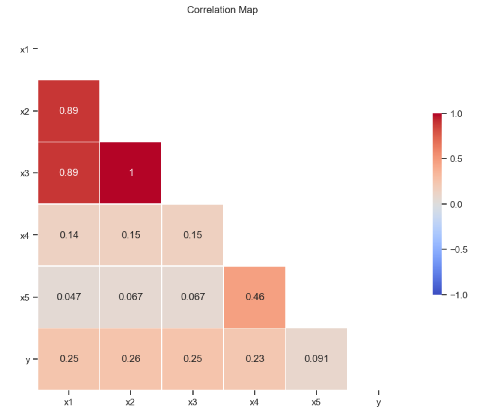
**SOAL BRI**

Perhatikan diagram di bawah untuk menjawab soal nomor 1:



1. Jika sebuah model linear akan dibangun menggunakan data yang ditampilkan pada diagram di atas, variabel apa yang paling tidak tepat untuk dibuang?

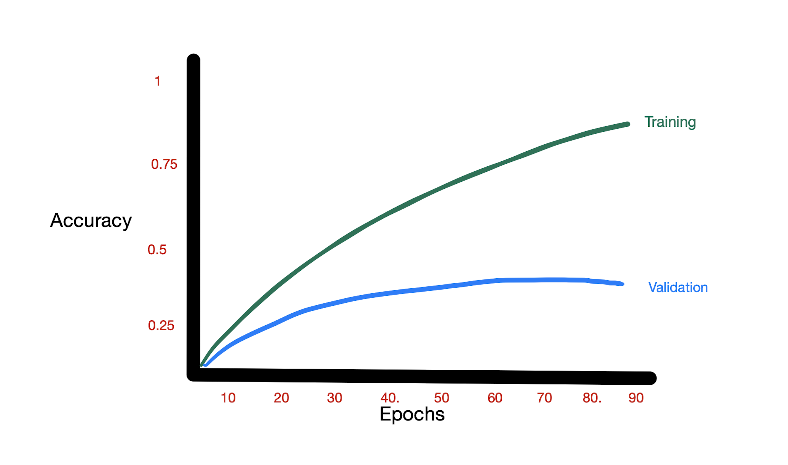
x1

x2

x3

x4

#### Perhatikan ilustrasi di bawah untuk menjawab soal nomor 2:



2. Selama training model menggunakan neural network, log performance yang didapatkan menyerupai ilustrasi di atas. Kesimpulan yang dapat diambil berdasarkan log performance plot tersebut adalah:

Model terindikasi underfitting

Model terindikasi overfitting

Model terjebak dalam local optima

Model terjebak dalam global optima

3. Dalam sebuah perusahaan media, terdapat sebuah divisi yang bertugas untuk mencari berita palsu. Anda sebagai data scientist diminta untuk membuat sebuah model machine learning yang dapat mendeteksi sebuah berita palsu (*positive class*) secara otomatis. Berita palsu yang terdeteksi kemudian akan ditangani lebih lanjut oleh divisi tersebut untuk memutuskan kebutuhan untuk menindak lanjuti pengunggah berita.  
  
Divisi tersebut berharap untuk mengoptimalkan model machine learning  yang Anda bangun agar dapat dengan tepat mengkategorikan berita palsu guna mengurangi dampak risiko yang tidak diinginkan dari proses tindak lanjut yang hendak dilakukan.  
  
Metrik yang tepat untuk digunakan dalam mengukur performa model pada kasus tersebut di atas:

Accuracy

Precision

Recall

Mean Squared Error

4. Proses normalisasi data adalah salah satu tahap pre-processing yang dibutuhkan untuk dapat meningkatkan performa model. Meskipun begitu, beberapa model machine learning menggunakan perhitungan yang memperhatikan unit variabel prediktor yang digunakan sehingga proses normalisasi menjadi hal yang harus dilakukan. Dari model-model di bawah model manakah yang membutuhkan proses normalisasi data?

Multiple Linear Regression

Random Forest

Binomial Logistic Regression

k-NN

5. Apabila proses normalisasi data dilakukan pada sebuah dataset, manakah pernyataan berikut yang tepat:

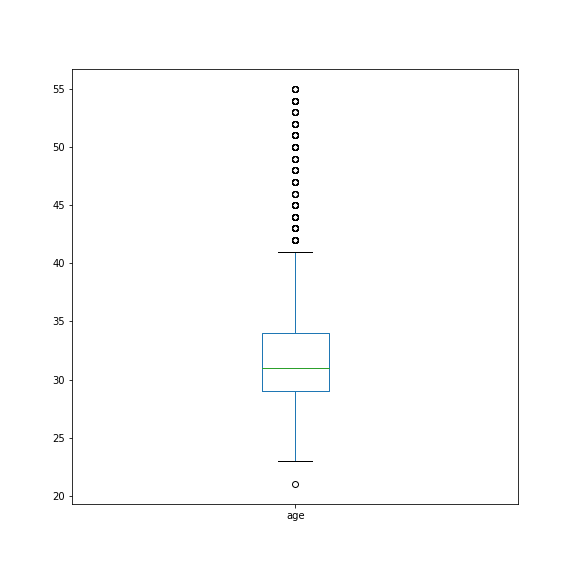
Nilai korelasi antarvariabel tidak akan berubah

