UJIAN TENGAH SEMESTER GANJIL 2019-2020
PROGRAM STUDI INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS AL AZHAR INDONESIA

MATA KULIAH	:	DATA MINING
DOSEN	:	ALI AKBAR SEPTIANDRI
KELAS	:	IF16A
HARI / TANGGAL	:	JUMAT, 08 NOVEMBER 2019
WAKTU	:	07.00-09.30
SIFAT UJIAN	:	BUKA BUKU, KALKULATOR

Peraturan

1. Bacalah doa terlebih dahulu dan bacalah soal dengan teliti
2. Dilarang pinjam-meminjam alat tulis kepada peserta ujian lain
3. Dilarang bertanya/memberitahukan kepada peserta ujian lain
4. Tidak diperkenankan keluar ruangan selama ujian berlangsung kecuali selesai mengerjakan soal ujian
5. Bacalah kembali soal dan jawaban sebelum meninggalkan ruangan ujian
6. Wajib mematuhi instruksi dari Dosen/Pengawas Ujian
7. Jika melakukan pelanggaran terhadap hal-hal tersebut diatas maka nilai ujian akan dikurangi atau nilai langsung *Failed* (E=0)
8. Notasi pemisah ribuan adalah koma (,), sedangkan desimal ditulis dengan titik (.)

1 Prediksi Tip

Sebuah perusahaan ojek online sedang mencoba memprediksi tips yang didapatkan pengemudi berdasarkan beberapa variabel dari setiap perjalanan. Ada enam pilihan nilai tips yang bisa diberikan pengguna: Rp1,000, Rp2,500, Rp5,000, Rp10,000, Rp15,000, dan Rp20,000. Beberapa variabel yang digunakan untuk memprediksi nilai tips adalah:

- Jarak perjalanan (km)
- Nilai pengemudi (1-5 bintang)
- Jenis kelamin penumpang (laki-laki/perempuan)
- Hari perjalanan (Senin-Minggu)

Berdasarkan informasi tersebut:

a. Tergolong kasus apakah contoh di atas? Klasifikasi, regresi, atau *clustering*? Berikan alasan Anda. [2 poin]

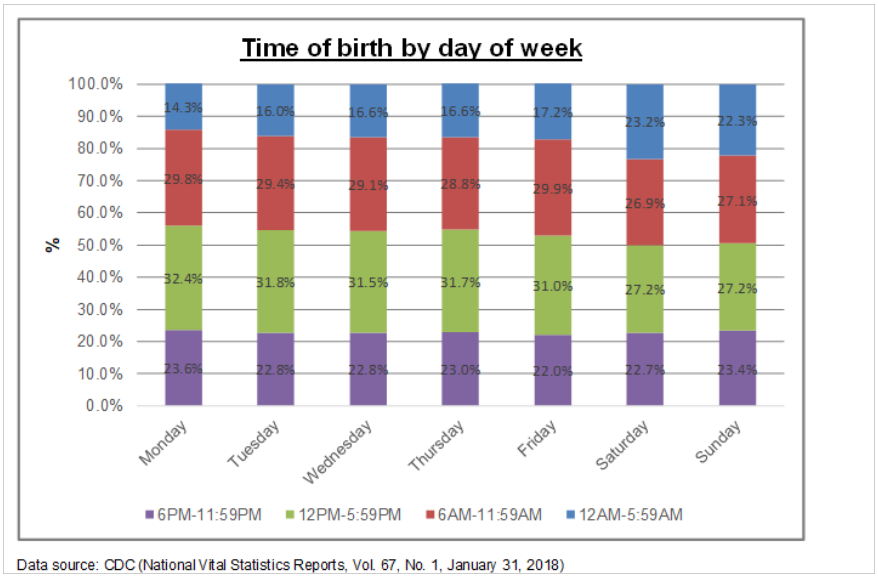
b. Apakah tipe dari atribut **hari perjalanan**? [1 poin]

c. Prapemrosesan apa yang biasanya dilakukan untuk dapat menggunakan model linear dengan tipe atribut seperti pada soal 1.b? Berikan contoh hasil prapemrosesan tersebut. [3 poin]

d. Variabel tambahan apa saja yang menurut Anda dapat membantu Anda memprediksi nilai tips dengan lebih baik di luar empat variabel yang telah disebutkan di atas? Berikan minimal 2 contoh variabel tambahan dan sertakan alasan dan prapemrosesan yang mungkin perlu dilakukan. [4 poin]

2 Visualisasi Data

Diberikan data waktu kelahiran bayi seperti pada Gambar 1.



Gambar 1: Data dari CDC (National Vital Statistics Reports, Vol. 67, No. 1, January 31, 2018)

a. Tuliskan 4 perbedaan tujuan visualisasi sebagai metode komunikasi (*explanatory*) dan analisis (*exploratory*). [4 poin]

b. Deskripsikan informasi menarik yang Anda temukan dari Gambar 1. [3 poin]

c. Jelaskan 3 hal yang dapat dilakukan untuk membuat visualisasi tersebut lebih efektif dalam memberikan informasi atau wawasan kepada pembaca. Anda dapat menghubungkan penjelasan Anda dengan jawaban dari soal 2.b. [3 poin]

3 Model Linear

a. Apa yang dimaksud sebagai *multicollinearity* dalam kasus regresi linear? [2 poin]

b. Apakah dampak dari *multicollinearity* terhadap model: *overfitting* atau *underfitting*? [1 poin]

c. Apa yang terjadi pada koefisien model Lasso saat diberikan penalti yang sangat besar? [2 poin]

d. Diberikan data pasangan tinggi dan berat badan sebagai berikut:

$$\text{tinggi} = [167, 150, 150, 165, 161, 186, 160, 167, 170, 163]$$

dan

$$\text{berat} = [67, 36, 54, 33, 83, 66, 45, 47, 59, 57]$$

Hitunglah korelasi antara kedua variabel tersebut. [5 poin]

Verifikasi dari Ketua Program Studi Informatika	Dosen Pengampu,
(Riri Safitri, S.Si., MT.)	(Ali Akbar Septiandri)