

Modul 06 - Classification

Roni Yunis

10/23/2023

Pengantar

Analisis Prediktif bisa menggunakan dua metode yaitu;

1. Classification: suatu bentuk dasar dari analisis data dimana datanya diklasifikasi ke dalam kelas-kelas.
2. Regression: memprediksi nilai variabel numerik, misalnya angka pendapatan perusahaan atau angka penjualan.

Untuk mendukung kedua hal tersebut, maka diperlukan **Machine Learning (ML)**, ML adalah tentang bagaimana membuat komputer belajar dan melakukan tugas dengan lebih baik berdasarkan data historis masa lalu. Pembelajaran selalu dilakukan berdasarkan pada observasi dari data yang tersedia. Ada dua jenis ML.

1. Supervised Learning: yaitu mesin membuat model prediktif dengan bantuan sekumpulan data pelatihan (data training) atau sering disebut juga pembelajaran dengan pengawasan, karena variabel yang digunakan sudah ditentukan
2. Unsupervised Learning: yaitu model prediktif yang tidak menggunakan data pelatihan (data training), tidak ada variabel target, sehingga sering disebut dengan pembelajaran tanpa pengawasan, karena variabel yang digunakan tidak ditentukan.

Untuk menghasilkan model prediksi yang baik, sehingga dataset bisa dibagi menjadi 3 bagian, yaitu:

- a. **Training Data Partition:** Partisi data pelatihan (data training) digunakan untuk melatih model. Rincian variabel hasil sudah diketahui. Untuk masalah klasifikasi, kelas variabel hasil sudah ditentukan dan terkadang dibuat secara manual dengan campur tangan manusia.
- b. **Test Data Partitions:** Partisi data pengujian (data testing) adalah bagian dari kumpulan data yang tidak ada dalam kumpulan pelatihan. Ini digunakan untuk menilai kinerja model untuk data baru. Partisi ini terkadang disebut partisi testing. Model harus bekerja dengan baik untuk data set pelatihan dan data pengujian.
- c. **Validation Data Partition:** Partisi data validasi digunakan untuk menyempurnakan kinerja model dan mengurangi masalah overfitting. Partisi ini dapat digunakan untuk menilai beberapa model dan memilih model terbaik. Kumpulan data ini tidak digunakan untuk membangun model. Jadi, model tersebut belum pernah melihat kumpulan data ini sebelumnya. Ini membantu menyempurnakan kinerja model dan mengurangi overfitting.

Overfitting adalah suatu keadaan dimana data yang digunakan untuk pelatihan itu adalah yang terbaik. Sehingga apabila dilakukan tes dengan menggunakan data yang berbeda dapat mengurangi akurasi (wikipedia)

Pada pembahasan kali ini, kita akan membahas Klasifikasi, Klasifikasi adalah metode analisis data yang digunakan untuk menemukan pola dalam data. Klasifikasi memprediksi kelas kategorikal, sedangkan regresi memprediksi fungsi nilai kontinyu. Contoh penerapan model klasifikasi adalah untuk memprediksi hasil dari proses persetujuan aplikasi kartu kredit (disetujui atau ditolak) atau untuk menentukan klain asuransi. Ada banyak algoritma klasifikasi yang sudah dikembangkan dan dapat kita gunakan untuk menunjang prediksi yang akan dilakukan. Berikut algoritma klasifikasi yang bisa digunakan, diantaranya yaitu Naive Bayes, Decision Tree, K-Nearest Neighbor (KNN), Random Forest, dll. Dalam pembahasan pada Modul 05 ini, kita akan bahas bagaimana cara menggunakan Algoritma Decision Tree dan Random Forest untuk memprediksi dan mengklasifikasi dataset *credit.csv*

Load Packages

Untuk mendukung klasifikasi yang akan dilakukan, maka ada beberapa packages/library yang diperlukan.

```
# Package untuk manipulasi data
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
```

```
# Package untuk visualisasi data
library(ggplot2)
# package untuk praktisi data
library(caret)
```

```
## Loading required package: lattice
```

```
# package untuk klasifikasi
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##   margin
```

```
## The following object is masked from 'package:dplyr':
##
##      combine
```

```
# package untuk mengukur perfomansi model klasifikasi
library(e1071)
# package untuk menguji kehandalan dari model prediksi
library(ROCit)
# package untuk decision tree
library(rpart)
# package untuk memodelkan pohon keputusan
library(rpart.plot)
```

Data Preparation

Import data dan melihat struktur data

```
credit <- read.csv("data/credit.csv")
glimpse(credit)
```

```
## Rows: 1,000
## Columns: 21
## $ checking_balance      <chr> "< 0 DM", "1 - 200 DM", "unknown", "< 0 DM", "< 0~
## $ months_loan_duration <int> 6, 48, 12, 42, 24, 36, 24, 36, 12, 30, 12, 48, 12~
## $ credit_history        <chr> "critical", "repaid", "critical", "repaid", "dela~
## $ purpose              <chr> "radio/tv", "radio/tv", "education", "furniture",~
## $ amount               <int> 1169, 5951, 2096, 7882, 4870, 9055, 2835, 6948, 3~
## $ savings_balance      <chr> "unknown", "< 100 DM", "< 100 DM", "< 100 DM", "<~
## $ employment_length    <chr> "> 7 yrs", "1 - 4 yrs", "4 - 7 yrs", "4 - 7 yrs",~
## $ installment_rate     <int> 4, 2, 2, 2, 3, 2, 3, 2, 2, 4, 3, 3, 1, 4, 2, 4, 4~
## $ personal_status      <chr> "single male", "female", "single male", "single m~
## $ other_debtors        <chr> "none", "none", "none", "guarantor", "none", "non~
## $ residence_history     <int> 4, 2, 3, 4, 4, 4, 4, 2, 4, 2, 1, 4, 1, 4, 4, 2, 4~
## $ property            <chr> "real estate", "real estate", "real estate", "bui~
## $ age                 <int> 67, 22, 49, 45, 53, 35, 53, 35, 61, 28, 25, 24, 2~
## $ installment_plan     <chr> "none", "none", "none", "none", "none", "none", "~
## $ housing              <chr> "own", "own", "own", "for free", "for free", "for~
## $ existing_credits     <int> 2, 1, 1, 1, 2, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 1, 2~
## $ default             <int> 1, 2, 1, 1, 2, 1, 1, 1, 1, 2, 2, 2, 1, 2, 1, 2, 1~
## $ dependents          <int> 1, 1, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ telephone           <chr> "yes", "none", "none", "none", "none", "yes", "no~
## $ foreign_worker       <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", ~
## $ job                 <chr> "skilled employee", "skilled employee", "unskille~
```

