# Assessment Test Solution

Candidate: Ramma Hayu Fitra Saleh

Update: 8 augustus 2024

---

## 1. Solusi Pertanyaan Nomor 1

Dari Pernyataan yang diajukan dapat disimpulkan beberapa tujuan yang akan dicapai yaitu sebagai berikut:

**Tujuan**

1. Melakukan analisis statistik terhadap data aditif untuk membuktikan perbedaan signifikan antara formulasi.
2. Memberikan deskripsi statistik, analisis grafis, dan clustering untuk memahami data dan mengelompokkan formulasi.

Untuk mencapai tujuan yang disebutkan diatas ada beberapa langkah yang dapat dilakukan.

1. Melakukan analisa statistik
   Mendeskripsikan statistik dasar. Hasil deskripsi statistik dapat dilihat pada gambar 1.

| | a | b | c | d | e | f | g | h | i |
|---|---|---|---|---|---|---|---|---|---|
| count | 214.000000 | 214.000000 | 214.000000 | 214.000000 | 214.000000 | 214.000000 | 214.000000 | 214.000000 | 214.000000 |
| mean | 1.518365 | 13.407850 | 2.684533 | 1.444907 | 72.650935 | 0.497056 | 8.956963 | 0.175047 | 0.057009 |
| std | 0.003037 | 0.816604 | 1.442408 | 0.499270 | 0.774546 | 0.652192 | 1.423153 | 0.497219 | 0.097439 |
| min | 1.511150 | 10.730000 | 0.000000 | 0.290000 | 69.810000 | 0.000000 | 5.430000 | 0.000000 | 0.000000 |
| 25% | 1.516522 | 12.907500 | 2.115000 | 1.190000 | 72.280000 | 0.122500 | 8.240000 | 0.000000 | 0.000000 |
| 50% | 1.517680 | 13.300000 | 3.480000 | 1.360000 | 72.790000 | 0.555000 | 8.600000 | 0.000000 | 0.000000 |
| 75% | 1.519157 | 13.825000 | 3.600000 | 1.630000 | 73.087500 | 0.610000 | 9.172500 | 0.000000 | 0.100000 |
| max | 1.533930 | 17.380000 | 4.490000 | 3.500000 | 75.410000 | 6.210000 | 16.190000 | 3.150000 | 0.510000 |

Gambar 1. Hasil deskripsi statistik

2. Melakukan uji normalitas. Untuk memastikan apakah data aditif berdistribusi normal. Dari hasil evaluasi pengamatan grafik uji normalitas zat aditif ‘d’ dan ‘h’ (catatan untuk data ‘h’ bernilai 0 untuk semua baris menurut penulis ini tidak normal atau informatif) memenuhi distribusi normal selebihnya tidak memenuhi distribusi normal (berdasarkan analisa hasil uji Shapiro-Wilk).
3. Berdasarkan uji ANOVA dan uji Korelasi. Disimpulkan bawah semua formulasi bensin yang diuji memberikan hasil pembakaran yang mirip dan tidak ada perbedaan yang jelas dalam pola pembakaran antara formulasi-formulasi

tersebut. Diperlihatkan dari p-value 1.000. yang berarti tidak ada perbedaan yang cukup besar untuk dianggap signifikan secara statistik antara setiap pasangan formulasi. Namun, berdasarkan uji korelasi disimpulkan bahwa terdapat korelasi tinggi antara beberapa formulasi menunjukkan bahwa karakteristik pembakaran mereka cenderung bergerak bersama, baik dalam arah yang sama atau berlawanan. Namun, banyak formulasi memiliki korelasi yang rendah, yang menunjukkan bahwa karakteristik pembakaran mereka tidak saling terkait secara signifikan. Dapat disimpulkan bahwa secara statistik tidak ada perbedaan yang signifikan, analisis korelasi memberikan wawasan tambahan tentang hubungan antar formulasi yang dapat berguna dalam evaluasi karakteristik pembakaran.

Gambaran ringkas hasil uji ANOVA dan korelasi dapat diperlihatkan pada Gambar 2 dan Gambar 3 berturut - turut.

