PENGGALIAN DATA
TUGAS KELOMPOK II
CLUSTERING

A. Permasalahan

Tugas kelompok kedua berkaitan dengan analisis data melalui proses clustering menggunakan R. Data yang diberikan relatif sangat kecil dan ditujukan agar mahasiswa dapat melakukan analisis dengan baik dan komprehensip. Sangat diharapkan bahwa tugas kelompok ini dikerjakan sendiri-sendiri dalam kelompoknya dan TIDAK berusaha untuk menyontek hasil perkerjaan dari kelompok lain, sehingga tugas ini dapat memberikan manfaat yang semaksimal mungkin. Adanya pertanyaan teknis penggunaan R sehubungan dengan implementasi tugas ini dapat dikonsultasikan dengan Asisten tutorial.

B. Deskripsi data

Data diperoleh dari bechmark internasional dan terkait dengan *abalone*, yaitu nama umum untuk kelompok siput laut yang dapat dimakan. Nama umum lainnya dari abalone adalah kerang kuping atau kuping laut.

1. **Title of Database:** Abalone data
2. **Sources**:
   (a) Original owners of database: Marine Resources Division; Marine Research Laboratories – Taroona, Department of Primary Industry and Fisheries, Tasmania; GPO Box 619F, Hobart, Tasmania 7001, Australia.
   (b) Donor of database: Sam Waugh (Sam.Waugh@cs.utas.edu.au); Department of Computer Science, University of Tasmania GPO Box 252C, Hobart, Tasmania 7001, Australia.
   (c) Date received: December 1995
3. **Relevant Information Paragraph:**
   The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.

   From the original data examples with missing values were removed (the majority having the predicted value missing), and the ranges of the continuous values have been scaled for use with an ANN (by dividing by 200).
4. **Number of Instances:** 4177
5. **Number of Attributes:** 8
6. **Attribute information:** Given is the attribute name, attribute type, the measurement unit and a brief description.

Name	Data Type	Measurement	Description
Sex	nominal	---	M, F, and I (infant)
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Whole weight	continuous	grams	whole abalone
Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut weight (after bleeding)
Shell weight	continuous	grams	after being dried
Rings	integer	---	+1.5 gives the age in years

7. **Missing Attribute Values:** None

C. Tugas

1. Pelajari beberapa referensi pendukung untuk memahami makna dari data dan juga untuk keperluan praproses dan analisis.
2. Lakukan eksplorasi data dari berbagai perspektif untuk memahami karakteristik data bisnis ritel tersebut. Gambakan hasil eksplortasi dalam berbagai bentuk grafik/chart yang menurut anda paling sesuai untuk menggambarkan karateristik data.
3. Lakukan praproses data, jika diperlukan (dapat menggunakan praproses yang disediakan dalam library R atau menggunakan praproses manual/menggunakan *spreadsheet*). Anda juga dapat memodifikasi bentuk data, sehingga sesuai dengan tujuan analisis yang akan dilakukan berdasarkan pemahaman terhadap referensi yang digunakan pada poin nomor 1 (gunakan referensi yang kredibel dan dapat diakses secara online).
4. Lakukan proses penghapusan penilan (*outlier*). Pertama identifikasi calon pencilan menggunakan cara mendefinisikan *outliers* dalam algoritma DBScan. Kemudian pastikan apakah calon pencilan benar-benar dapat dikategorikan sebagai pencilan definitif dengan menggunakan algoritma LOF.
5. Lakukan proses clustering (dari data hasil penghapusan pencilan) menggunakan empat pendekatan yang berbeda: **partisional (K-means)**, **Vector Quantization (VQ)**, **metode hirarki (MIN, MAX, dan AVERAGE)**, dan **metode berbasis densitas (DBScan).** Gunakan library R yang sesuai untuk masing-masing metode. Untuk metode K-means, lakukan eksperimen dengan berbagai titik pusat awal yang berbeda. Untuk metode VQ, lakukan eksperimen untuk beberapa nilai *rho*. Untuk metode DBScan, lakukan eskperimen untuk beberapa nilai *minimum points* dan *epsilon* yang berbeda.
6. Lakukan perbandingan hasil clustering dari ketiga metode yang digunakan. Berikan penjelasan secukupnya untuk mendukung hasil perbandingan.
7. Untuk hasil analisis, tentukan jumlah klaster yang paling tepat (misalnya menggunakan metode silhoute) dan juga analsis mendalam terhadap masing-masing klaster yang dihasilkan.

D. Laporan dan Batas Waktu

Laporan ditulis pada kertas berukuran A4 dengan spasi tunggal. Laporan dalam format PDF diserahkan per kelompok dan diunggah dalam menu “Assignment” pada aplikasi TEAMS **paling lambat pada tanggal 5 Mei 2020 pukul 16.00 WIB** (hanya satu orang dari setiap kelompok yang mengunggah). Masukkan *screenshot* dari script R yang digunakan disertai penjelasan seperlunya. Penilaian akan didasarkan pada aspek: sistematika penulisan dan kelengkapan laporan (25%), eksplorasi dan praproses data (20%), dan hasil clustering, ketajaman serta kedalaman analisis, termasuk rujukan terhadap referensi analisis CRM yang digunakan (55%). Tugas ini akan memberikan kontribusi 35% dari keseluruhan nilai tugas mata kuliah. Isi laporan yang mengindikasikan adanya plagiarisme tidak akan dinilai.

----------oooOooo----------