Melihat ringkasan dari data

```
summary(credit)
```

```
##  checking_balance  months_loan_duration  credit_history      purpose
##  Length:1000      Min.      : 4.0          Length:1000      Length:1000
```

```
## Class :character 1st Qu.:12.0      Class :character  Class :character
## Mode :character Median :18.0      Mode :character  Mode :character
##                               Mean :20.9
##                               3rd Qu.:24.0
##                               Max. :72.0
## amount savings_balance employment_length installment_rate
## Min. : 250 Length:1000 Length:1000 Min. :1.000
## 1st Qu.: 1366 Class :character Class :character 1st Qu.:2.000
## Median : 2320 Mode :character Mode :character Median :3.000
## Mean : 3271 Mean :2.973
## 3rd Qu.: 3972 3rd Qu.:4.000
## Max. :18424 Max. :4.000
## personal_status other_debtors residence_history property
## Length:1000 Length:1000 Min. :1.000 Length:1000
## Class :character Class :character 1st Qu.:2.000 Class :character
## Mode :character Mode :character Median :3.000 Mode :character
##                               Mean :2.845
##                               3rd Qu.:4.000
##                               Max. :4.000
## age installment_plan housing existing_credits
## Min. :19.00 Length:1000 Length:1000 Min. :1.000
## 1st Qu.:27.00 Class :character Class :character 1st Qu.:1.000
## Median :33.00 Mode :character Mode :character Median :1.000
## Mean :35.55 Mean :1.407
## 3rd Qu.:42.00 3rd Qu.:2.000
## Max. :75.00 Max. :4.000
## default dependents telephone foreign_worker
## Min. :1.0 Min. :1.000 Length:1000 Length:1000
## 1st Qu.:1.0 1st Qu.:1.000 Class :character Class :character
## Median :1.0 Median :1.000 Mode :character Mode :character
## Mean :1.3 Mean :1.155
## 3rd Qu.:2.0 3rd Qu.:1.000
## Max. :2.0 Max. :2.000
## job
## Length:1000
## Class :character
## Mode :character
##
##
##
```

Melihat 6 baris teratas dari data credit

```
head(credit)
```

```
## checking_balance months_loan_duration credit_history purpose amount
## 1 < 0 DM 6 critical radio/tv 1169
## 2 1 - 200 DM 48 repaid radio/tv 5951
## 3 unknown 12 critical education 2096
## 4 < 0 DM 42 repaid furniture 7882
## 5 < 0 DM 24 delayed car (new) 4870
## 6 unknown 36 repaid education 9055
## savings_balance employment_length installment_rate personal_status
```

```

## 1      unknown      > 7 yrs      4      single male
## 2      < 100 DM      1 - 4 yrs      2      female
## 3      < 100 DM      4 - 7 yrs      2      single male
## 4      < 100 DM      4 - 7 yrs      2      single male
## 5      < 100 DM      1 - 4 yrs      3      single male
## 6      unknown      1 - 4 yrs      2      single male
##      other_debtors residence_history      property age installment_plan
## 1      none      4      real estate 67      none
## 2      none      2      real estate 22      none
## 3      none      3      real estate 49      none
## 4      guarantor      4 building society savings 45      none
## 5      none      4      unknown/none 53      none
## 6      none      4      unknown/none 35      none
##      housing existing_credits default dependents telephone foreign_worker
## 1      own      2      1      1      yes      yes
## 2      own      1      2      1      none      yes
## 3      own      1      1      2      none      yes
## 4 for free      1      1      2      none      yes
## 5 for free      2      2      2      none      yes
## 6 for free      1      1      2      yes      yes
##      job
## 1      skilled employee
## 2      skilled employee
## 3 unskilled resident
## 4      skilled employee
## 5      skilled employee
## 6 unskilled resident

```

Melihat 6 baris terakhir dari data credit

```
tail(credit)
```

```

##      checking_balance months_loan_duration credit_history      purpose amount
## 995      unknown      12      repaid car (new) 2390
## 996      unknown      12      repaid furniture 1736
## 997      < 0 DM      30      repaid car (used) 3857
## 998      unknown      12      repaid radio/tv 804
## 999      < 0 DM      45      repaid radio/tv 1845
## 1000      1 - 200 DM      45      critical car (used) 4576
##      savings_balance employment_length installment_rate personal_status
## 995      unknown      > 7 yrs      4      single male
## 996      < 100 DM      4 - 7 yrs      3      female
## 997      < 100 DM      1 - 4 yrs      4      divorced male
## 998      < 100 DM      > 7 yrs      4      single male
## 999      < 100 DM      1 - 4 yrs      4      single male
## 1000      101 - 500 DM      unemployed      3      single male
##      other_debtors residence_history      property age
## 995      none      3      other 50
## 996      none      4      real estate 31
## 997      none      4 building society savings 40
## 998      none      4      other 38
## 999      none      4      unknown/none 23
## 1000      none      4      other 27

```

```
##      installment_plan  housing existing_credits default dependents telephone
## 995          none      own              1      1      1      yes
## 996          none      own              1      1      1      none
## 997          none      own              1      1      1      yes
## 998          none      own              1      1      1      none
## 999          none for free            1      2      1      yes
## 1000         none      own              1      1      1      none
##      foreign_worker              job
## 995          yes      skilled employee
## 996          yes      unskilled resident
## 997          yes mangement self-employed
## 998          yes      skilled employee
## 999          yes      skilled employee
## 1000         yes      skilled employee
```

Kalau kita lihat ada beberapa variabel yang type datanya kategorikal, yaitu variabel `checking_balance`, `saving_balance`, `employment_length`, `personal_status`, `other_debtors`, `property`, `installment_plan`, `housing`, `telephone`, `foreign_worker`, `credit_history`, `purpose`, dan `job`

EDA dan Visualisasi

Melihat data kosong atau missing value (NA's)

```
colSums(is.na(credit))
```

```
##      checking_balance months_loan_duration      credit_history
##              0              0              0
##      purpose              amount      savings_balance
##              0              0              0
##      employment_length installment_rate      personal_status
##              0              0              0
##      other_debtors residence_history      property
##              0              0              0
##      age      installment_plan      housing
##              0              0              0
##      existing_credits      default      dependents
##              0              0              0
##      telephone      foreign_worker      job
##              0              0              0
```

Bisa dilihat bahwa tidak ada data kosong atau NA's

Kita lanjutkan melihat kategorikal dari beberapa variabel

```
# melihat kategori dari checking_balance
table(credit$checking_balance)
```

```
##
##      < 0 DM      > 200 DM 1 - 200 DM      unknown
##      274          63          269          394
```

Bisa dilihat ada 4 kategori, yaitu <0, 1-200, >200, dan unknown

```
# melihat kategori dari savings_balance
table(credit$savings_balance)
```

```
##
##      < 100 DM      > 1000 DM  101 - 500 DM  501 - 1000 DM      unknown
##           603             48           103           63           183
```

Bisa dilihat ada 5 kategori, <100, 101-500, 501-100, >1000, dan unknown

```
# melihat kategori dari housing
table(credit$housing)
```

```
##
## for free      own      rent
##       108       713      179
```

Ternyata housing terbanyak adalah untuk kategori own

```
# melihat kategori dari property
table(credit$property)
```

```
##
## building society savings      other      real estate
##              232              332              282
##      unknown/none
##              154
```

```
# Melihat kategori dari month_loan_duration dan purpose
table(credit$months_loan_duration, credit$purpose)
```

```
##
##      business car (new) car (used) domestic appliances education furniture
##  4           0           3           0                   0           0           1
##  5           1           0           0                   0           0           0
##  6           2          25           2                   2           5          11
##  7           0           0           0                   0           0           0
##  8           1           2           0                   0           0           1
##  9           2          11           1                   1           6          10
## 10           0          13           2                   1           1           6
## 11           1           5           0                   0           0           1
## 12           9          49          10                   2          10          35
## 13           1           0           0                   0           0           0
## 14           1           3           0                   0           0           0
## 15           3          13           7                   3           4          12
## 16           0           2           0                   0           0           0
## 18          12          23           6                   1           4          29
## 20           0           2           3                   0           0           2
## 21           5           9           4                   0           2           6
## 22           0           1           0                   0           0           0
```

##	24	18	38	29	0	4	36
##	26	0	0	1	0	0	0
##	27	6	1	2	0	0	1
##	28	0	1	1	0	0	0
##	30	6	4	6	0	0	9
##	33	1	0	1	0	0	1
##	36	9	16	14	1	8	14
##	39	0	0	2	0	1	1
##	40	0	0	0	0	1	0
##	42	2	0	2	0	0	2
##	45	1	0	1	0	0	0
##	47	0	1	0	0	0	0
##	48	13	7	8	1	3	3
##	54	1	0	1	0	0	0
##	60	2	5	0	0	1	0
##	72	0	0	0	0	0	0
##							
##		others	radio/tv	repairs	retraining		
##	4	0	2	0	0		
##	5	0	0	0	0		
##	6	0	24	2	2		
##	7	0	5	0	0		
##	8	1	2	0	0		
##	9	0	17	1	0		
##	10	0	4	0	1		
##	11	0	2	0	0		
##	12	0	55	4	5		
##	13	0	3	0	0		
##	14	0	0	0	0		
##	15	0	18	4	0		
##	16	0	0	0	0		
##	18	0	34	4	0		
##	20	1	0	0	0		
##	21	0	4	0	0		
##	22	0	1	0	0		
##	24	5	51	2	1		
##	26	0	0	0	0		
##	27	0	2	1	0		
##	28	0	1	0	0		
##	30	0	14	1	0		
##	33	0	0	0	0		
##	36	1	18	2	0		
##	39	0	1	0	0		
##	40	0	0	0	0		
##	42	0	4	1	0		
##	45	0	3	0	0		
##	47	0	0	0	0		
##	48	3	10	0	0		
##	54	0	0	0	0		
##	60	1	4	0	0		
##	72	0	1	0	0		

```
# melihat kategori purpose
table(credit$purpose)
```



```
##
##          business          car (new)          car (used) domestic appliances
##              97              234              103              12
##          education          furniture          others              radio/tv
##              50              181              12              280
##          repairs          retraining
##              22              9
```

