Model_Ensemble

Roni Yunis

2023-11-01

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- 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(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:dplyr':
##
##     combine
##
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(e1071)
library(Metrics)
```

```
##
## Attaching package: 'Metrics'
##
## The following objects are masked from 'package:caret':
##
##     precision, recall
```

```r
library(readxl)
```

Konsep Model Ensemble

Model Ensemble adalah teknik yang digunakan dalam analitika data untuk menggabungkan hasil dari beberapa model prediksi atau algoritma berbeda menjadi satu model yang lebih optimal. **Tujuan** umum dari model ensemble adalah meningkatkan akurasi prediksi dan mengurangi risiko overfitting yang mungkin terjadi jika hanya menggunakan satu model tunggal, serta meningkatkan stabilitas dari model sehingga hasil prediksi lebih konsisten dan dapat dihandalkan.

Manfaat model ensemble diantaranya adalah:

1. *Meningkatkan Prediksi*: Dengan menggabungkan kekuatan berbagai model, Ensemble dapat memberikan prediksi yang lebih akurat dan dapat diandalkan.
2. *Pengambilan Keputusan yang Lebih Baik*: Dengan hasil yang lebih akurat, bisnis dapat membuat keputusan yang lebih baik dalam hal perencanaan strategi, manajemen risiko, dan alokasi sumber daya.
3. *Fleksibilitas*: Model Ensemble dapat digunakan dalam berbagai jenis masalah bisnis, seperti klasifikasi, regresi, atau segmentasi pelanggan, sehingga memberikan fleksibilitas dalam menerapkan analitika data untuk berbagai keperluan.
4. *Mengatasi Heterogenitas Data*: Dalam beberapa kasus, data mungkin sangat heterogen atau tidak sesuai dengan model tunggal tertentu. Ensemble dapat membantu mengatasi tantangan ini dengan menggabungkan model yang berbeda untuk berbagai bagian data.

Contoh Model Ensemble

Berikut ini kita akan menggunakan model ensemble untuk prediksi produksi padi, dengan menggabungkan 2 model tunggal untuk mendapatkan tingkat akurasi yang lebih baik.

Obstain Data

```r
padi_sumut <- read_excel("data/bps_padi_sumut.xlsx")
head(padi_sumut)
```

```
## # A tibble: 6 x 5
##   Tahun `Kabupaten Kota` `Rata-rata produksi` Produksi           `Luas Panen`
##   <dbl> <chr>            <chr>                <chr>              <chr>
## 1  2022 Asahan           61,64                62786.65           10185.41
## 2  2022 Batu Bara        55,4                 71050.570000000007 12827.29
## 3  2022 Binjai           50,34                6266.34            1244.910000000~
## 4  2022 Dairi            49,2                 38714.36           7868.1
## 5  2022 Deli Serdang     60,91                328854.78999999998 53984.69
## 6  2022 Gunungsitoli     51,6                 11017.47           2135.570000000~
```

```
#Melihat struktur data
glimpse(padi_sumut)
```

```
## Rows: 429
## Columns: 5
## $ Tahun                <dbl> 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2~
## $ `Kabupaten Kota`     <chr> "Asahan", "Batu Bara", "Binjai", "Dairi", "Deli S~
## $ `Rata-rata produksi` <chr> "61,64", "55,4", "50,34", "49,2", "60,91", "51,6"~
## $ Produksi             <chr> "62786.65", "71050.570000000007", "6266.34", "387~
## $ `Luas Panen`         <chr> "10185.41", "12827.29", "1244.9100000000001", "78~
```

Scrub Data

a. Merubah type data

```
# Merubah type data karakter menjadi numeric
padi_sumut$`Rata-rata produksi` <- as.numeric(padi_sumut$`Rata-rata produksi`)
```

```
## Warning: NAs introduced by coercion
```

```
padi_sumut$Produksi <- as.numeric(padi_sumut$Produksi)
```

```
## Warning: NAs introduced by coercion
```

```
padi_sumut$`Luas Panen` <- as.numeric(padi_sumut$`Luas Panen`)
```

```
## Warning: NAs introduced by coercion
```

```
padi_sumut
```

```
## # A tibble: 429 x 5
##    Tahun `Kabupaten Kota`   `Rata-rata produksi` Produksi `Luas Panen`
##    <dbl> <chr>                             <dbl>    <dbl>        <dbl>
##  1  2022 Asahan                               NA   62787.       10185.
##  2  2022 Batu Bara                            NA   71051.       12827.
##  3  2022 Binjai                               NA    6266.        1245.
##  4  2022 Dairi                                NA   38714.        7868.
##  5  2022 Deli Serdang                         NA  328855.       53985.
```

```
##  6  2022 Gunungsitoli                          NA   11017.        2136.
##  7  2022 Humbang Hasundutan                    NA   75462.       17992.
##  8  2022 Karo                                  NA   69058.        9834.
##  9  2022 Labuanbatu Utara                      NA   80204.       19868.
## 10  2022 Labuhan Batu                          NA   83641.       21456.
## # i 419 more rows
```

```
# Merubah type Tahun menjadi Date
padi_sumut$Tahun <- make_date(padi_sumut$Tahun)
glimpse(padi_sumut)
```

