b)
- **PercentMarried**: Percent of county residents who are married (b)
- **PctNoHS18_24**: Percent of county residents ages 18-24 highest education attained: less than high school (b)
- **PctHS18_24**: Percent of county residents ages 18-24 highest education attained: high school diploma (b)
- **PctSomeCol18_24**: Percent of county residents ages 18-24 highest education attained: some college (b)
- **PctBachDeg18_24**: Percent of county residents ages 18-24 highest education attained: bachelor's degree (b)

- **PctHS25_Over**: Percent of county residents ages 25 and over highest education attained: high school diploma (b)
- **PctBachDeg25_Over**: Percent of county residents ages 25 and over highest education attained: bachelor's degree (b)
- **PctEmployed16_Over**: Percent of county residents ages 16 and over employed (b)
- **PctUnemployed16_Over**: Percent of county residents ages 16 and over unemployed (b)
- **PctPrivateCoverage**: Percent of county residents with private health coverage (b)
- **PctPrivateCoverageAlone**: Percent of county residents with private health coverage alone (no public assistance) (b)
- **PctEmpPrivCoverage**: Percent of county residents with employee-provided private health coverage (b)
- **PctPublicCoverage**: Percent of county residents with government-provided health coverage (b)
- **PctPubliceCoverageAlone**: Percent of county residents with government-provided health coverage alone (b)
- **PctWhite**: Percent of county residents who identify as White (b)
- **PctBlack**: Percent of county residents who identify as Black (b)
- **PctAsian**: Percent of county residents who identify as Asian (b)
- **PctOtherRace**: Percent of county residents who identify in a category which is not White, Black, or Asian (b)
- **PctMarriedHouseholds**: Percent of married households (b)
- **BirthRate**: Number of live births relative to number of women in county (b)

Sumber dataset:
https://data.world/nrippner/ols-regression-challenge (dimodifikasi)

Soal Tugas 3

Diberikan sebuah dataset cancer_reg, tujuan akhir dari pemrosesan data nantinya adalah memprediksi kolom TARGET_deathRate. Untuk mempersiapkan data tersebut, kerjakan soal-soal berikut!

[25] Preprocessing

1. [10] Berikan ringkasan mengenai data tersebut terkait dengan deskripsi setiap atribut, jumlah atribut (numerik & kategorik), jumlah *missing values*, jumlah duplikasi data, dan kemungkinan adanya *outliers* pada data!
2. [5] Berdasarkan eksplorasi anda pada nomor 1, lakukan data preparation hingga data tersebut menurut anda cukup "clean" dan dapat memberikan hasil regresi yang maksimal.
3. [10] Menurut Anda, apakah perlu dilakukan normalisasi terhadap data sebelum pemrosesan lebih lanjut, atau cukup menggunakan data asli? Jika ya, bentuk normalisasi apa yang tepat digunakan pada data? Jelaskan secara singkat alasan Anda!

[45] Regresi

(Penggunaan *library* diperbolehkan). Gunakan **TARGET_deathRate** sebagai kolom target.

1. [15] Implementasikan Linear Regression pada data hasil *preprocessing.* Gunakan Method dari library sklearn.
    a. Tampilkan visualisasi hasil prediksi
    b. Tampilkan nilai MSE, MAE, RMSE, dan R2 Square
2. [15] Implementasikan Ridge Regression pada data hasil *preprocessing.* Gunakan method dari library sklearn.
    a. Tampilkan nilai R2 Square
    b. Coba ubah parameter alpha dengan nilai yang lebih besar dan analisis bagaimana hubungan perubahan parameter tersebut dengan kualitas hasil regresi.
3. [15] Implementasikan Lasso Regression pada data hasil *preprocessing.* Gunakan method dari library sklearn.
    a. Tampilkan nilai R2 Square
    b. Coba ubah parameter alpha dengan nilai yang lebih besar dan analisis bagaimana hubungan perubahan parameter tersebut dengan kualitas hasil regresi.

    NOTE: Parameter alpha adalah parameter yang merepresentasikan "kekuatan" dari regularisasi yang dilakukan oleh model. Regularisasi adalah teknik modifikasi yang digunakan untuk mengurangi *generalization error,* dengan kata lain menghindari *overfitting*. Hal ini dilakukan dengan memberikan batasan/*constraint* atau penambahan penalti pada parameter/atribut yang kita gunakan. Untuk lebih jelas, silahkan baca dokumentasi model Ridge dan Lasso yang ada pada link berikut : API Reference — scikit-learn 1.0.1 documentation

[30] Logistic and Softmax Regression

1. [15] Implementasikan Logistic Regression dengan target yang digunakan adalah kolom **TARGET_logistic**. Gunakan method dari library sklearn.
    a. Visualisasikan hasil regresi yang didapat dengan membandingkan hasil prediksi dan nilai target aktual.
    b. Tampilkan rata-rata akurasi yang didapat (baca method .score() pada dokumentasi

sklearn)

2. [15] Implementasikan Softmax Regression dengan target yang digunakan adalah kolom **TARGET_softmax**. Gunakan method dari library sklearn.
    a. Visualisasikan hasil regresi yang didapat dengan membandingkan hasil prediksi dan nilai target aktual.
    b. Tampilkan rata-rata akurasi yang didapat.