PENGGALIAN DATA
TUGAS KELOMPOK III
ASSOCIATION RULES MINING

A. Permasalahan

Tugas ketiga berkaitan dengan analisis data melalui penggalian aturan asosiasi (*association rules mining*) menggunakan R. Data yang digunakan diambil dari *the census bureau database* di USA dan diperoleh dari *UCI Machine Laerning Repository*. Data tersebut dapat digunakan untuk menganalisis asosiasi antar beberapa atribut non-kelas dengan tingkat penghasilan yang diperoleh.

B. Deskripsi data

Data yang digunakan terdiri 14 attribut non-kelas (atribut ke-1 s.d. ke-14) dan satu atribut label kelas (atribut ke-15). Jenis atribut bertipe campuran (kontinyu dan diskrit). Dalam data juga terdapat atribut yang nilainya tidak diketahui (*missing attribute*) yang ditandai dengan karakter tanda tanya (?). Terdapat dua nilai dari label kelas (<= 50K dan > 50K), yaitu apakah seseorang mempunyai penghasilan lebih kecil atau sama dengan USD 50.000 per bulan atau lebih besar dari USD 50.000. Deskripsi dari data dapat dilihat dalam tabel berikut.

No	Nama Atribut	Nilai Atribut
1	age	continuous
2	workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
3	fnlwgt	continuous
4	education	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
5	education-num	continuous
6	marital-status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
7	occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
8	relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
9	race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
10	sex	Female, Male
11	capital-gain	continuous
12	capital-loss	continuous
13	hours-per-week	continuous
14	cative-country	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands
15	**class-label**	**<= 50K, > 50K**

Jumlah baris data (*instances*) adalah 48.842 baris (termasuk baris yang di dalamnya terdapat nilai atribut yang tidak diketahui) atau 45.222 baris (mengabaikan baris yang mengandung nilai tidak diketahui).

C. Tugas

1. Lakukan eksplorasi data dari berbagai perspektif untuk memahami karakteristik data. Gambakan hasil eksplortasi dalam berbagai bentuk grafik/chart yang menurut anda paling sesuai untuk menggambarkan karateristik data.
2. Lakukan praproses data (dapat menggunakan praproses yang disediakan dalam library R atau menggunakan praproses manual/menggunakan *spreadsheet*). Praproses difokuskan untuk mentrasformasikan data agar dapat diperlakukan sebagai item, sehingga dapat dilakukan analisis asosiasi.
3. Lakukan proses pembangkitan *frequent itemsets* dengan menggunakan algoritma *apriori* dan *FP-growth*. Lakukan perbandingan yang diperoleh menggunakan kedua algoritma tersebut. Gunakan library R yang tersedia untuk keduanya. Lakukan uji coba untuk berbagai nilai ambang batas support dan tentukan nilai ambang batas support yang pas menurut hasil uji coba anda.
4. Bangkitkan sejumlah aturan asosiasi (*association rules*) yang menarik dari satu set *frequent itemsets* yang diperoleh sebelumnya, di mana atribut "class-label" sebagai target bersama-sama dengan misalnya rata-rata usia sebagai bagian dari target. Lakukan analisis kemenarikan (*interestingness*) dari aturan yang dihasilkan menggunakan berbagai ukuran kemenarikan (selain *confidence*) yang dapat dilakukan menggunalan R. Lakukan uji coba untuk berbagai nilai ambang batas ukuran kemenarikan dan buat kesimpulan dari hasil uji coba tersebut.

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 hari Jumat, tanggal 13 Desember 2019 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 (25%), dan hasil clustering, ketajaman dan kedalaman analisis, termasuk rujukan terhadap referensi analisis CRM yang digunakan (50%). Tugas ini akan memberikan kontribusi 25% dari keseluruhan nilai tugas mata kuliah. Isi laporan yang mengindikasikan adanya plagiarisme tidak akan dinilai.

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