```
## Rows: 429
## Columns: 5
## $ Tahun                <date> 2022-01-01, 2022-01-01, 2022-01-01, 2022-01-01, ~
## $ `Kabupaten Kota`     <chr> "Asahan", "Batu Bara", "Binjai", "Dairi", "Deli S~
## $ `Rata-rata produksi` <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ Produksi             <dbl> 62786.65, 71050.57, 6266.34, 38714.36, 328854.79,~
## $ `Luas Panen`         <dbl> 10185.41, 12827.29, 1244.91, 7868.10, 53984.69, 2~
```

```
# Merubah nama variabel Kabupaten Kota, Rata-rata Produksi dan Luas Panen
names(padi_sumut)[names(padi_sumut) == "Kabupaten Kota"] <- "Kabupaten_kota"
names(padi_sumut)[names(padi_sumut) == "Rata-rata produksi"] <- "Rata_rata_produksi"
names(padi_sumut)[names(padi_sumut) == "Luas Panen"] <- "Luas_panen"
glimpse(padi_sumut)
```

```
## Rows: 429
## Columns: 5
## $ Tahun              <date> 2022-01-01, 2022-01-01, 2022-01-01, 2022-01-01, 20~
## $ Kabupaten_kota     <chr> "Asahan", "Batu Bara", "Binjai", "Dairi", "Deli Ser~
## $ Rata_rata_produksi <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ Produksi           <dbl> 62786.65, 71050.57, 6266.34, 38714.36, 328854.79, 1~
## $ Luas_panen         <dbl> 10185.41, 12827.29, 1244.91, 7868.10, 53984.69, 213~
```

b. Menghapus Data Kosong

```
# Menampikan variabel dengan baris kosong
colSums(is.na(padi_sumut))
```

```
##              Tahun     Kabupaten_kota Rata_rata_produksi           Produksi
##                  0                  0                 61                  2
##         Luas_panen
##                  3
```

```
# Menghapus data NA's
padisumut_clean <- na.omit(padi_sumut)
summary(padisumut_clean)
```

```
##      Tahun            Kabupaten_kota     Rata_rata_produksi    Produksi
##  Min.   :2010-01-01   Length:367         Min.   :28.40      Min.   :    258
```

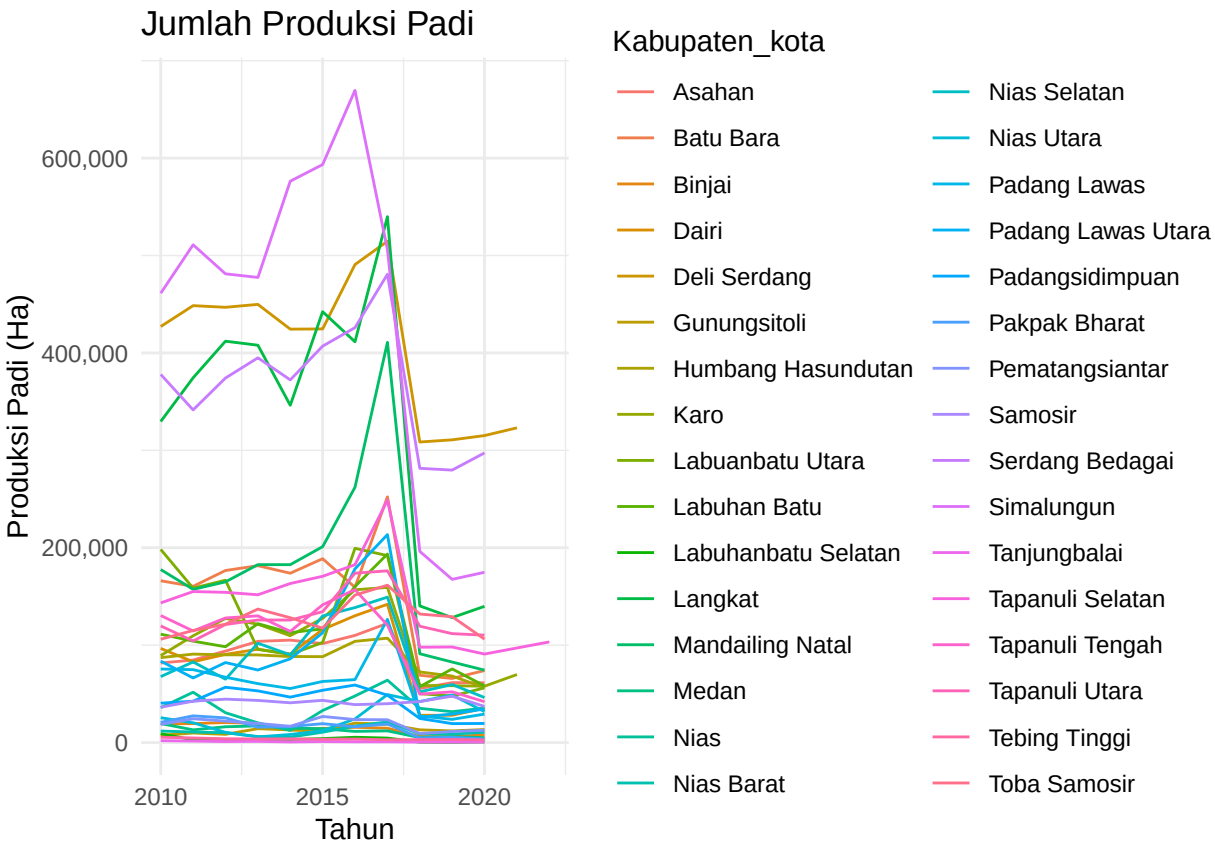
```
##  1st Qu.:2012-01-01   Class :character   1st Qu.:42.84      1st Qu.:  19563
##  Median :2015-01-01   Mode  :character   Median :47.95      Median :  73939
##  Mean   :2015-01-25                      Mean   :47.72      Mean   : 215385
##  3rd Qu.:2018-01-01                      3rd Qu.:52.35      3rd Qu.: 151569
##  Max.   :2022-01-01                      Max.   :71.00      Max.   :5136186
##    Luas_panen
##  Min.   :    68
##  1st Qu.:  4046
##  Median : 17087
##  Mean   : 42714
##  3rd Qu.: 30932
##  Max.   :988068
```

