Latihan Responsi

2210512015 - Deni Permana

2024-11-16

Latihan Responsi

Kerjakan soal-soal berikut dengan teliti!

Import Library dan *Dataset*

1. *Import Library* (5 poin)

Library yang dibutuhkan adalah `tidyverse`, `tidymodels`, dan`here`.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 1.2.0 --
## v broom        1.0.7     v rsample      1.2.1
## v dials        1.3.0     v tune         1.2.1
## v infer        1.0.7     v workflows    1.1.4
## v modeldata    1.4.0     v workflowsets 1.1.0
## v parsnip      1.2.1     v yardstick    1.3.1
## v recipes      1.1.0
## -- Conflicts ------------------------------------------ tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Dig deeper into tidy modeling with R at https://www.tmwr.org
```

```r
library(here)
```

```
## here() starts at E:/PMM JOGJA/Prak Data Science/Prak-DS-DP-UPNVYK-2024
```

2. *Import Dataset* (5 poin)

Import *dataset* **airquality1.csv** dan **airquality2.csv** dengan menggunakan *library* `here`, lalu tampilkan 10 data pertama.

```r
# airquality1.csv
airquality1 <- read_csv(here("Pertemuan 9/datasetP9/airquality1.csv"))
```

```
## Rows: 153 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (4): X, Ozone, Solar.R, Wind
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
head(airquality1, 10)
```

```
## # A tibble: 10 x 4
##        X Ozone Solar.R  Wind
##    <dbl> <dbl>   <dbl> <dbl>
##  1     1    41     190   7.4
##  2     2    36     118   8
##  3     3    12     149  12.6
##  4     4    18     313  11.5
##  5     5    NA      NA  14.3
##  6     6    28      NA  14.9
##  7     7    23     299   8.6
##  8     8    19      99  13.8
##  9     9     8      19  20.1
## 10    10    NA     194   8.6
```

```r
# airquality2.csv
airquality2 <- read_csv(here("Pertemuan 9/datasetP9/airquality2.csv"))
```

```
## Rows: 153 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (4): X, Temp, Month, Day
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
head(airquality2, 10)
```

```
## # A tibble: 10 x 4
##        X  Temp Month   Day
##    <dbl> <dbl> <dbl> <dbl>
##  1     1    67     5     1
##  2     2    72     5     2
##  3     3    74     5     3
##  4     4    62     5     4
##  5     5    56     5     5
##  6     6    66     5     6
##  7     7    65     5     7
##  8     8    59     5     8
##  9     9    61     5     9
## 10    10    69     5    10
```

Data Preprocessing

3. *Data Imputation* (15 poin)

Dari soal sebelumnya, dapat dilihat bahwa *dataset* **airquality1** memiliki nilai N/A pada beberapa kolom. Hapus nilai N/A atau lakukan imputasi data sederhana untuk mengisi nilai N/A, lalu tampilkan 10 data pertama.

```
# Hapus N/A
airquality1 <- na.omit(airquality1)
head(airquality1, 10)
```

```
## # A tibble: 10 x 4
##        X Ozone Solar.R  Wind
##    <dbl> <dbl>   <dbl> <dbl>
##  1     1    41     190   7.4
##  2     2    36     118   8
##  3     3    12     149  12.6
##  4     4    18     313  11.5
##  5     7    23     299   8.6
##  6     8    19      99  13.8
##  7     9     8      19  20.1
##  8    12    16     256   9.7
##  9    13    11     290   9.2
## 10    14    14     274  10.9
```

4. *Joining Table* (10 poin)

Perhatikan *dataset* **airquality1** dan **airquality2**, ada satu kolom yang sama dari kedua *dataset* tersebut. Gunakan kolom tersebut untuk menyatukan kedua *dataset* ke dalam variabel baru bernama **airquality**. Tampilkan 6 data terakhirnya.

```
library(dplyr)

airquality <- inner_join(airquality1, airquality2, by = "X")
tail(airquality)
```

```
## # A tibble: 6 x 7
##       X Ozone Solar.R  Wind  Temp Month   Day
##   <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1   147     7      49  10.3    69     9    24
## 2   148    14      20  16.6    63     9    25
## 3   149    30     193   6.9    70     9    26
## 4   151    14     191  14.3    75     9    28
## 5   152    18     131   8      76     9    29
## 6   153    20     223  11.5    68     9    30
```

5. Pemilihan Kolom (10 poin)

Buat kolom baru bernama Date yang menyimpan kombinasi tanggal dari kolom Month dan Day dengan format yyyy-MM-dd (tahun = 1973). Gunakan fungsi `paste` untuk menggabungkan string dan fungsi `as.POSIXct` untuk mengubah string menjadi tanggal. Manfaatkan fungsi *help* sebaik-baiknya.