Nilai kovarian antarvariabel tidak akan berubah

Nilai rata-rata tiap variabel tidak akan berubah

Nilai standar deviasi tiap variabel tidak akan berubah

#### Perhatikan plot di bawah untuk menjawab soal nomor 6:



6. Berdasarkan boxplot di atas, jika dalam suatu sampel yang diambil secara random satu observasi, manakah di antara usia observasi yang paling sesuai untuk dapat dikategorikan sebagai pencilan?

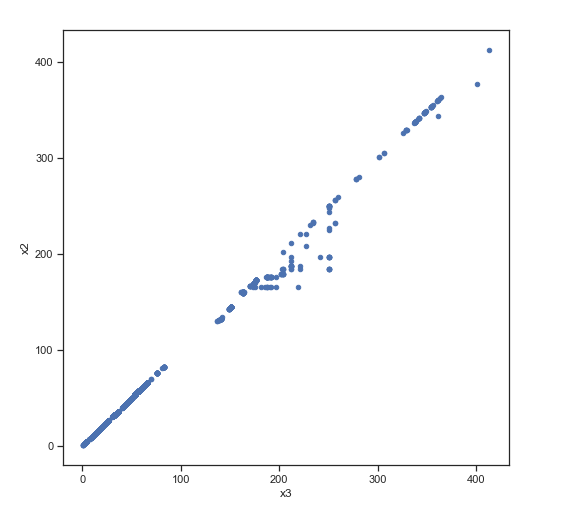
20

30

35

40

#### Perhatikan plot di bawah untuk menjawab soal nomor 7:



7. Bedasarkan scatter plot di atas, pernyataan yang paling sesuai mengenai dua variabel yang ditampilkan adalah:

Variabel x2 dan x3 memiliki hubungan non linear

Variabel x2 dan x3 memiliki hubungan linear yang lemah

Variabel x2 dan x3 berpotensi untuk menjadi variabel prediktor yang redundan

Tidak ada jawaban yang tepat

#### 8. Seorang nasabah di suatu perbankan melakukan transaksi penarikan saldo dalam 30 hari terakhir. Dalam 28 hari terakhir nasabah tersebut rutin melakukan penarikan tunai sebesar Rp100,000. Pada dua hari terakhir, nasabah tersebut melakukan penarikan sebesar Rp7,500,000 dan Rp5,000,000. Berdasarkan data transaksi tersebut, nilai manakah yang paling tepat untuk menggambarkan sentral tendensi dari sebaran data penarikan tunai nasabah tersebut?

Rp510,000

Rp100,000

Rp275,000

Rp426,000

#### 9. Lozy merupakan sebuah usaha yang bergerak di bidang fashion. Untuk memasarkan produknya, Lozy menggunakan sebuah digital platform dan merekrut influencer A sebagai seorang brand ambassador. Sebelum menjalankan campaign tersebut, rata-rata social media engagement yang diterima Lozy adalah sebesar 14,000 likes dengan standar deviasi sebesar 5,500. Setelah beberapa waktu, bedasarkan data yang yang dikumpulkan dari 20 posting social media, rata-rata likes yang diterima oleh Lozy meningkat menjadi 20,000. Menggunakan nilai taraf signifikansi 5%, apakah strategi kampanye Lozy dapat dikatakan meningkatkan social media engagement secara signifikan?

Strategi kampanye dengan influencer A tidak meningkatkan social media engagement Lozy

Strategi kampanye dengan influencer A meningkatkan social media engagement Lozy

Strategi kampanye dengan influencer A kurang signifikan dalam meningkatkan social media engagement Lozy

Strategi kampanye dengan influencer A tidak berpengaruh terhadap nilai social media engagement Lozy

#### 10. Berdasarkan sebuah studi terdahulu, diketahui bahwa dividend payout ratio dari sebuah industri menyerupai distribusi normal dengan standar deviasi 2.4%. Seorang manajer investasi mengumpulkan data dari 81 perusahaan di sebuah industri dan mendapatkan rata-rata dividend payout ratio dari 81 perusahaan tersebut adalah 11.8%. Apabila manajer investasi tersebut ingin membangun 95% interval keyakinan untuk rata-rata dividend payout ratio populasi di industri tersebut, berapakah estimasi yang didapatkan?