```
padisumutfilter <- padisumut_clean[padisumut_clean$Kabupaten_kota != "Sumatera Utara",]
padisumutfilter
```

```
## # A tibble: 355 x 5
##    Tahun      Kabupaten_kota     Rata_rata_produksi Produksi Luas_panen
##    <date>     <chr>                           <dbl>    <dbl>      <dbl>
##  1 2022-01-01 Tapanuli Selatan                 50    103327.     20802.
##  2 2021-01-01 Deli Serdang                     60    323108      53981
##  3 2021-01-01 Karo                             71     69829       9844
##  4 2020-01-01 Asahan                           57.1   61350.     10737.
##  5 2020-01-01 Batu Bara                        56.9   73939.     12988.
##  6 2020-01-01 Binjai                           54.0    7870.      1456.
##  7 2020-01-01 Dairi                            53.9   35311.      6546.
##  8 2020-01-01 Deli Serdang                     63.5  315156.     49658.
##  9 2020-01-01 Gunungsitoli                     56.8   13352.      2349.
## 10 2020-01-01 Humbang Hasundutan               47.1   56390.     11969.
## # i 345 more rows
```

Ekplorasi Data Analysis

```
plot_1 <- ggplot(padisumutfilter, aes(x = Tahun, y = Produksi, group = Kabupaten_kota, color = Kabupaten
  geom_line() +
  labs(title = "Jumlah Produksi Padi", x = "Tahun", y = "Produksi Padi (Ha)") +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal() +
  theme(legend.position = "right")
plot_1
```



Model

Tujuan adalah mengembangkan model prediksi untuk memperkiraan jumlah produksi padi.

```
# Menyiapkan data pelatihan dan data pengujian
set.seed(123)  # Untuk hasil yang dapat direproduksi
splitIndex <- createDataPartition(padisumutfilter$Produksi, p = 0.7, list = FALSE)
data_train <- padisumutfilter[splitIndex, ]  # Data pelatihan (70%)
data_test <- padisumutfilter[-splitIndex, ]  # Data pengujian (30%)
dim(data_train)
```

```
## [1] 251   5
```

```
dim(data_test)
```

```
## [1] 104   5
```

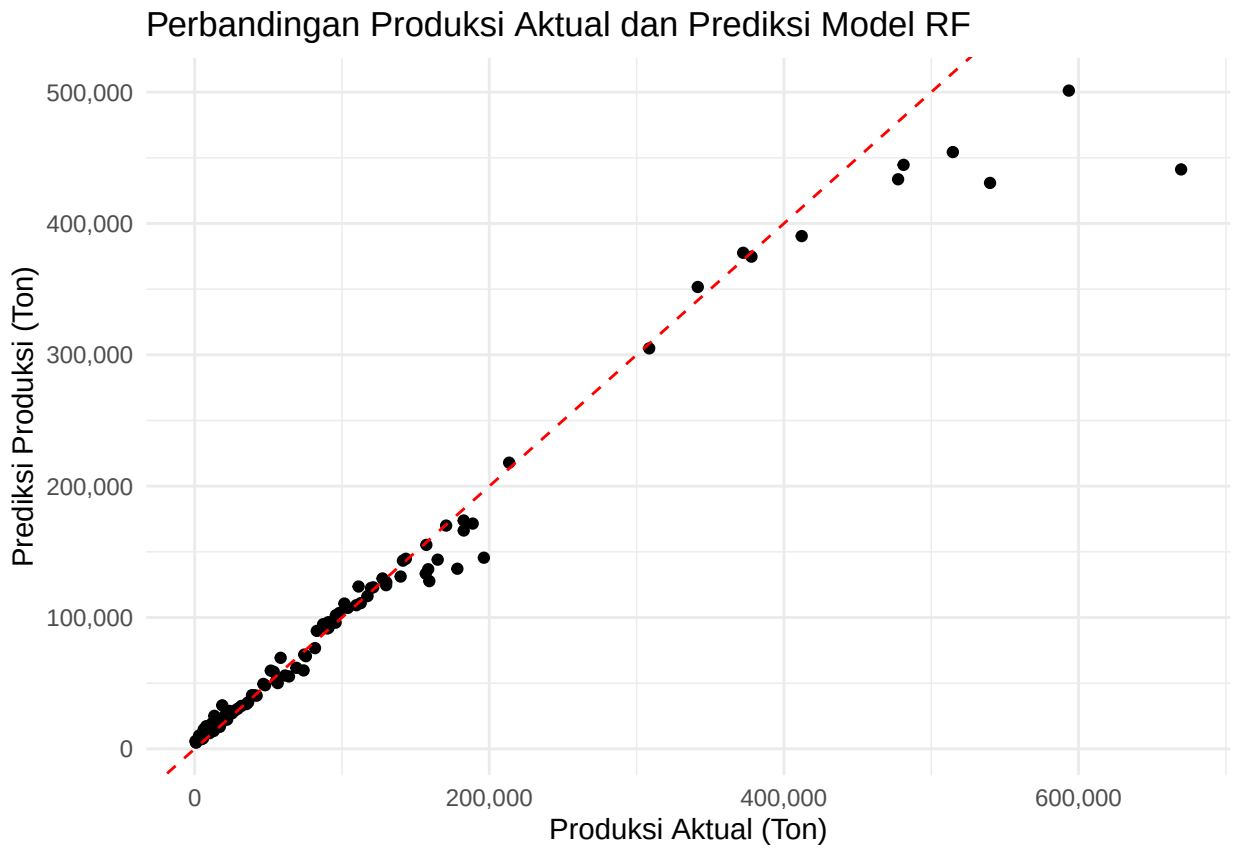
Model Random Forest

