

TUGAS INDIVIDU

Note: Tugas dikumpulkan dalam bentuk .pdf dan diupload dalam bentuk link GDrive yang diberi akses public. Isian tugas adalah screenshoot dari Perancangan Modelnya di RapidMiner dan Hasil Performancenya. Contoh screenshot dapat dilihat pada [halaman lampiran](#).


Soal 1 (ada di pertemuan 4 hal 39):



Latihan

- Karena bantuan data mining sebelumnya, Sarah akhirnya mendapatkan **promosi menjadi VP marketing**, yang mengelola ratusan marketer
- Sarah ingin para marketer dapat memprediksi pelanggan potensial mereka masing-masing secara mandiri. Masalahnya, data **HeatingOil.csv** hanya boleh diakses oleh level **VP (Sarah)**, dan tidak diperbolehkan diakses oleh marketer secara langsung
- Sarah ingin masing-masing marketer membuat proses yang dapat mengestimasi kebutuhan konsumsi minyak dari *client* yang mereka *approach*, dengan menggunakan model yang sebelumnya dihasilkan oleh Sarah, meskipun **tanpa mengakses data training (HeatingOil.csv)**
- Asumsikan bahwa data **HeatingOil-Marketing.csv** adalah data calon pelanggan yang berhasil di *approach* oleh salah satu marketingnya
- Yang harus dilakukan **Sarah** adalah membuat proses untuk:
 1. Mengkomparasi algoritma yang menghasilkan model yang memiliki akurasi tertinggi (LR, NN, SVM), gunakan 10 Fold X Validation
 2. Menyimpan model ke dalam suatu file (operator **Write Model**)
- Yang harus dilakukan **Marketer** adalah membuat proses untuk:
 1. Membaca model yang dihasilkan Sarah (operator **Read Model**)
 2. Menerapkannya di data **HeatingOil-Marketing.csv** yang mereka miliki
- Mari kita bantu Sarah dan Marketer membuat dua proses tersebut

Soal 2 (ada di pertemuan 5, hal 42):



Latihan: Prediksi Kelulusan Mahasiswa

1. Lakukan **training** pada data mahasiswa (**datakelulusanmahasiswa.xls**) dengan menggunakan DT, NB, K-NN
2. Lakukan dimension reduction dengan **Forward Selection** untuk ketiga algoritma di atas
3. Lakukan **pengujian** dengan menggunakan 10-fold X Validation
4. Uji beda dengan **t-Test** untuk mendapatkan model terbaik

	DT	NB	K-NN	DT+FS	NB+FS	K-NN+FS
Accuracy						
AUC						

Soal 3 (ada di pertemuan 5, hal 43):



Latihan

- Lakukan **training** pada data eReader Adoption (eReader-Training.csv) dengan menggunakan DT dengan 3 alternative **criterion** (Gain Ratio, Information Gain dan Gini Index)
- Lakukan feature selection dengan **Forward Selection** untuk ketiga algoritma di atas
- Lakukan pengujian dengan menggunakan 10-fold X Validation
- Dari model terbaik, tentukan **faktor (atribut) apa saja yang berpengaruh** pada tingkat adopsi eReader

	DTGR	DTIG	DTGI	DTGR+FS	DTIG+FS	DTGI+FS
Accuracy	58.39	51.01	31.01	61.41	56.73	31.01

Soal 4 (ada di pertemuan 5, hal 56):

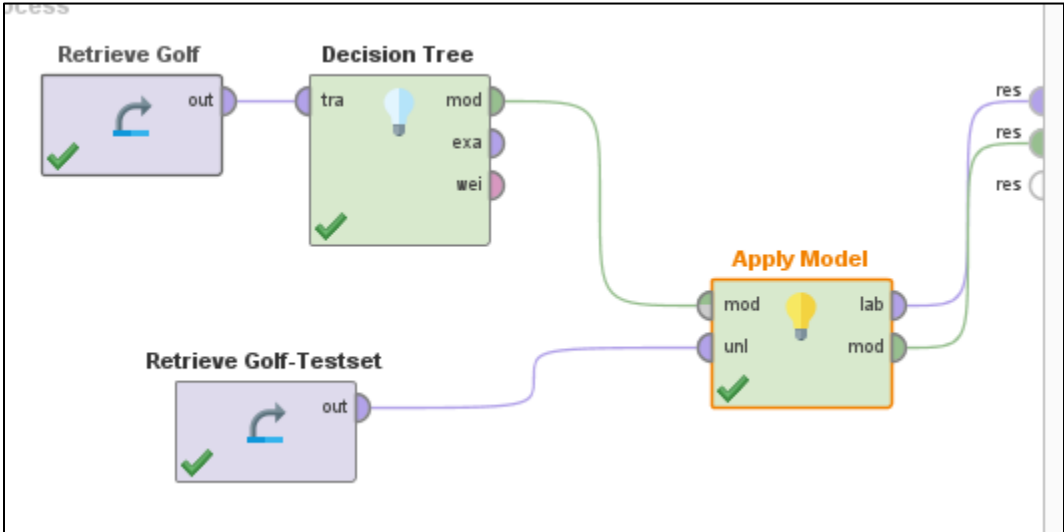


Latihan

- Lakukan eksperimen mengikuti buku Matthew North, Data Mining for the Masses, 2012, **Chapter 7 Discriminant Analysis**, pp. 105-125
- Datasets: **SportSkill-Training.csv** dan **SportSkill-Scoring.csv**
- Analisis **metode preprocessing** apa saja yang digunakan dan mengapa perlu dilakukan pada dataset tersebut!

Halaman Lampiran

Model:



Hasil:

