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

Tugas 1: *Data Preparation & Dimensionality Reduction*
Tenggat Waktu: 2 Oktober 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_Tugas1_1706979341_Lulu Ilmaknun Qurotaini.ipynb
4. Kumpulkan dokumen tersebut pada submisi yang telah disediakan di SCeLe sesuai dengan kelas masing-masing sebelum **2 Oktober 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 untuk pengerjaan Tugas 1 dapat diakses pada link berikut:
   https://colab.research.google.com/drive/10_oYeE_GIUaoQenq2bP3b-eILfMoaSmh?usp=sharing

Deskripsi dataset *startup_data*:

Dataset berisi data *startup* yang berada di Amerika. Di dalamnya terdapat beberapa fitur atau spesifikasi dari suatu startup beserta dengan statusnya. Berikut merupakan deskripsi dari setiap atribut pada dataset:

- `state_code`: kode state startup
- `latitude`: posisi latitude startup
- `longitude`: posisi longitude startup
- `founded_at`: tanggal ketika startup tersebut didirikan
- `age_first_funding_year`: umur startup dalam tahun ketika pertama kali mendapatkan funding
- `age_last_funding_year`: umur startup dalam tahun ketika terakhir kali mendapatkan funding
- `funding_rounds`: banyaknya funding yang diterima oleh startup
- `funding_total_usd`: jumlah funding yang diterima oleh startup dalam USD
- `category_code`: bidang yang menjadi fokus dari startup
- `has_VC`: apakah startup tersebut memiliki venture capital
- `has_angel`: apakah startup tersebut memiliki angel investor
- `has_seriesA`: apakah startup tersebut mendapatkan funding series A
- `has_seriesB`: apakah startup tersebut mendapatkan funding series B
- `has_seriesC`: apakah startup tersebut mendapatkan funding series C
- `has_seriesD`: apakah startup tersebut mendapatkan funding series D
- `avg_participants`: rata-rata banyak pengguna dari startup tersebut dalam juta
- `is_top500`: apakah startup tersebut pernah masuk ke dalam 500 startup dengan peringkat teratas di Amerika
- `status` (target): status dari startup tersebut sekarang, *acquired* berarti startup tersebut berhasil karena diakuisisi oleh organisasi lain, sebaliknya, *closed* berarti startup tersebut sudah berhenti beroperasi dan gagal

Sumber dataset: https://www.kaggle.com/manishkc06/startup-success-prediction
(dimodifikasi oleh Asisten Dosen KASDD Ganjil 2021/2022)

Deskripsi dataset *cancer_reg*:

Dataset yang diperoleh sensus penduduk Amerika Serikat untuk memprediksi tingkat kematian warga-warga Amerika Serikat akibat kanker. Berikut merupakan deskripsi dari setiap atribut pada dataset:

- `avgAnnCount`: jumlah rata-rata kasus kanker yang dilaporkan yang didiagnosis setiap tahun
- `avgDeathsPerYear`: rata-rata jumlah kematian yang dilaporkan karena kanker
- `incidenceRate`: rata-rata per kapita (100.000) diagnosis kanker
- `medIncome`: pendapatan rata-rata per kabupaten
- `popEst2015`: populasi kabupaten
- `povertyPercent`: persentase kemiskinan penduduk
- `studyPerCap`: jumlah uji klinis terkait kanker per kapita per kabupaten
- `MedianAge`: usia rata-rata penduduk kabupaten
- `MedianAgeMale`: usia rata-rata penduduk kabupaten laki-laki
- `MedianAgeFemale`: usia rata-rata penduduk kabupaten perempuan
- `AvgHouseholdSize`: rata-rata ukuran rumah tangga kabupaten
- `PercentMarried`: persentase penduduk kabupaten yang menikah
- `BirthRate`: jumlah kelahiran relatif terhadap jumlah wanita di kabupaten
- `TARGET_deathRate` (target): rata-rata per kapita (100.000) kematian akibat kanker

Sumber dataset: https://data.world/nrippner/ols-regression-challenge
(dimodifikasi oleh Asisten Dosen KASDD Ganjil 2021/2022)

Soal Tugas 1

[50] Preprocessing

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

1. [15] 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. [10] Lakukan eksplorasi sederhana pada data dan ceritakan *insight* yang Anda dapatkan dari data tersebut! (Hint: Anda bisa mencari hubungan antar atribut atau melakukan visualisasi sederhana dari atribut tertentu)
3. [15] Lakukan penanganan terhadap *missing values*, duplikasi data, dan *outliers* jika ada!
4. [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!

[50] Dimensionality reduction

Diberikan sebuah dataset *cancer_reg*, lalu lakukanlah dimensionality reduction dengan mengikuti instruksi berikut!

5. [10] Visualisasikan dataset tersebut dengan menggunakan t-SNE! (2 komponen)
6. [20] Implementasikan *step-by-step* PCA secara manual pada data hasil *preprocessing*!
   Pilih jumlah komponen utama yang menurut Anda sebaiknya digunakan sehingga dapat menggambarkan data dengan baik. Sertakan juga alasan yang mendasari Anda melakukan pemilihan tersebut! (Hint: Anda bisa menggunakan rasio kumulatif dari nilai eigen ke-i sebagai persentase *variance* yang dapat di-*cover* oleh i nilai eigen tertinggi)
   Catatan: penggunaan *library* yang diperbolehkan pada implementasi hanya library *numpy* dan *pandas*
7. [10] Implementasikan PCA menggunakan *library scikit-learn* dengan:
   a. Jumlah komponen utama sebanyak 2. Visualisasikan hasil transformasi dengan menggunakan *scatter plot*!
   b. Jumlah komponen utama yang sama pada nomor 6. Tampilkan hasil transformasi beserta dengan nilai eigen dari implementasi tersebut!
8. [10] Berikan analisis Anda secara singkat mengenai perbedaan:
   a. Hasil yang Anda dapatkan pada nomor 5 dan nomor 7a!
   b. Hasil yang Anda dapatkan pada nomor 6 dan nomor 7b!