Bisa dilihat ada 10 kategori. Kategori yang paling banyak adalah radio/tv

```
# melihat kategori dari foreign worker
table(credit$foreign_worker)
```

```
##
## no yes
## 37 963
```

Ternyata kategori untuk pekerja asing yang paling banyak yaitu sebanyak 963

```
# melihat kategori credit_history
table(credit$credit_history)
```

```
##
##          critical          delayed          fully repaid
##              293              88              40
## fully repaid this bank          repaid
##              49              530
```

Kategori credit history yang paling banyak adalah repaid.

```
# melihat asosiasi antara purpose dan credit history
table(credit$purpose, credit$credit_history)
```

```
##
##          critical delayed fully repaid fully repaid this bank
## business          19    23          15              7
## car (new)          78    17           7             12
## car (used)         36     8           3              5
## domestic appliances  1     0           0              1
## education          19     5           0              3
## furniture          50    10           7              8
## others              3     2           1              2
## radio/tv           80    20           4              9
## repairs            6     3           2              0
## retraining         1     0           1              2
##
##          repaid
## business          33
## car (new)         120
## car (used)         51
## domestic appliances 10
## education          23
```

```
## furniture          106
## others              4
## radio/tv           167
## repairs            11
## retraining          5
```

Ternyata bisa dilihat bahwa repaid dan untuk tujuan radio/tv adalah yang paling banyak yaitu sebanyak 167

kita akan filter, credit purpose = "radio/tv"

```
radiotv <- filter(credit, purpose == "radio/tv")
head(radiotv)
```

```
## checking_balance months_loan_duration credit_history purpose amount
## 1 < 0 DM 6 critical radio/tv 1169
## 2 1 - 200 DM 48 repaid radio/tv 5951
## 3 unknown 12 repaid radio/tv 3059
## 4 1 - 200 DM 12 repaid radio/tv 1567
## 5 < 0 DM 24 repaid radio/tv 1282
## 6 unknown 24 critical radio/tv 2424
## savings_balance employment_length installment_rate personal_status
## 1 unknown > 7 yrs 4 single male
## 2 < 100 DM 1 - 4 yrs 2 female
## 3 > 1000 DM 4 - 7 yrs 2 divorced male
## 4 < 100 DM 1 - 4 yrs 1 female
## 5 101 - 500 DM 1 - 4 yrs 4 female
## 6 unknown > 7 yrs 4 single male
## other_debtors residence_history property age installment_plan
## 1 none 4 real estate 67 none
## 2 none 2 real estate 22 none
## 3 none 4 real estate 61 none
## 4 none 1 other 22 none
## 5 none 2 other 32 none
## 6 none 4 building society savings 53 none
## housing existing_credits default dependents telephone foreign_worker
## 1 own 2 1 1 yes yes
## 2 own 1 2 1 none yes
## 3 own 1 1 1 none yes
## 4 own 1 1 1 yes yes
## 5 own 1 2 1 none yes
## 6 own 2 1 1 none yes
## job
## 1 skilled employee
## 2 skilled employee
## 3 unskilled resident
## 4 skilled employee
## 5 unskilled resident
## 6 skilled employee
```

Kita akan melihat berapa banyak pekerja asing yang mengajukan credit utk tujuan radio/tv

```
radiotv %>%
  group_by(foreign_worker) %>%
  count() %>%
  arrange(-n)
```

```
## # A tibble: 2 x 2
## # Groups:   foreign_worker [2]
##   foreign_worker     n
##   <chr>           <int>
## 1 yes             275
## 2 no              5
```

Bisa kita lihat bahwa pekerja asing dengan tujuan credit utk radio/tv ada sebanyak 275
 Sekarang kita akan melihat berapa jumlah pengajuan credit dilihat dari jenis pekerjaan (job)

```
radiotv %>%
  group_by(job) %>%
  count() %>%
  arrange(-n)
```

```
## # A tibble: 4 x 2
## # Groups:   job [4]
##   job              n
##   <chr>          <int>
## 1 skilled employee 195
## 2 unskilled resident 57
## 3 mangement self-employed 26
## 4 unemployed non-resident 2
```

Jenis pekerjaan yang paling banyak mengajukan credit utk radio/tv adalah *skilled employee*

```
radiotv %>%
  group_by(personal_status) %>%
  count() %>%
  arrange(-n)
```

```
## # A tibble: 4 x 2
## # Groups:   personal_status [4]
##   personal_status     n
##   <chr>           <int>
## 1 single male      146
## 2 female          85
## 3 married male    42
## 4 divorced male    7
```

Jumlah pekerja dengan status *single male* ada sebanyak 146 orang
 Kita akan melihat hubungan antara jenis pekerjaan dengan personal status