```
# Membuat model RF
rf_model <- randomForest(Produksi ~ Luas_panen + Rata_rata_produksi, data = data_train, ntree = 100)
```

```r
# Melakukan Prediksi terhadap Data Pengujian
predictions_rf <- predict(rf_model, data_test)
```

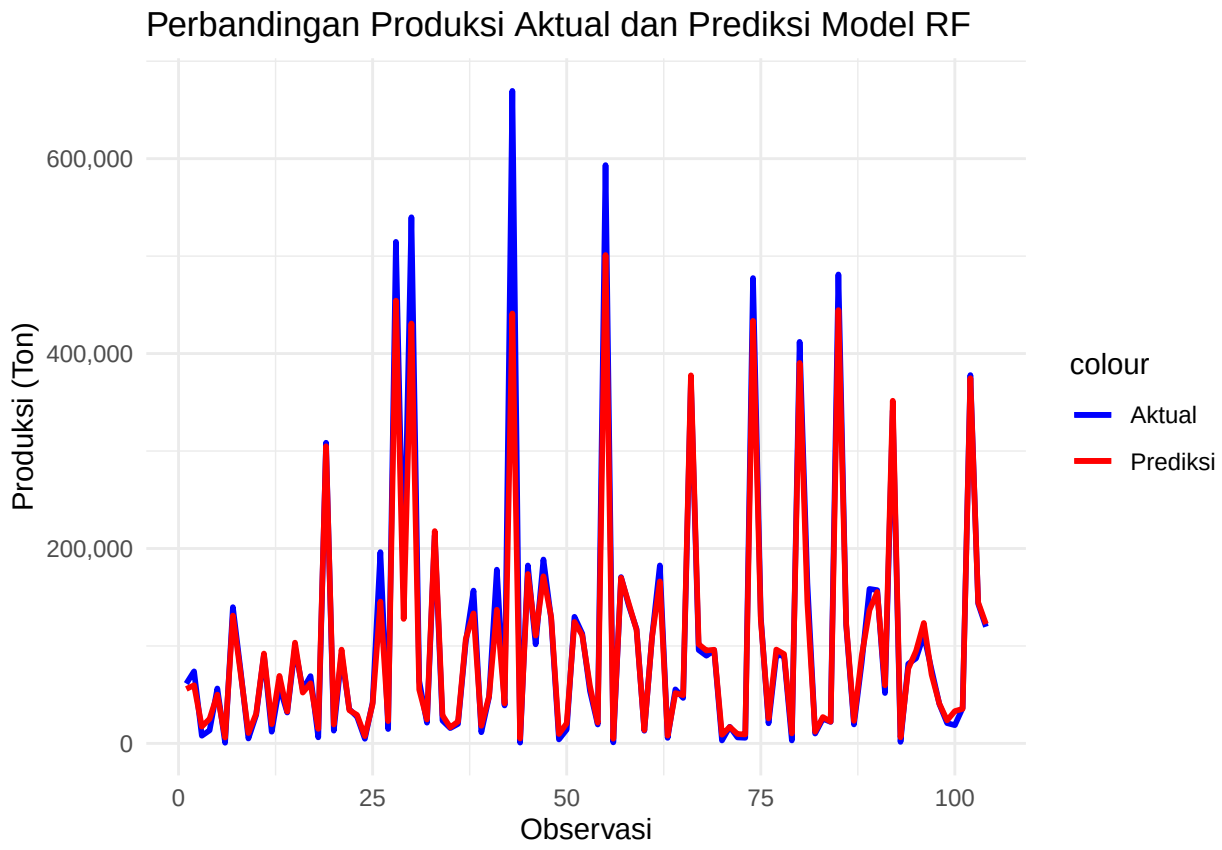
```r
# Visualisasi Sales Aktual dengan Hasil Prediksi
result_data_rf <- data.frame(Produksi = data_test$Produksi, Predictions = predictions_rf)

