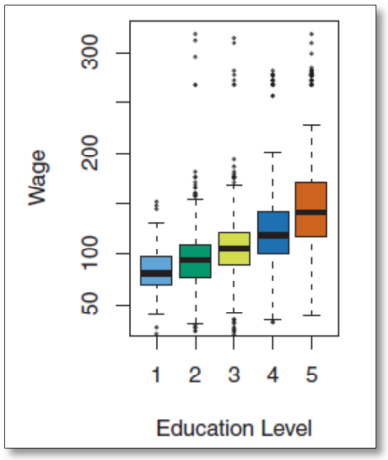
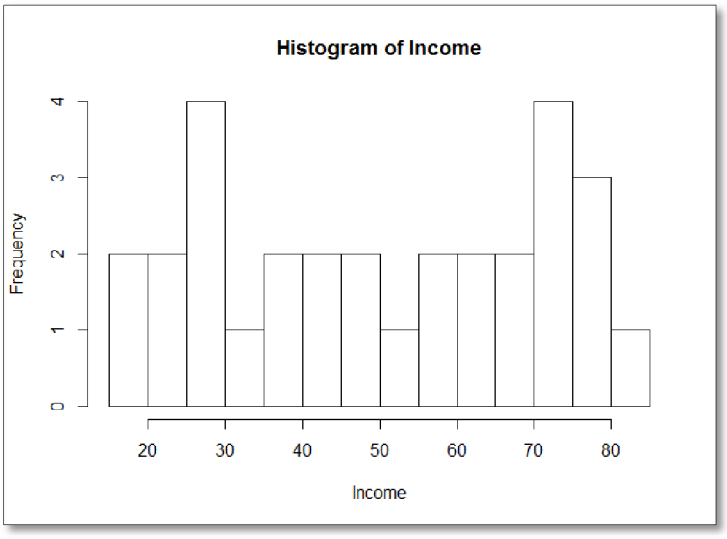
Kuis Besar UTS

IF074 - Pembelajaran Mesin

1. Berdasarkan gambar berikut ini. Bagaimanakah menurut Anda kaitan antara variabel *Wage* (upah) dengan *Education* (tingkat pendidikan)?



2. Apakah yang dimaksudkan dengan histogram?Jelaskan selengkap mungking hal-hal apa sajakah yang diilustrasikan dalam histogram di bawah ini?



3. Berikan dua contoh nyata untuk tipe-tipe data berikut ini:

   a. Numerik
   b. Nominal

4. Apakah perbedaan dan persamaan antara Pembelajaran Mesin dan Pembelajaran Statistik?

5. Berikan masing-masing dua contoh untuk *supervised* dan *unsupervised learning*.

6. Apakah keuntungan dan kerugian pendekatan fleksibel dalam klasifikasi dan regresi?

7. Apakah yang dimaksudkan dengan *Mean Square Error* (MSE), dan bagaimanakah konsep tersebut digunakan dalam evaluasi sebuah model? Ilustrasikan contoh anda.

8. Gunakan klasifikasi 3-NN (*nearest neighbor*) untuk data training dan testing berikut ini. Ke dalam kelas manakah testing diarahkan dengan 3-NN?

Tr	age	prescription	astigmatic	tear_rate	class
	2	1	1	3	no
	2	1	1	2	yes
	2	0	0	3	no
	1	0	1	3	no
	1	0	1	2	yes
Test:	2	0	0	2	?

