**Soal Ujian**

**Soal Pertama**

Budi adalah manajer perusahaan penyedia jasa asuransi kesehatan. Berdasarkan artikel dan makalah ilmiah yang dibacanya, Budi menemukan bahwa ada korelasi antara berat badan, jenis kelamin dan kolesterol pada pertumbuhan penyakit jantung koroner. Budi mengusulkan kepada perusahaan tempat dia bekerja untuk membuatkan program pengelolaan berat badan dan kolesterol bagi pelanggan asuransi kesehatan. Tindakan awal dari program tersebut adalah mengelompokkan pelanggan asuransi kesehatan yang memiliki resiko rendah, sedang dan tinggi terkena penyakit jantung koroner. Dengan pengelompokan pelanggan menjadi tiga tingkat resiko, perusahaan asuransi dimana Budi bekerja dapat lebih akurat dalam penentuan tarif asuransi, persiapan budget untuk klaim pelanggan asuransi, dan berbagai program lainnya. Tindakan berikutnya adalah Budi ingin mengetahui, korelasi antar atribut (faktor) dan juga khususnya korelasi berbagai atribut itu ke penyakit jantung koroner.

1. Untuk keperluan pengelompokan tersebut Budi mendapatkan 547 data yang dipilih secara random dari database pelanggan (*Koroner.csv*). Data pelanggan asuransi yang didapatkan Budi memiliki masalah karena beberapa ada yang kosong (hilang) dan juga ada yang *noisy* atau nilainya tidak wajar.
2. Analisis masalah di atas dengan menggunakan tahapan CRISP-DM dan mari kita bantu Budi menyelesaikan masalahnya dimulai dari pemahaman masalah, persiapan data, pemodelan, evaluasi dan *deployment*.
3. Perlu dicatat bahwa data yang kosong harus diganti dengan nilai rata-rata, sedangkan data *noisy* diperbolehkan untuk dihapus.
4. Gunakan algoritma *k-means*, dan lakukan analisis terhadap hasil yang didapatkan.
5. Tentukan nilai *k* paling optimal dengan melakukan ujicoba pengubahan nilai *k*. Lakukan pengukuran performance dengan menggunakan *Cluster Distance Performance*, untuk mendapatkan nilai *Davies Bouldin Index* (DBI). Nilai DBI semakin rendah berarti cluster yang kita bentuk semakin baik
6. Gambarkan grafik (*chart*) dan pilih *Scatter* untuk menggambarkan data hasil klastering yang telah dilakukan.
7. Simpan data hasil cluster yang memuat cluster sebagai atribut baru dengan operator *Write CSV* (*Koroner-Cluster.csv*). Gunakan operator *Replace* untuk mengubah nama cluster menjadi numerik (0, 1, 2, dsb), dan kemudian gunakan Correlation Matrix untuk melihat hubungan antar atribut. Gunakan operator *Rename* utk mengubah attribut cluster menjadi Jantung Koroner
8. Rangkumkan dalam bentuk slide hasil dari tahapan CRISP-DM yang dilakukan sehingga mudah dipahami dan presentasikan di depan kelas. Ambil *printscreen* dari Rapidminer yang diperlukan untuk mempermudah penjelasan.

**Soal Kedua**

Setelah sukses dengan program di atas, Budi melanjutkan program berikutnya yaitu program pengelolaan kesehatan sehingga dapat menghindarkan pelanggan dari kemungkinan terkena serangan jantung. Untuk bisa sukses melakukan program ini, Budi harus menemukan pola, pelanggan asuransi dengan profile seperti apa yang kemungkinan terkena serangan jantung.

1. Untuk keperluan di atas, Budi mengambil data pelanggan asuransi sebanyak 138 orang (*SeranganJantung.csv*). Data pelanggan asuransi yang didapatkan Budi memiliki masalah karena beberapa ada yang kosong (hilang) dan juga ada yang *noisy* atau nilainya tidak wajar.
2. Analisis masalah di atas dengan menggunakan tahapan CRISP-DM dan mari kita bantu Budi menyelesaikan masalahnya dimulai dari pemahaman masalah, persiapan data, pemodelan, evaluasi dan *deployment*.
3. Perlu dicatat bahwa karena terbatasnya jumlah data, maka data yang kosong tidak diperbolehkan untuk dihapus, tapi harus diganti dengan nilai rata-rata dari data yang lain. Data noisy boleh dihapus.
4. Pada fase pemodelan lakukan berbagai hal sebagai berikut:
   1. Analisis korelasi antar atribut dari dataset dengan *Correlation Matrix* dan pengamatan terhadap *chart* khususnya menggunakan *Scatter*
   2. Lakukan komparasi minimal 4 algoritma data mining, gunakan 10-fold X Validation dan pilih satu yang terbaik.
5. Simpan model yang terbaik dengan menggunakan operator *Write Model*. Gunakan model yang dihasilkan untuk diterapkan di 10 data pelanggan baru yang masuk ke Budi (*SeranganJantung-PelangganBaru.cs*v), analisis hasilnya.
6. Rangkumkan dalam bentuk slide hasil dari tahapan CRISP-DM yang dilakukan sehingga mudah dipahami dan presentasikan di depan kelas. Ambil *printscreen* dari Rapidminer yang diperlukan untuk mempermudah penjelasan.

Kirimkan semua file berhubungan dengan jawaban Soal Pertama dan Kedua, baik slide dan proses Rapidminer ke adiwjj@gmail.com

Keterangan dataset:

Koroner.csv

|  |  |  |
| --- | --- | --- |
| Weight | Pounds (lbs) | |
| Cholesterol | Tingkat Kolesterol | |
| Gender | 0 = perempuan | 1 = Laki-Laki |

SeranganJantung.csv

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Status Pernikahan | 0 = belum menikah | 1 = menikah | 2 = duda | 3 = janda |
| Jenis Kelamin | 0 = perempuan | 1 = laki-laki |  |  |
| Kategori Berat Badan | 0 = normal | 1 = kelebihan berat badan | 2 = obesitas |  |
| Pelatihan Pengelolaan Stress | 0 = tidak ikut | 1 = ikut |  |  |
| Tingkat Stres | 0-100 |  |  |  |
| Serangan Jantung | 0 = no 1 = yes | | | |