```r
airquality$Date <- as.POSIXct(paste(1973, airquality$Month, airquality$Day, sep = "-"),
                              format = "%Y-%m-%d")
head(airquality)
```

```
## # A tibble: 6 x 8
##       X Ozone Solar.R  Wind  Temp Month   Day Date
##   <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dttm>
## 1     1    41     190   7.4    67     5     1 1973-05-01 00:00:00
## 2     2    36     118   8      72     5     2 1973-05-02 00:00:00
## 3     3    12     149  12.6    74     5     3 1973-05-03 00:00:00
## 4     4    18     313  11.5    62     5     4 1973-05-04 00:00:00
## 5     7    23     299   8.6    65     5     7 1973-05-07 00:00:00
## 6     8    19      99  13.8    59     5     8 1973-05-08 00:00:00
```

Setelah itu, buang kolom X, Day, dan Month yang tidak akan digunakan untuk membuat model. Kemudian, ubah nama kolom Solar.R menjadi Solar_Radiation. Gunakan operator pipeline.

```r
airquality <- airquality %>%
  select(-X, -Day, -Month) %>%
  rename(Solar_Radiation = Solar.R)

head(airquality)
```

```
## # A tibble: 6 x 5
##   Ozone Solar_Radiation  Wind  Temp Date
##   <dbl>           <dbl> <dbl> <dbl> <dttm>
## 1    41             190   7.4    67 1973-05-01 00:00:00
## 2    36             118   8      72 1973-05-02 00:00:00
## 3    12             149  12.6    74 1973-05-03 00:00:00
## 4    18             313  11.5    62 1973-05-04 00:00:00
## 5    23             299   8.6    65 1973-05-07 00:00:00
## 6    19              99  13.8    59 1973-05-08 00:00:00
```

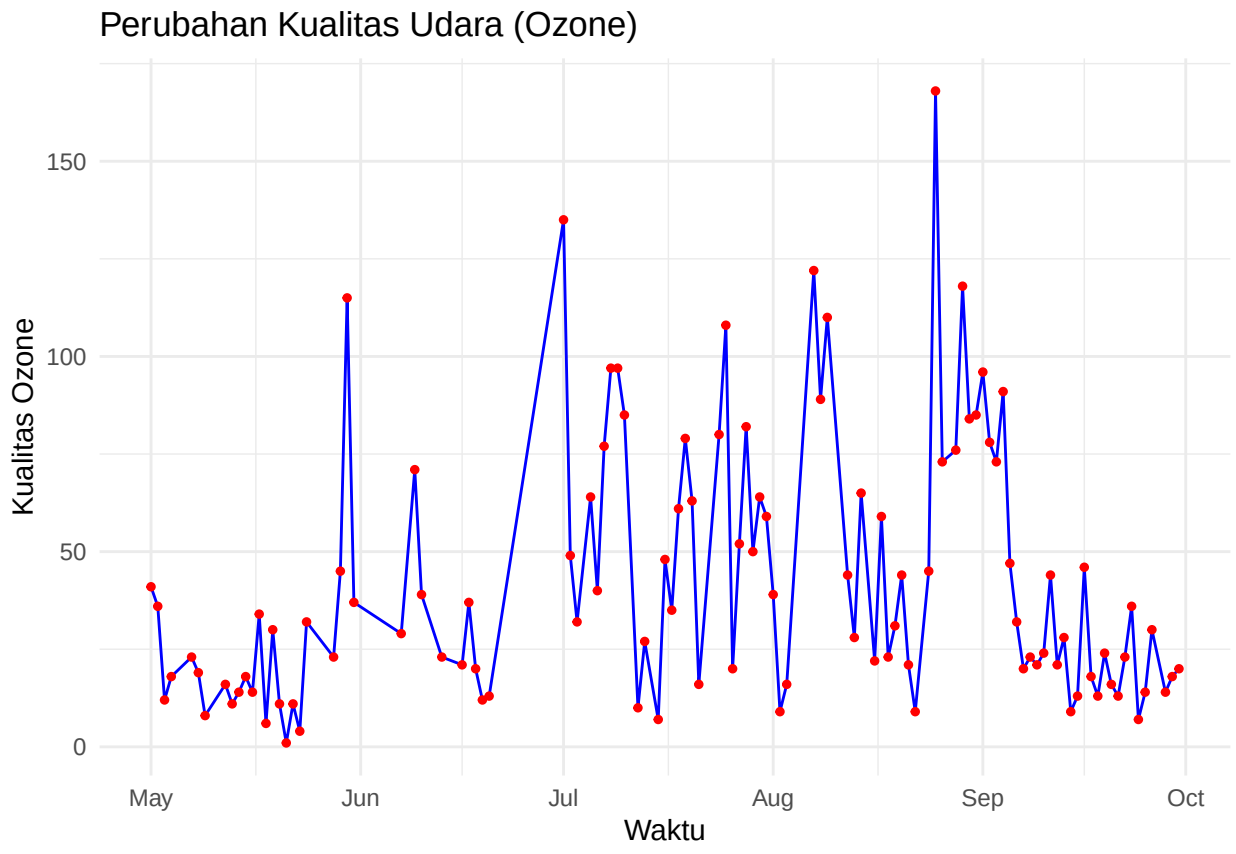
Data Visualization

6. Perubahan Kualitas Udara (10 poin)

Gambarkan perubahan kualitas udara (Ozone) setiap harinya dengan menggunakan ggplot2. Kombinasikan 2 jenis geom yang sesuai dengan data yang ada.