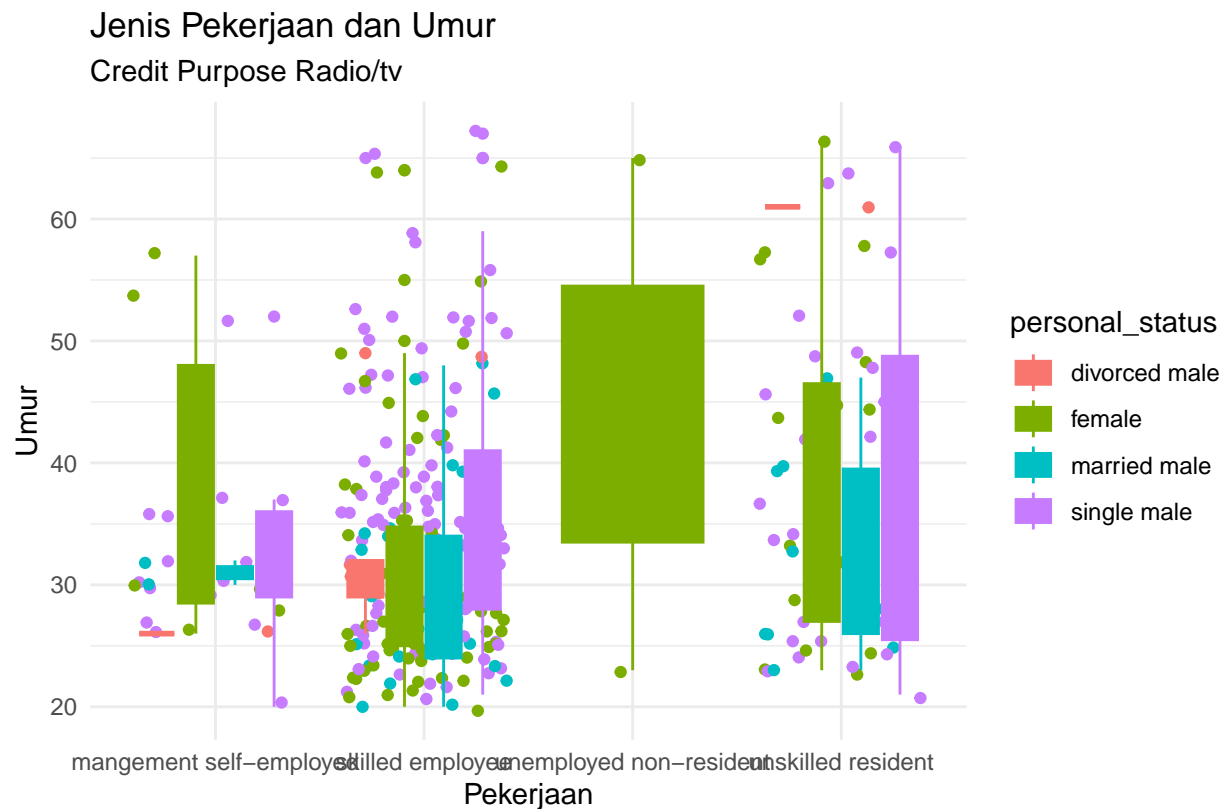
```
table(radiotv$job, radiotv$personal_status)
```

```
##
##               divorced male female married male single male
## mangement self-employed         1      6          2      17
## skilled employee             5     58         29     103
## unemployed non-resident       0      2          0      0
## unskilled resident            1     19         11     26
```

Bisa dilihat bahwa jenis pekerjaan *skill employee* dengan status *single male* yang paling banyak yaitu 103 orang

Visualisasi yang mengajukan credit dengan tujuan radio/tv dilihat dari umur dan jenis pekerjaan
radiotv %>%

```
ggplot(aes(x=job, y=age, col=personal_status, fill=personal_status)) +
  geom_jitter() +
  geom_boxplot() +
  labs(
    title = "Jenis Pekerjaan dan Umur",
    subtitle = "Credit Purpose Radio/tv",
    caption = "by: Roni Yunis",
    x = "Pekerjaan",
    y = "Umur"
  ) +
  theme_minimal()
```



by: Roni Yunis

Membagi Dataset

```
set.seed(100) #pengambilan data secara random
#untuk data training diambil 70%, sisanya untuk data testing, berdasarkan variabel foreign_worker
index_train <- createDataPartition(credit$foreign_worker,
                                   p = 0.7,list = FALSE)

data.train <- credit[index_train,]
data.test <- credit[-index_train,]
```

```
dim(data.train)
```

```
## [1] 701  21
```

```
dim(data.test)
```

```
## [1] 299  21
```

Setelah kita bagi, maka bisa dijelaskan bahwa untuk data training ada 701 baris data dan untuk data testing ada 299 baris data yang kita gunakan utk mendukung klasifikasi yang akan dilakukan.

Model Klasifikasi dengan Decision Tree

Memodelkan klasifikasi

```
modelTree <- rpart(data=data.train,
                   foreign_worker~.,
                   control = rpart.control(cp=0, minsplit=15))
```

opsi minsplit = 15 mengandung pengertian bahwa jika ada node yang berukuran kurang dari 15, maka algoritma dihentikan