# Visualisasi Perbandingan Produksi Aktual dengan Hasil Prediksi
ggplot(data = result_data_rf, aes(x = Produksi, y = Predictions)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  scale_x_continuous(labels = scales::comma) +
  scale_y_continuous(labels = scales::comma) +
  labs(x = "Produksi Aktual (Ton)", y = "Prediksi Produksi (Ton)") +
  ggtitle("Perbandingan Produksi Aktual dan Prediksi Model RF") +
  theme_minimal()
```




```r
# Visualisasi dengan Plot Line
ggplot(data = result_data_rf, aes(x = 1:length(Produksi))) +
  geom_line(aes(y = Produksi, color = "Aktual"), size = 1) +
  geom_line(aes(y = Predictions, color = "Prediksi"), size = 1) +
  labs(x = "Observasi", y = "Produksi (Ton)") +
  scale_color_manual(values = c("Aktual" = "blue", "Prediksi" = "red")) +
  ggtitle("Perbandingan Produksi Aktual dan Prediksi Model RF") +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```




```
# Evaluasi RF Model

# Hitung MAE
mae_value_rf <- mae(data_test$Produksi, predictions_rf)

# Hitung MSE
mse_value_rf <- mse(data_test$Produksi, predictions_rf)

# Hitung MAPE
mape_value_rf <- mape(data_test$Produksi, predictions_rf)

# Tampilkan hasil evaluasi
cat(paste("MAE: ", mae_value_rf, "\n"))
```

```
## MAE:  11703.7717535256
```

```
cat(paste("MSE: ", mse_value_rf, "\n"))
```

```
## MSE:  865358224.16532
```

```r
cat(paste("MAPE: ", mape_value_rf, "%\n"))
```

```
## MAPE:  0.415700899907106 %
```

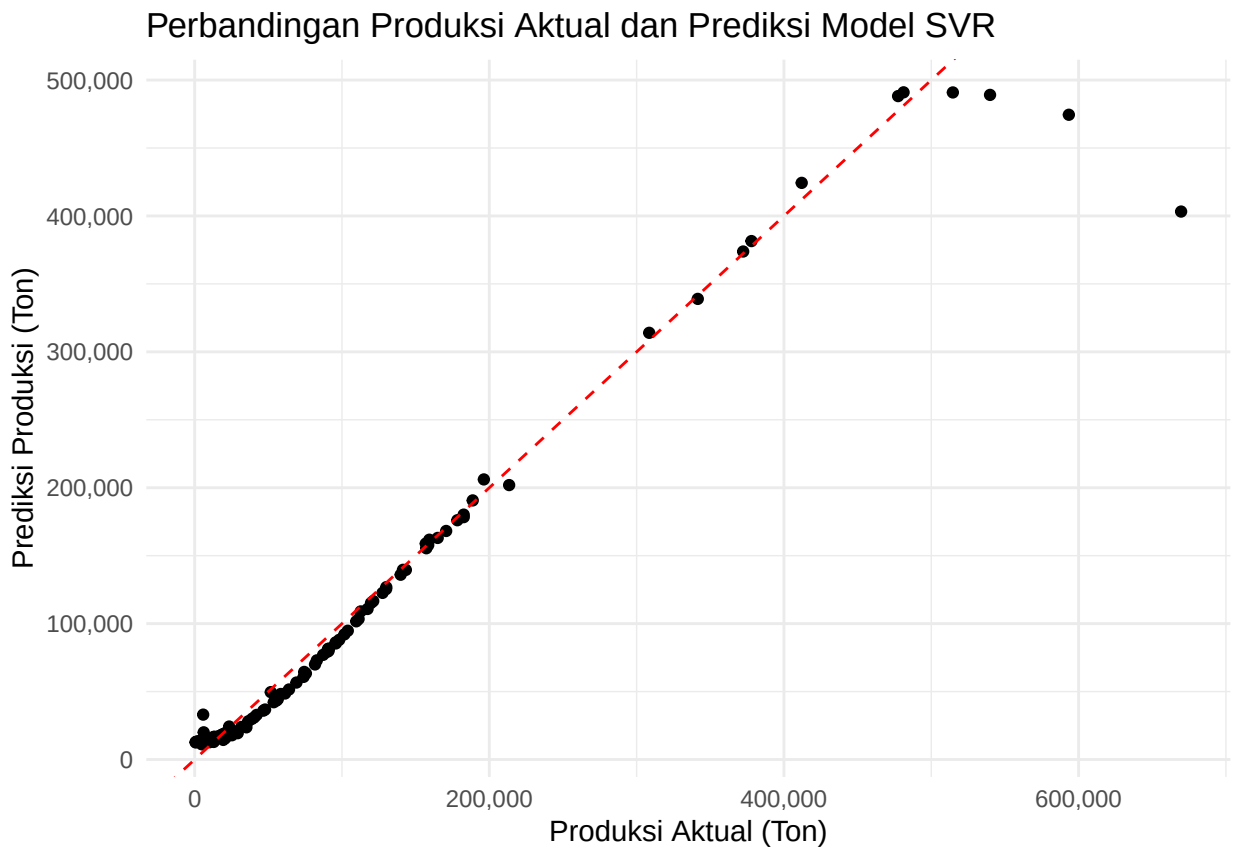
Model SVR

```r
# Membuat model SVR
svr_model <- svm(Produksi ~ Luas_panen + Rata_rata_produksi, data = data_train, kernel = "radial", cost
```