9. Jelaskan mengenai masing-masing cara evaluasi model regresi linear berikut ini:

a. *Residual standard error*
b. *R^2 statistic*
c. *F statistic*

10. Diketahui sebuah hasil regresi linear yang merelasikan antara upah seseorang dan tingkat pendidikan, sebagai berikut:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -39.4463     4.7248  -8.349  4.4e-09 ***
Education     5.5995     0.2882  19.431  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.653 on 28 degrees of freedom
Multiple R-squared:  0.931,     Adjusted R-squared:  0.9285
F-statistic: 377.6 on 1 and 28 DF,  p-value: < 2.2e-16
```

Menurut anda, apakah nilai estimasi koefisien untuk atribut *Education* memiliki nilai kepercayaan yang baik? Mengapa?

11. Apakah yang dimaksudkan dengan *collinearity problem*? Berikan dua cara penanganannya dalam regresi linear.

12. Apakah keterbatasan regresi linear? Bagaimanakah regresi linear mempengaruhi pembentukan model dalam regresi logistik?

13. Diketahui sebuah kumpulan data sebagai berikut:

Person	Hair Length	Weight	Age	Class
Homer	0″	250	36	M
Marge	10″	150	34	F
Bart	2″	90	10	M
Lisa	6″	78	8	F
Maggie	4″	20	1	F
Abe	1″	170	70	M
Selma	8″	160	41	F
Otto	10″	180	38	M
Krusty	6″	200	45	M

Ubahlah nilai atribut numerik *'Hair Length'* menjadi:

- 0-5 = pendek
- 6-10 = panjang

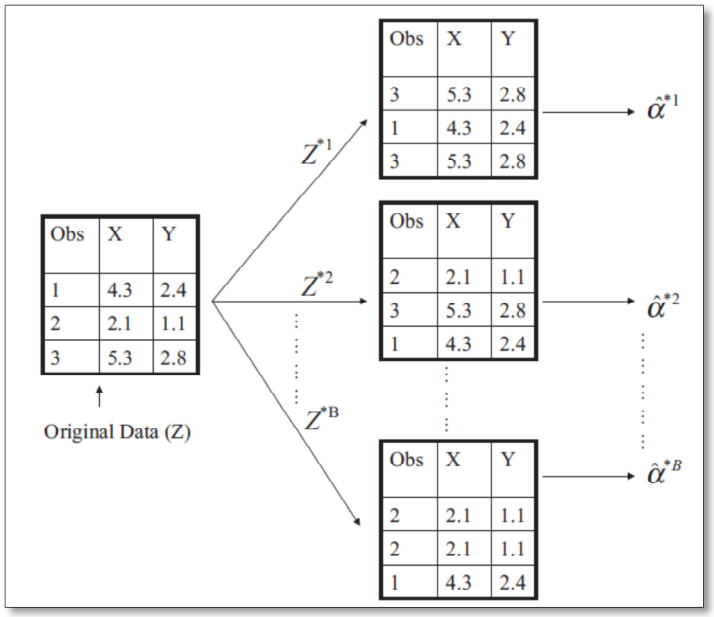
Berikan tabel yang berisi *probability, odds* dan *logit* untuk atribut *'Hair Length'* tersebut.

14. Apakah yang dimaksudkan dengan pembentukan validation set dengan konsep sebagai berikut (silakan menggunakan ilustrasi jika dianggap perlu):

   a. *Hold-out*
   b. *Leave one out cross validation (LOOCV)*
   c. *k-fold cross validation*

15. Apakah keuntungan dan kerugian menggunakan LOOCV dibandingkan *k*-fold CV?

16. Jelaskan penggunaan bootstrap pada contoh ilustrasi di bawah ini:



Yang dikumpulkan pada saat UTS adalah soal No. 1-16 secara tertulis.

No. 17-20 di bawah ini hanya untuk latihan, di saat UTS anda akan diminta untuk melakukannya.

17. Gunakan data **Heart.csv** (bisa diambil dari sitoba.itmaranatha.org)

a. Berikan *summary* dari data tersebut.
b. Atribut apa sajakah yang merupakan data nominal dan numerik?
c. Ada berapa banyak instans di dalam data tersebut?
d. Apakah data ini memiliki komposisi kelas AHD yang seimbang? Mengapa?

18. Lakukan regresi linear dari data Heart.csv untuk menebak kaitan antara tingkat kolesterol (Chol) dan atribut RestBP.

a. Berapakah nilai estimasi koefisien RestBP?
b. Berapakah tingkat kepercayaan estimasi tersebut?

19. Regresi linear dengan gabungan atribut:

a. Jika prediksi terhadap Chol dilakukan dengan menggunakan atribut MaxHL, apakah model menjadi lebih baik? Mengapa?
b. Jika prediksi terhadap Chol dilakukan dengan menggabungkan atribut RestBP dan MaxHL, apakah model menjadi lebih baik dibandingkan dengan hanya menggunakan salah satunya saja? Apakah RestBP dan MaxHL dapat disebutkan memiliki *collinearity problem*? Jelaskan.

20. Lakukan regresi logistik untuk data Heart

a. Gunakan semua varibel yang ada. Variabel manakah yang memberikan hasil prediksi koefisien paling baik secara nilai kepercayaan?
b. Bentuklah *confusion matrix* untuk hasil prediksi model terhadap data Heart, berapakah nilai akurasi dari model yang terbentuk tersebut?