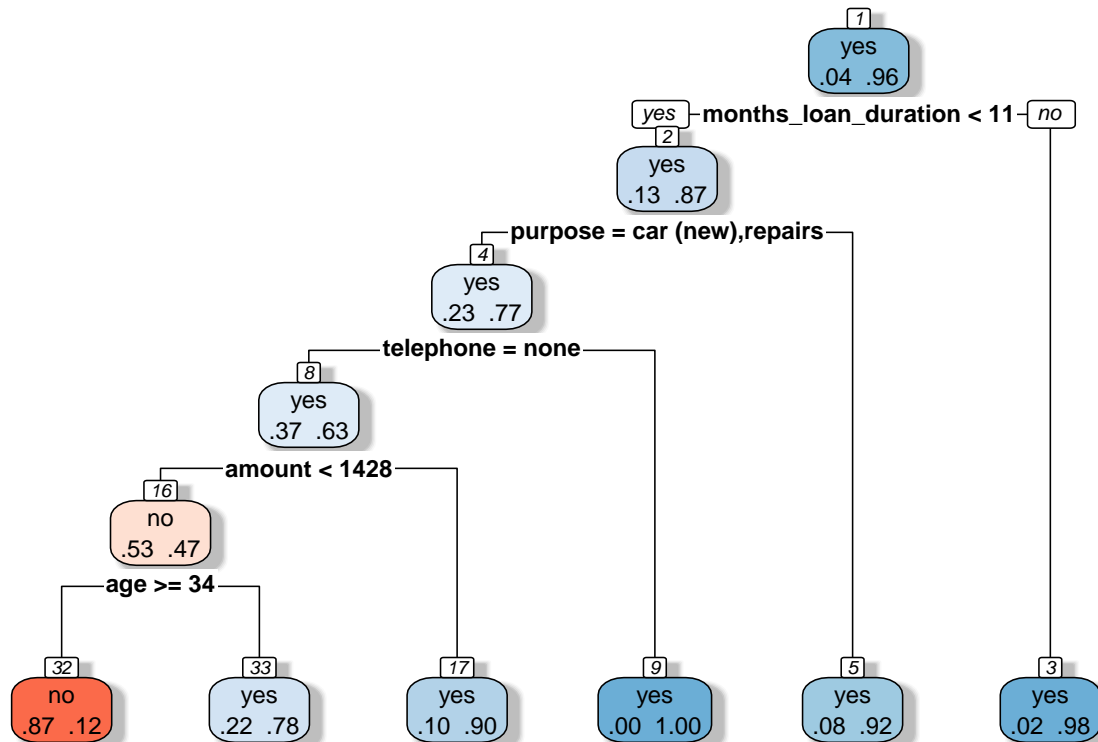
```
modelTree
```

```
## n= 701
##
## node), split, n, loss, yval, (yprob)
##      * denotes terminal node
##
## 1) root 701 26 yes (0.03708987 0.96291013)
##    2) months_loan_duration< 10.5 122 16 yes (0.13114754 0.86885246)
##      4) purpose=car (new),repairs 43 10 yes (0.23255814 0.76744186)
##        8) telephone=none 27 10 yes (0.37037037 0.62962963)
##          16) amount< 1427.5 17  8 no (0.52941176 0.47058824)
##            32) age>=33.5 8   1 no (0.87500000 0.12500000) *
##            33) age< 33.5 9   2 yes (0.22222222 0.77777778) *
##          17) amount>=1427.5 10  1 yes (0.10000000 0.90000000) *
##        9) telephone=yes 16   0 yes (0.00000000 1.00000000) *
##      5) purpose=business,car (used),domestic appliances,education,furniture,others,radio/tv,retraining
##    3) months_loan_duration>=10.5 579 10 yes (0.01727116 0.98272884) *
```

Visualisasi Model Klasifikasi

```
# Menampilkan pohon klasifikasi
```

```
rpart.plot(modelTree, extra=4, box.palette="RdBu", shadow.col="gray", nn=TRUE)
```



Dari gambaran visualisasi diatas bisa dijelaskan bahwa keputusan terbaik untuk foreign_worker yang mengajukan credit adalah dengan durasi lama pinjaman < 11 bulan, dgn tujuan pinjaman utk membeli mobil baru dengan peluang sebesar 0,87.

Mengukur Kinerja Prediksi

```
prediksiTree <- predict(modelTree, data.test)
head(prediksiTree, n=10)
```

```
##          no          yes
## 3  0.01727116 0.9827288
## 5  0.01727116 0.9827288
## 7  0.01727116 0.9827288
## 8  0.01727116 0.9827288
## 17 0.01727116 0.9827288
## 20 0.01727116 0.9827288
## 22 0.07594937 0.9240506
## 23 0.10000000 0.9000000
```

```
## 27 0.07594937 0.9240506
## 31 0.01727116 0.9827288
```

```
prediksi.status.t <- ifelse(prediksiTree[,2] > 0.5, "yes", "no")
#menghitung ukuran kinerja prediksi
confusionMatrix(as.factor(prediksi.status.t), as.factor(data.test$foreign_worker))
```