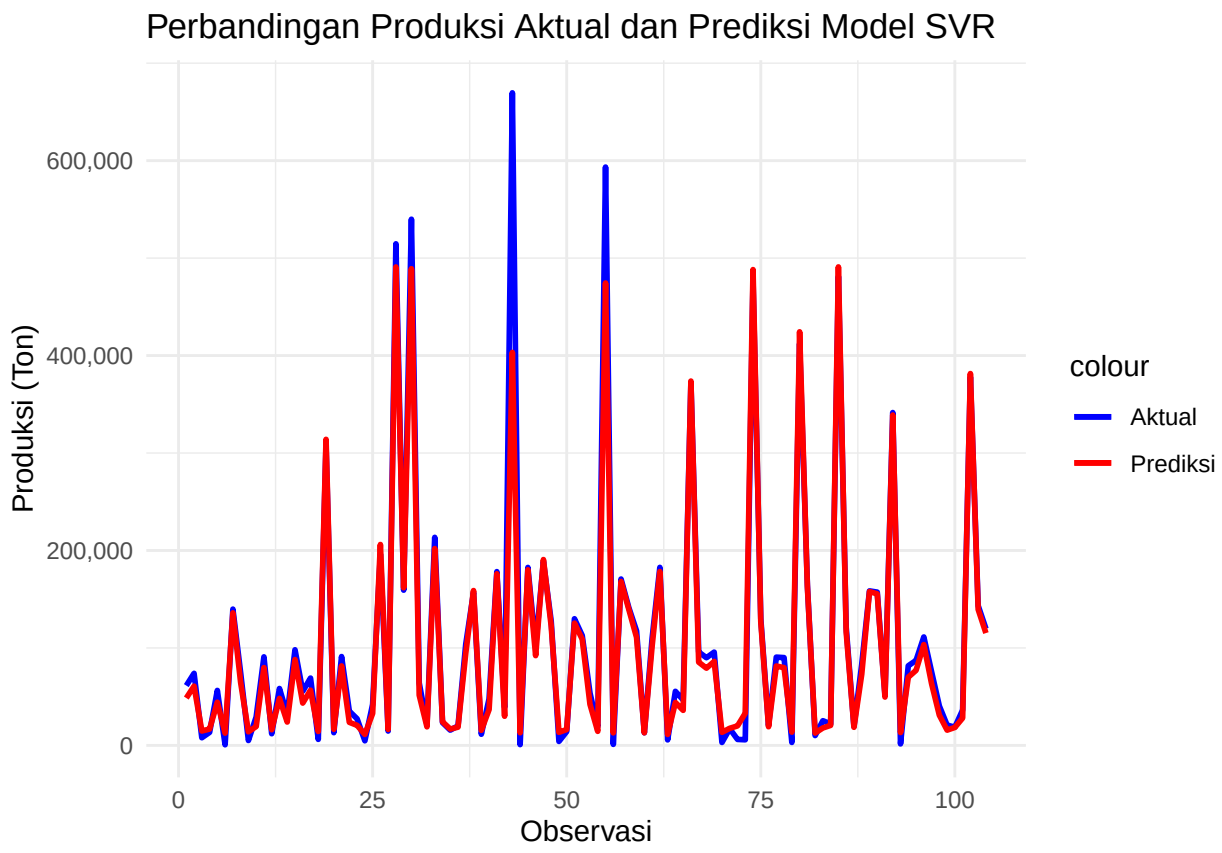
```r
# Melakukan prediksi dengan Data Testing
predictions_svr <- predict(svr_model, data_test)
```

```r
# Visualisasi hasil prediksi
result_data_svr <- data.frame(Produksi = data_test$Produksi, Predictions = predictions_svr)

# Visualisasi Perbandingan Produksi Aktual dengan Hasil Prediksi
ggplot(data = result_data_svr, aes(x = Produksi, y = Predictions)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  labs(x = "Produksi Aktual (Ton)", y = "Prediksi Produksi (Ton)") +
  scale_x_continuous(labels = scales::comma) +
  scale_y_continuous(labels = scales::comma) +
  ggtitle("Perbandingan Produksi Aktual dan Prediksi Model SVR") +
  theme_minimal()
```



```r
# Buat grafik garis dengan ggplot2
ggplot(data = result_data_svr, aes(x = 1:length(Produksi))) +
  geom_line(aes(y = Produksi, color = "Aktual"), size = 1) +
  geom_line(aes(y = Predictions, color = "Prediksi"), size = 1) +
  labs(x = "Observasi", y = "Produksi (Ton)") +
  scale_color_manual(values = c("Aktual" = "blue", "Prediksi" = "red")) +
  ggtitle("Perbandingan Produksi Aktual dan Prediksi Model SVR") +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal()
```




```r
# Evaluasi SVR Model

# Hitung MAE
mae_value_svr <- mae(data_test$Produksi, predictions_svr)

# Hitung MSE
mse_value_svr <- mse(data_test$Produksi, predictions_svr)

# Hitung MAPE
mape_value_svr <- mape(data_test$Produksi, predictions_svr)

# Tampilkan hasil evaluasi
cat(paste("MAE: ", mae_value_svr, "\n"))
```

```
## MAE:  11226.3816094294
```

```r
cat(paste("MSE: ", mse_value_svr, "\n"))
```

```
## MSE:  916184747.895429
```

```r
cat(paste("MAPE: ", mape_value_svr, "%\n"))
```

```
## MAPE:  0.832844950718474 %
```