```
library(ggplot2)

ggplot(
  data = airquality,
  aes(
    x = Date,
    y = Ozone)
  ) + geom_line(color = "blue") +
  geom_point(color = "red", size = 1) +
  labs(title = "Perubahan Kualitas Udara (Ozone)",
       x = "Waktu",
       y = "Kualitas Ozone") +
  theme_minimal()
```



Data Modelling

7. *Scaling Data* (5 poin)

Variabel pada dataset ini memiliki range yang berbeda-beda. Lakukan scaling agar berada di range yang sama.

```
library(scales)

airquality_scaled <- airquality %>%
  mutate(across(c(Ozone, Solar_Radiation, Wind), ~ rescale(.)))

head(airquality_scaled, 5)
```

```
## # A tibble: 5 x 5
##    Ozone Solar_Radiation  Wind  Temp Date
##    <dbl>           <dbl> <dbl> <dbl> <dttm>
## 1 0.240            0.560 0.277    67 1973-05-01 00:00:00
## 2 0.210            0.339 0.310    72 1973-05-02 00:00:00
## 3 0.0659           0.434 0.560    74 1973-05-03 00:00:00
## 4 0.102            0.936 0.5      62 1973-05-04 00:00:00
## 5 0.132            0.893 0.342    65 1973-05-07 00:00:00
```

8. Data Splitting (5 poin)

Bagi dataset untuk *training* dan *testing* dengan proporsi *training* 80%. Pastikan dataset diacak sebelum dibagi, dan pastikan hasil acak akan tetap konsisten walaupun dijalankan berkali-kali dari perangkat berbeda sekalipun.

```
set.seed(125)

data_split <- initial_split(airquality_scaled, prop = 0.8)
airquality_train <- training(data_split)
airquality_test <- testing(data_split)

head(airquality_train, 5)
```

```
## # A tibble: 5 x 5
##    Ozone Solar_Radiation   Wind  Temp Date
##    <dbl>           <dbl>  <dbl> <dbl> <dttm>
## 1 0.216            0.847 1         72 1973-06-17 00:00:00
## 2 0.0359           0.128 0.435     69 1973-09-24 00:00:00
## 3 0.725            0.758 0.0924    89 1973-08-07 00:00:00
## 4 0.180            0.725 0.467     78 1973-08-19 00:00:00
## 5 0.216            0.832 0.277     76 1973-05-31 00:00:00
```

```
head(airquality_test, 5)
```

```
## # A tibble: 5 x 5
##    Ozone Solar_Radiation  Wind  Temp Date
##    <dbl>           <dbl> <dbl> <dbl> <dttm>
```

```
## 1 0.240            0.560 0.277    67 1973-05-01 00:00:00
## 2 0.132            0.893 0.342    65 1973-05-07 00:00:00
## 3 0.108            0.281 0.625    59 1973-05-08 00:00:00
## 4 0.0778           0.817 0.467    68 1973-05-14 00:00:00
## 5 0.0599           0.957 0.777    73 1973-05-22 00:00:00
```

9. Buat Resep (15 poin)

Buat resep untuk *training* data. Tentukan 3 variabel yang menjadi prediktor dan 1 variabel yang menjadi *outcome*. Biarkan variabel Date sebagai ID.