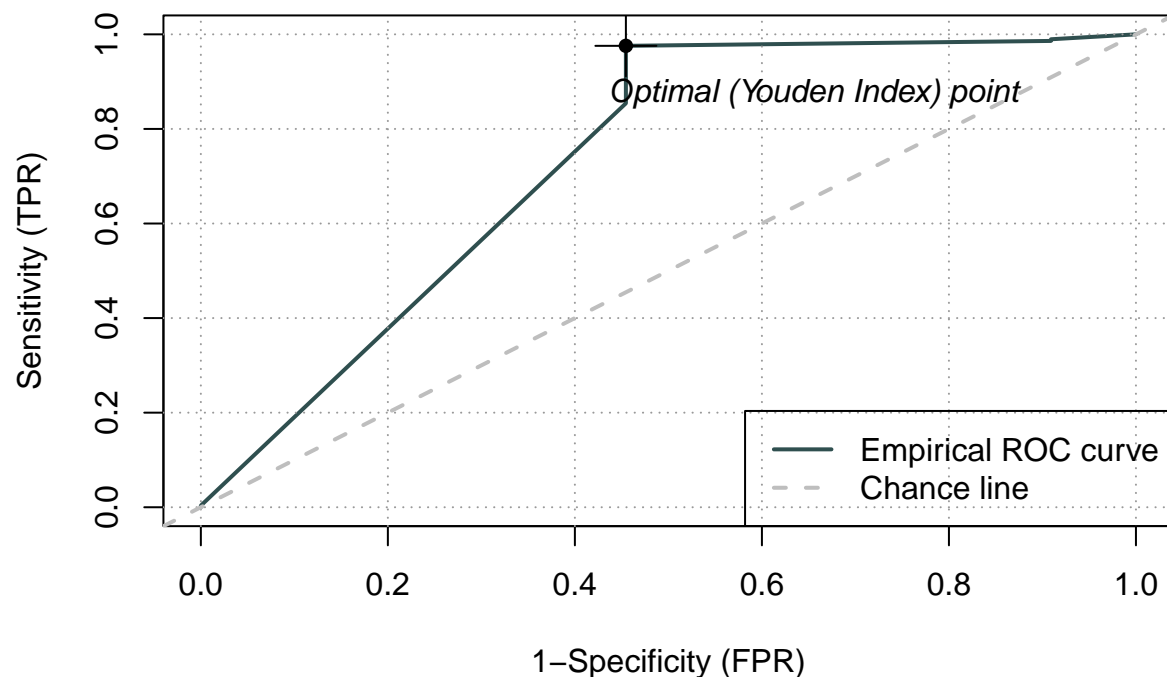
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  no yes
##          no   1   3
##          yes 10 285
##
##           Accuracy : 0.9565
##           95% CI : (0.9268, 0.9766)
##       No Information Rate : 0.9632
##       P-Value [Acc > NIR] : 0.78480
##
##           Kappa : 0.116
##
##  Mcnemar's Test P-Value : 0.09609
##
##           Sensitivity : 0.090909
##           Specificity : 0.989583
##           Pos Pred Value : 0.250000
##           Neg Pred Value : 0.966102
##           Prevalence : 0.036789
##           Detection Rate : 0.003344
##       Detection Prevalence : 0.013378
##           Balanced Accuracy : 0.540246
##
##           'Positive' Class : no
##
```

Berdasarkan hasil diatas bisa lihat bahwa nilai akurasi sebesar 95,6%

Hitung Nilai Performance dari Prediksi

Kurva ini digunakan untuk menilai hasil prediksi

```
ngitungROct <- rocit(score=prediksiTree[,2],class=data.test$foreign_worker)
plot(ngitungROct)
```



Setelah didapatkan nilai curva, maka langkah selanjutnya adalah menghitung Area Under Curve (AUC) yang nantinya dijadikan sebagai dasar untuk menentukan ketepatan prediksi klasifikasi yang sudah lakukan. Nilai AUC bisa dikelompokkan atas: a. 0.90 - 1.00 = Excellence Classification b. 0.80 - 0.90 = Good Classification c. 0.70 - 0.80 = Fair Classification d. 0.60 - 0.70 = Poor Classification e. 0.50 - 0.60 = Failur

Dalam banyak kasus, nilai AUC ini juga digunakan untuk mengukur perbedaan performansi metode klasifikasi.

```
# Menghitung Area Under Curve (AUC)
AUCtree <- ngitungROct$AUC
AUCtree
```

```
## [1] 0.7312184
```

Nilai AUC nya adalah 73,1%, artinya klasifikasi yang dihasilkan termasuk pada **fair classification**

Model Klasifikasi dengan Random Forest

Memodelkan klasifikasi

```
set.seed(123) #menentukan nilai acak dari data
modelForest <- randomForest(data=data.train,
                             as.factor(foreign_worker)~.,
                             ntree=100, mtry=3)
```



```
modelForest
```

```
##
## Call:
## randomForest(formula = as.factor(foreign_worker) ~ ., data = data.train,      ntree = 100, mtry = 3,
##               Type of random forest: classification
##               Number of trees: 100
## No. of variables tried at each split: 3
##
## OOB estimate of error rate: 3.71%
## Confusion matrix:
##      no yes class.error
## no   0  26          1
## yes  0 675          0
```

Tingkat kesalahan sebesar 3,71% atau dengan akurasi sebesar 96,29%

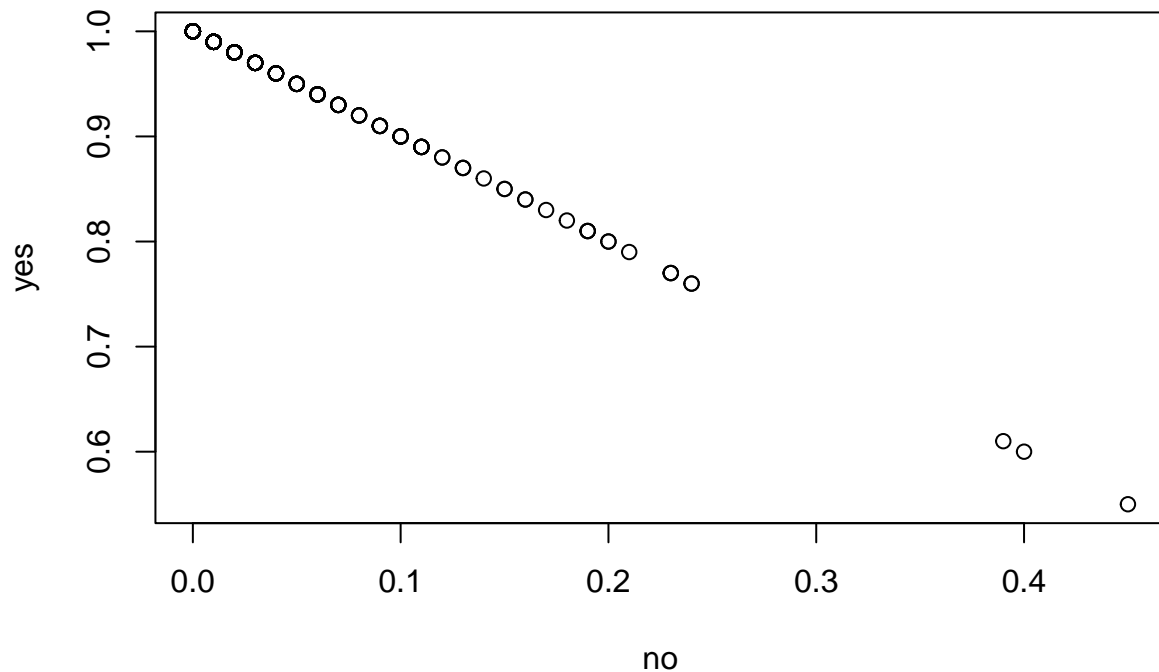
Mengukur kinerja prediksi