Hybrid/Ensemble Model SVR dan RF

```r
#Hybrid Model SVR dan rf
predictions_hybrid_svr_rf <- (predictions_svr + predictions_rf) / 2
```

```r
# Hitung MAE
mae_value_hybrid_svr_rf <- mae(data_test$Produksi, predictions_hybrid_svr_rf)

# Hitung MSE
mse_value_hybrid_svr_rf <- mse(data_test$Produksi, predictions_hybrid_svr_rf)

# Hitung MAPE
mape_value_hybrid_svr_rf <- mape(data_test$Produksi, predictions_hybrid_svr_rf)

# Tampilkan hasil evaluasi
cat(paste("MAE: ", mae_value_hybrid_svr_rf, "\n"))
```

```
## MAE:  9747.46748699991
```

```r
cat(paste("MSE: ", mse_value_hybrid_svr_rf, "\n"))
```

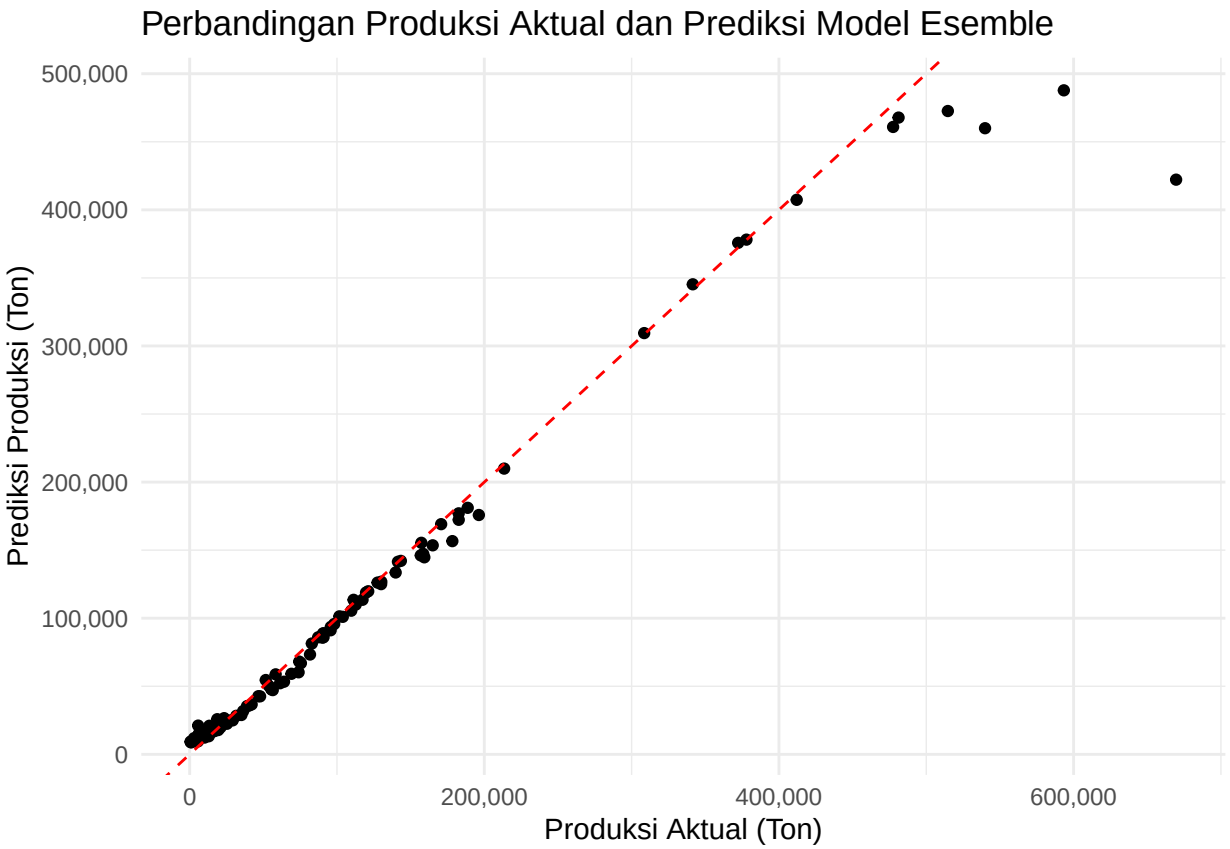
```
## MSE:  820065063.540825
```

```r
cat(paste("MAPE: ", mape_value_hybrid_svr_rf, "%\n"))
```

