IS184943 - PENGGALIAN DATA
TUGAS KELOMPOK I
KLASIFIKASI

A. Permasalahan

Tugas kelompok pertama berkaitan dengan implementasi beberapa model pengklasifikasi menggunakan R. Data yang digunakan berkaitan dengan pemeringkatan film dan diperoleh dari Internet.

B. Deskripsi data

SUMMARY

These files contain 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users who joined MovieLens in 2000.

RATINGS FILE DESCRIPTION

All ratings are contained in the file "ratings.dat" and are in the following format:

UserID::MovieID::Rating::Timestamp

- UserIDs range between 1 and 6040
- MovieIDs range between 1 and 3952
- Ratings are made on a 5-star scale (whole-star ratings only)
- Timestamp is represented in seconds since the epoch as returned by time(2)
- Each user has at least 20 ratings

USERS FILE DESCRIPTION

User information is in the file "users.dat" and is in the following format:

UserID::Gender::Age::Occupation::Zip-code

All demographic information is provided voluntarily by the users and is not checked for accuracy. Only users who have provided some demographic. Information are included in this data set.

- Gender is denoted by a "M" for male and "F" for female

- Age is chosen from the following ranges:
* 1: "Under 18"
* 18: "18-24"
* 25: "25-34"
* 35: "35-44"
* 45: "45-49"
* 50: "50-55"
* 56: "56+"

- Occupation is chosen from the following choices:
* 0: "other" or not specified
* 1: "academic/educator"
* 2: "artist"
* 3: "clerical/admin"
* 4: "college/grad student"
* 5: "customer service"
* 6: "doctor/health care"
* 7: "executive/managerial"

* 8: "farmer"
* 9: "homemaker"
* 10: "K-12 student"
* 11: "lawyer"
* 12: "programmer"
* 13: "retired"
* 14: "sales/marketing"
* 15: "scientist"
* 16: "self-employed"
* 17: "technician/engineer"
* 18: "tradesman/craftsman"
* 19: "unemployed"
* 20: "writer"

MOVIES FILE DESCRIPTION

Movie information is in the file "movies.dat" and is in the following format:

MovieID::Title::Genres

- Titles are identical to titles provided by the IMDB (including year of release)

- Genres are pipe-separated and are selected from the following genres:
* Action
* Adventure
* Animation
* Children's
* Comedy
* Crime
* Documentary
* Drama
* Fantasy
* Film-Noir
* Horror
* Musical
* Mystery
* Romance
* Sci-Fi
* Thriller
* War
* Western

- Some MovieIDs do not correspond to a movie due to accidental duplicate entries and/or test entries
- Movies are mostly entered by hand, so errors and inconsistencies may exist

C. Tugas

1. Lakukan eksplorasi data dari berbagai perspektif untuk memahami karakteristik data agar anda akan mempunyai persepsi yang baik terhadap data yang akan diolah. Gambakan hasil eksplortasi dalam berbagai bentuk grafik/*chart* yang menurut anda paling sesuai untuk menggambarkan karateristik data secara komprehensip dan mudah dipahami.
2. Lakukan praproses data yang menurut anda diperlukan sebelum dilakukan proses penggalian data.

3. Lakukan proses pembuatan model pengklasifikasi menggunakan metode: (a) *Decision Tree***,** (b) *Nearest Neighbor* (NN), (c) *Bayesian Classifier*, dan (d) *Neural Network.* Untuk ini, gunakan *library* yang tersedia dalam R (dapat diperoleh dari berbagai sumber dan jelaskan library yang anda digunakan). Untuk masing model pengklasifikasi, lakukan evaluasi kinerja dari model pengklasifikasi dengan menggunakan (a) *repeated hold-out* sebanyak 100 kali putaran dan (b) *10-fold cross validation*.
4. Terkait dengan tugas nomor 3 di atas:
   a. Untuk *repeated hold-out*, pilih 2/3 data secara acak sebagai data pelatihan dan 1/3 sisanya sebagai data tes/validasi pada setiap putarannya.
   b. Pada setiap putaran (untuk setiap model pengklasifikasi), baik yang menggunakan *repeated hold-out* maupun yang menggunakan *10-fold cross validation*.
5. Lakukan proses pengujian model pengklasifikasi menggunakan data tes yang tersedia. Hitung kinerja hasil pengujian menggunakan *accuracy*, *precision*, *recall*, dan *F-measure*. Lakukan perbandingan kinerja dari keempat model pengklasifikasi tersebut. Selain itu, lakukan analisis yang komprehensip terhadap hasil klasifikasi (dapat disajikan menggunakan grafik/*chart* untuk mendukung analisis yang dibuat).

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 29 Oktober 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, kejelasan uraian laporan termasuk tata tulis, dan hasil akurasi dari model yang dibangun, serta kedalaman analisis hasil klasifikasi. Tugas ini akan memberikan kontribusi 40% dari keseluruhan nilai tugas mata kuliah. Isi laporan yang mengindikasikan adanya plagiarisme tidak akan dinilai.

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