```
hasilPrediksi <- predict(modelForest, data.test, type="prob")
head(hasilPrediksi, n=10)
```

```
##      no yes
## 3  0.03 0.97
## 5  0.02 0.98
## 7  0.01 0.99
## 8  0.00 1.00
## 17 0.00 1.00
## 20 0.00 1.00
## 22 0.08 0.92
## 23 0.40 0.60
## 27 0.06 0.94
## 31 0.00 1.00
```

Menampilkan plot hasil prediksi

```
plot(hasilPrediksi )
```



```
prediksi.status.f <- ifelse(hasilPrediksi[,2] > 0.5, "yes", "no")

#menghitung ukuran kinerja prediksi
confusionMatrix(as.factor(prediksi.status.f), as.factor(data.test$foreign_worker))
```

```
## Warning in confusionMatrix.default(as.factor(prediksi.status.f),
## as.factor(data.test$foreign_worker)): Levels are not in the same order for
## reference and data. Refactoring data to match.
```

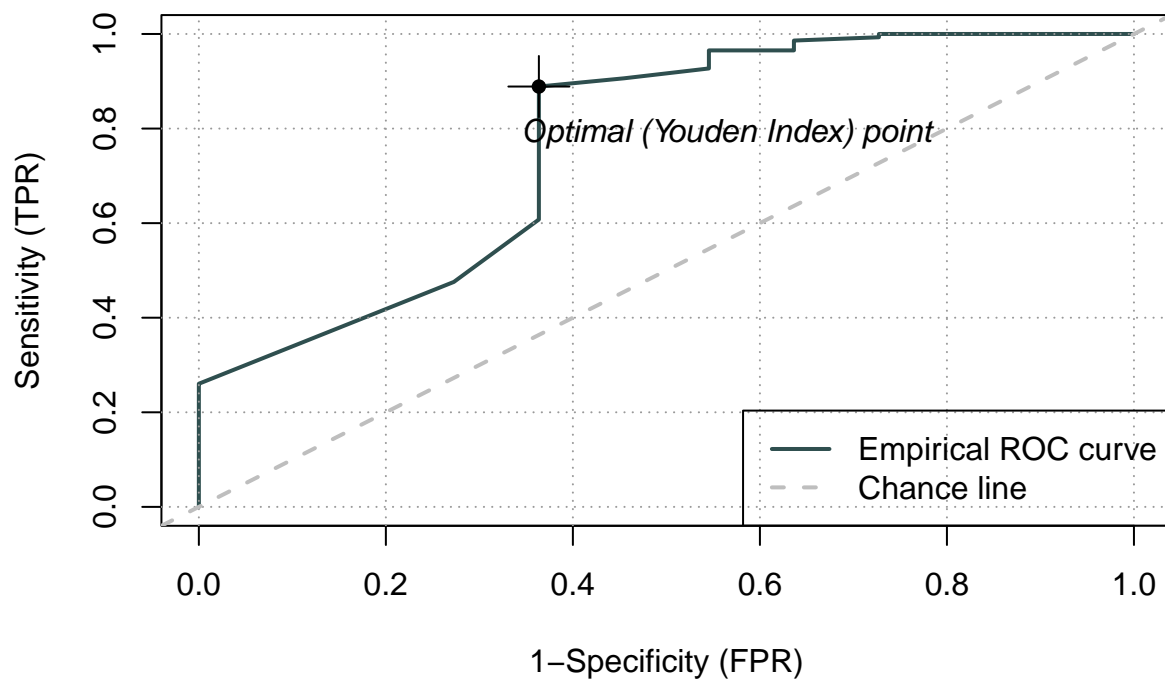
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  no  yes
##      no      0   0
##      yes    11 288
##
##           Accuracy : 0.9632
##           95% CI : (0.9351, 0.9815)
##      No Information Rate : 0.9632
##      P-Value [Acc > NIR] : 0.579281
##
##           Kappa : 0
##
##      Mcnemar's Test P-Value : 0.002569
##
```

```
##          Sensitivity : 0.00000
##          Specificity : 1.00000
##          Pos Pred Value :      NaN
##          Neg Pred Value : 0.96321
##          Prevalence : 0.03679
##          Detection Rate : 0.00000
##          Detection Prevalence : 0.00000
##          Balanced Accuracy : 0.50000
##
##          'Positive' Class : no
##
```

Berdasarkan hasil diatas bisa lihat bahwa nilai akurasi sebesar 96,3 %

Hitung Nilai Performance dari Prediksi

```
ngitungROCf <- rocit(score=hasilPrediksi[,2],class=data.test$foreign_worker)
plot(ngitungROCf)
```



```
AUCf <- ngitungROCf$AUC
AUCf
```

```
## [1] 0.7649937
```

Nilai AUC nya adalah 76,5%, artinya klasifikasi yang dihasilkan termasuk pada **fair classification**

Kalau kita bandingkan dari kedua model tersebut, kinerja dari klasifikasi dengan *Random Forest* **lebih baik sedikit** dibandingkan dengan *Decision Tree*