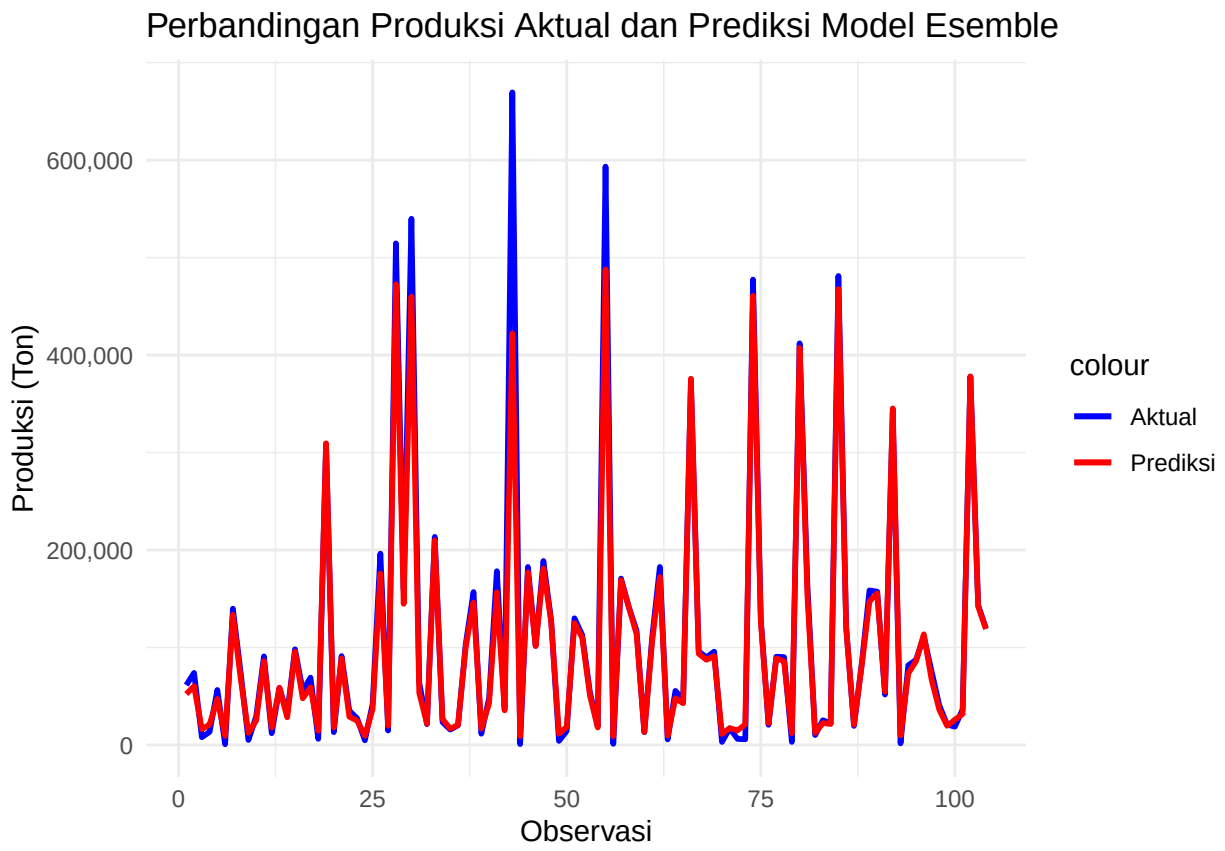
```
## MAPE:  0.603160482397144 %
```

```r
# Visualisasi hasil prediksi dan nilai aktual
result_data <- data.frame(
  Produksi = data_test$Produksi,
  Pred_Hybrid = predictions_hybrid_svr_rf
)

# Visualisasi Perbandingan Produksi Aktual dengan Hasil Prediksi
ggplot(data = result_data, aes(x = Produksi, y = Pred_Hybrid)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  labs(x = "Produksi Aktual (Ton)", y = "Prediksi Produksi (Ton)") +
  scale_x_continuous(labels = scales::comma) +
  scale_y_continuous(labels = scales::comma) +
  ggtitle("Perbandingan Produksi Aktual dan Prediksi Model Esemble") +
  theme_minimal()
```




```r
# Buat grafik garis dengan ggplot2
ggplot(data = result_data, aes(x = 1:length(Produksi))) +
  geom_line(aes(y = Produksi, color = "Aktual"), size = 1) +
  geom_line(aes(y = Pred_Hybrid, color = "Prediksi"), size = 1) +
  labs(x = "Observasi", y = "Produksi (Ton)") +
  scale_color_manual(values = c("Aktual" = "blue", "Prediksi" = "red")) +
  ggtitle("Perbandingan Produksi Aktual dan Prediksi Model Esemble") +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal()
```



iNterpret

```
model_performance <- data.frame(
  No = c(1:3),
  Model = c("Random Forest", "SVR", "Esemble RF & SVR"),
  MSE = c(mse_value_rf, mse_value_svr, mse_value_hybrid_svr_rf),
  MAE = c(mae_value_rf, mae_value_svr, mae_value_hybrid_svr_rf),
  MAPE = c(mape_value_rf, mape_value_svr, mape_value_hybrid_svr_rf),
  stringsAsFactors = FALSE

)
model_performance
```

```
##   No            Model       MSE       MAE      MAPE
## 1  1    Random Forest 865358224 11703.772 0.4157009
## 2  2              SVR 916184748 11226.382 0.8328450
## 3  3 Esemble RF & SVR 820065064  9747.467 0.6031605
```

Berdasarkan hasil diatas bisa disimpulkan bahwa, performansi dari model ensemble lebih baik dari pada model SVR dan RF dalam melakukan prediksi, dimana nilai MAE dari model ensemble adalah sebesar 9747.467 dan nilai MSE sebesar 820065064. Nilai MAE dan MSE dari model ensemble lebih kecil dari 2 model yang lain. Penggabungan model tersebut dapat meningkatkan performansi atau akurasi dalam melakukan prediksi.