```
airquality_resep <- recipe(Ozone ~ Solar_Radiation + Wind + Temp, data = airquality_train) %>%
  step_normalize(all_predictors())
airquality_train_prepped <- prep(airquality_resep, training = airquality_train)

summary(airquality_resep)
```

```
## # A tibble: 4 x 4
##   variable        type      role      source
##   <chr>           <list>    <chr>     <chr>
## 1 Solar_Radiation <chr [2]> predictor original
## 2 Wind            <chr [2]> predictor original
## 3 Temp            <chr [2]> predictor original
## 4 Ozone           <chr [2]> outcome   original
```

```
summary(airquality_train_prepped)
```

```
## # A tibble: 4 x 4
##   variable        type      role      source
##   <chr>           <list>    <chr>     <chr>
## 1 Solar_Radiation <chr [2]> predictor original
## 2 Wind            <chr [2]> predictor original
## 3 Temp            <chr [2]> predictor original
## 4 Ozone           <chr [2]> outcome   original
```

10. Terapkan Resep (5 poin)

Terapkan resep yang sudah dibuat ke data *training* dan *testing*.

```
# airquality_training
airquality_train_processed <- bake(airquality_train_prepped, new_data = airquality_train)

head(airquality_train_processed, 5)
```

```
## # A tibble: 5 x 4
##   Solar_Radiation   Wind     Temp  Ozone
##             <dbl>  <dbl>    <dbl>  <dbl>
## 1           1.04   3.05  -0.635   0.216
## 2          -1.54   0.121 -0.950   0.0359
## 3           0.720 -1.66   1.15    0.725
## 4           0.599  0.291 -0.00477 0.180
## 5           0.984 -0.697 -0.215   0.216
```

```r
# airquality_testing
airquality_test_processed <- bake(airquality_train_prepped, new_data = airquality_test)

head(airquality_test_processed, 5)
```

```
## # A tibble: 5 x 4
##   Solar_Radiation   Wind   Temp  Ozone
##             <dbl>  <dbl>  <dbl>  <dbl>
## 1         0.00637 -0.697 -1.16  0.240
## 2         1.20    -0.358 -1.37  0.132
## 3        -0.993    1.11  -2.00  0.108
## 4         0.929    0.291 -1.05  0.0778
## 5         1.43     1.90  -0.530 0.0599
```

11. *Training* Model (10 poin)

Training model berdasarkan data yang sudah diolah.

```r
hasil_model <- linear_reg() %>%
  set_engine("lm")

airquality_trained_model <- hasil_model %>%
  fit(Ozone ~ Solar_Radiation + Wind + Temp, data = airquality_train_processed)

summary(hasil_model)
```

```
##                       Length Class    Mode
## args                  2      -none-   list
## eng_args              0      quosures list
## mode                  1      -none-   character
## user_specified_mode   1      -none-   logical
## method                0      -none-   NULL
## engine                1      -none-   character
## user_specified_engine 1      -none-   logical
```

```r
summary(airquality_trained_model)
```

```
##              Length Class      Mode
## lvl           0     -none-     NULL
## spec          7     linear_reg list
## fit          12     lm         list
## preproc       1     -none-     list
## elapsed       1     -none-     list
## censor_probs  0     -none-     list
```

12. Evaluasi Model (5 poin)

Evaluasi performa model menggunakan data *testing* (performanya jelek juga gapapa)

```r
# Evaluasi performa model menggunakan data testing
airquality_predictions <- predict(airquality_trained_model, new_data = airquality_test_processed)

# Menghitung evaluasi metrik (misalnya RMSE)
library(yardstick)

airquality_results <- airquality_test_processed %>%
  bind_cols(airquality_predictions) %>%
  metrics(truth = Ozone, estimate = .pred)


print(airquality_results)
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       0.162
## 2 rsq     standard       0.547
## 3 mae     standard       0.102
```