Tugas Akhir Metode Numerik 2021/2022

Diberikan dataset `diamonds.csv`. Data yang tersedia pada `diamonds.csv` adalah data harga berlian dan berbagai informasi mengenai sebuah berlian. Keterangan untuk masing-masing atribut adalah sebagai berikut:

1. price: harga berlian dalam dollar US
2. carat: berat berlian
3. cut: kualitas dari potongan berlian (Fair, Good, Very Good, Premium,Ideal)
4. color: warna berlian (dari J(paling buruk) hingga D(paling baik))
5. clarity: ukuran seberapa bening berlian (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
6. x: panjang dalam mm
7. y: lebar dalam mm
8. z: kedalaman dalam mm
9. depth: persentase total kedalaman (z / average(x,y))
10. table: panjang terluas dari permukaan atas sebuah berlian

Dari dataset ini, carilah model regresi terbaik untuk menentukan harga sebuah berlian. Untuk mencari model regresi ini lakukanlah:

1. Eksplorasi data sederhana. Anda bisa membuat grafik-grafik yang dapat membantu Anda dalam menentukan hubungan antar atribut yang ada pada dataset `diamonds.csv`
2. Jika Anda tidak dapat menemukan atribut numerik yang kira-kira memiliki hubungan yang kuat dengan harga berlian, cobalah untuk memanfaatkan atribut yang tipenya kategorik (seperti cut, color, dan clarity). Anda dapat memisahkan dataset menjadi beberapa dataset berlian yang lebih kecil berdasarkan atribut cut, kemudian membuat grafik untuk membantu Anda mempelajari hubungan antar atribut numerik dengan harga berlian.

   Note: untuk mengambil sebagian isi dataframe berdasarkan elemen bari, Anda dapat menggunakan fungsi `<nama data frame>.loc[(<condition>), [<nama-nama kolom>]]`. Salah satu source untuk belajar fungsi `loc`: https://www.activestate.com/resources/quick-reads/how-to-slice-a-dataframe-in-pandas/ . Source untuk belajar Python tidak terbatas pada yang diberikan di modul, silakan pelajari lagi dari berbagai source yang tersedia online.

3. Tentukanlah atribut-atribut numerik yang akan digunakan untuk membuat model regresi. Jika memutuskan menggunakan lebih dari 1 atribut numerik, maka pastikan kedua atribut numerik tersebut tidak saling bergantung. Anda bisa melihat kebergantungan ini dari scatter plot kedua variabel tersebut. Jika dari scatter plot terlihat kedua variabel tidak memiliki hubungan, maka Anda bisa menggunakan kedua variabel tersebut untuk membuat multiple regression. Tapi jika kedua variabel memiliki hubungan, Anda cukup memilih salah satunya.
4. Jika Anda sudah menentukan calon atribut yang akan digunakkan untuk melakukan regresi, carilah model regresi terbaik dengan memanfaatkan S_r dan metode cross validation yang sudah Anda pelajari sebelumnya. Model regresi Anda bisa saja lebih dari satu (jika Anda membuat model regresi untuk data yang terpisah berdasarkan atribut kategorikalnya, misalnya ada model

regresi untuk data berlian yang kualitas potongan berliannya Fair, ada model regresi untuk yang kualitas potongannya Good, dll).

5. Gunakan model yang Anda dapatkan di nomor 3 untuk membuat fungsi yang akan mengembalikan nilai prediksi harga berlian dengan parameter dari fungsi tersebut adalah atribut-atribut yang digunakan untuk regresi.

Kerjakan pada Jupiter Notebook. Selama proses untuk menemukan model yang tepat, selalu berikan narasi berisi penjelasan proses yang sedang Anda lakukan dan apa hasilnya. Anda dapat menggunakan fungsi untuk mencari model regresi, menghitung hasil regresi, dan menghitung nilai S_r yang sudah pernah Anda buat sebelumnya.

Kerjakan dalam kelompok yang berisi **maksimal** 3 orang. Satu kelompok cukup mengumpulkan satu file notebook saja. Pada notebook Anda, tuliskan Nama dan NPM masing-masing anggota kelompok.

Penilaian dilakukan berdasarkan:

1. Eksplorasi yang dilakukan sebelum menentukan atribut yang akan digunakan untuk membuat model
2. Proses dalam membuat model regresi
3. Fungsi prediksi berdasarkan model regresi yang dipilih
4. Narasi untuk setiap bagiannya

Tugas akhir ini memiliki bobot 60% dari UAS.