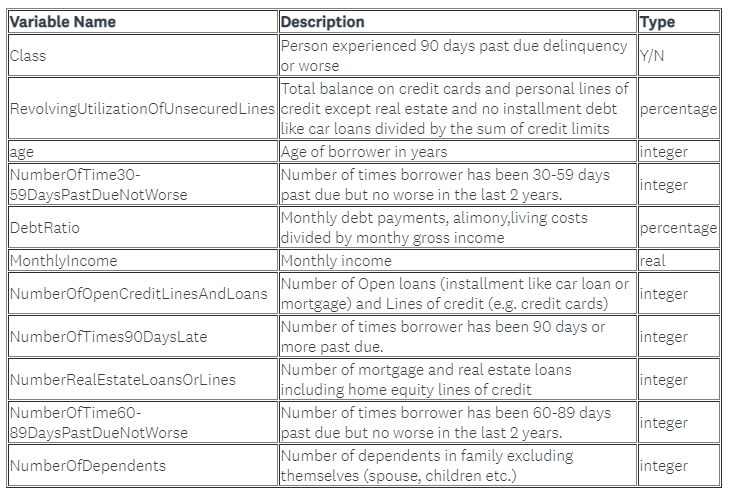
11.36% - 12.24%

11.28% - 12.32%

11.8% - 12.32%

11.36% - 12.8%

Untuk menjawab soal nomor 11 - 15 silahkan unduh dataset pada [tautan berikut](https://drive.google.com/file/d/1NXA7dGWeTs817GHkjA5v7IiOCHmj0N7S/view?usp=sharing" \t "_blank).  
  
Dataset ini merupakan sebuah dataset yang telah dimodifikasi dari sebuah [kompetisi terdahulu](https://www.kaggle.com/c/GiveMeSomeCredit/data?select=cs-training.csv" \t "_blank) pada platform Kaggle dan memilik data glossary sebagai berikut:



#### 11. Apabila model untuk memprediksi target Class dibangun dengan binomial logistic regression, dapat diketahui bahwa dengan model tidak mampu mencapai kondisi konvergen. Pernyataan yang paling sesuai untuk menjelaskan kondisi tidak konvergen adalah:

Iterasi yang dilakukan tidak mampu menghasilkan model dengan error yang paling optimal

Iterasi yang dilakukan tidak mampu bergerak menuju nilai estimasi parameter yang diharapkan

Iterasi yang dilakukan tidak mampu mendapatkan nilai information loss di bawah batas toleransi

Iterasi yang dilakukan tidak mengasilkan hasil model yang lebih baik dari tebakan acak

#### 12. Berdasarkan data tersebut, diketahui bahwa terdapat beberapa kolom yang memiliki nilai yang hilang (missing value). Beberapa strategi dapat dilakukan untuk menangani kondisi ini. Di bawah merupakan variabel yang membutuhkan penanganan missing value﻿:

NumberOfDependents

NumberOfTime60.89DaysPastDueNotWorse

DebtRatio

NumberOfTimes90DaysLate

#### 13. Dari variabel-variabel yang tersedia, dapat dilihat bahwa beberapa variabel terindikasi multicollinearity sehingga memiliki risiko redundansi informasi. Dari variabel-variabel berikut, manakah yang memiliki risiko terkecil terjadinya multicollinearity ketika digunakan?

NumberOfTimes90DaysLate

NumberOfTime60.89DaysPastDueNotWorse

NumberOfOpenCreditLinesAndLoans

NumberOfTime30.59DaysPastDueNotWorse

#### 14. Bangunlah sebuah model binomial logistic regression untuk memodelkan variabel Class dengan variabel prediktor berikut: age, DebtRatio, MonthlyIncome, NumberOfOpenCreditLinesAndLoans dan RevolvingUtilizationOfUnsecuredLines. Berdasarkan tes signifikansi dari masing-masing prediktor, variabel manakah yang disimpulkan sebagai variabel signifikan terhadap peluang nasabah mengalami keterlambatan pembayaran di atas 90 hari?

DebtRatio

age

RevolvingUtilizationOfUnsecuredLines

NumberOfOpenCreditLinesAndLoans

#### 15. Menggunakan model yang sama dengan pertanyaan sebelumnya, buatlah sebuah confusion matrix dengan menggunakan in-sample data. Hal yang dapat disimpulkan dari confusion matrix tersebut adalah?

Model membutuhkan model lain sebagai benchmark model

Model merupakan best-fit

Model terindikasi overfitting

Model terindikasi underfitting