ANOVA Results:

| | Column 1 | Column 2 | F-statistic | p-value | Significant Difference |
|---|---|---|---|---|---|
| 0 | a | b | 2.257884e-25 | 1.0 | No |
| 1 | a | c | 2.290123e-25 | 1.0 | No |
| 2 | a | d | 2.230433e-25 | 1.0 | No |
| 3 | a | e | 1.096409e-25 | 1.0 | No |
| 4 | a | f | 2.265284e-25 | 1.0 | No |
| 5 | a | g | 2.380350e-25 | 1.0 | No |
| 6 | a | h | 4.558190e-25 | 1.0 | No |
| 7 | a | i | 2.284734e-25 | 1.0 | No |
| 8 | b | c | 9.190165e-30 | 1.0 | No |
| 9 | b | d | 9.028226e-30 | 1.0 | No |

Gambar 2. Hasil uji ANOVA

Matriks Korelasi:

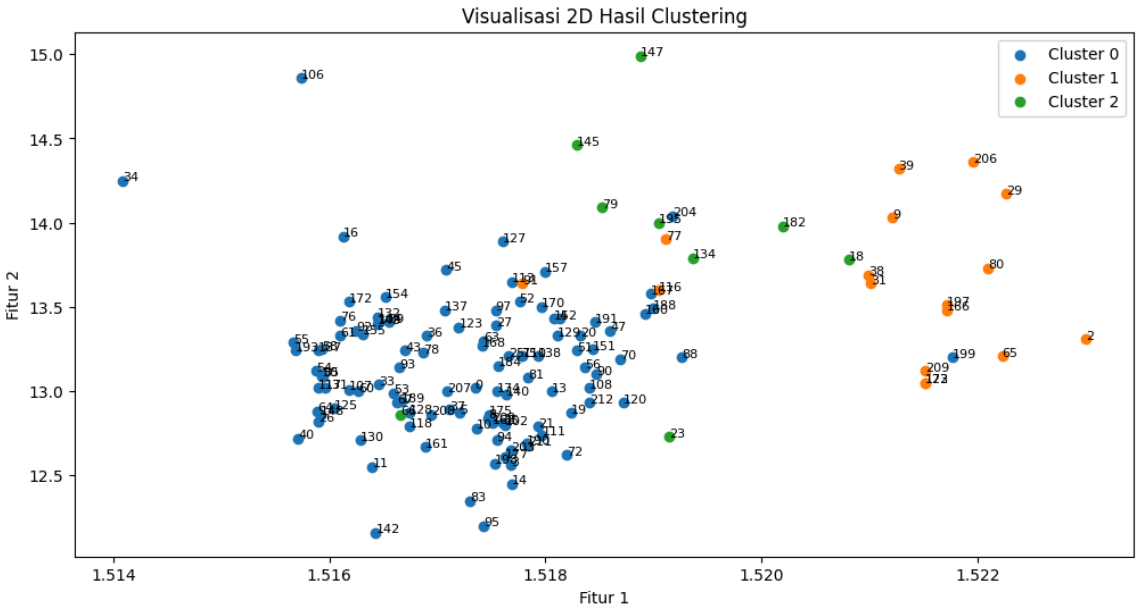
| | a | b | c | d | e | f | g | h | i |
|---|---|---|---|---|---|---|---|---|---|
| a | 1.000000 | 0.306245 | 0.041342 | -0.634328 | -0.630537 | -0.618174 | 0.762795 | NaN | 0.050620 |
| b | 0.306245 | 1.000000 | -0.120900 | -0.075339 | -0.674300 | -0.565736 | 0.143260 | NaN | -0.175159 |
| c | 0.041342 | -0.120900 | 1.000000 | -0.363735 | -0.193589 | 0.009072 | -0.470654 | NaN | -0.043090 |
| d | -0.634328 | -0.075339 | -0.363735 | 1.000000 | 0.224517 | 0.486540 | -0.445555 | NaN | -0.057493 |
| e | -0.630537 | -0.674300 | -0.193589 | 0.224517 | 1.000000 | 0.466566 | -0.362872 | NaN | 0.060168 |
| f | -0.618174 | -0.565736 | 0.009072 | 0.486540 | 0.466566 | 1.000000 | -0.542178 | NaN | 0.009638 |
| g | 0.762795 | 0.143260 | -0.470654 | -0.445555 | -0.362872 | -0.542178 | 1.000000 | NaN | 0.137881 |
| h | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| i | 0.050620 | -0.175159 | -0.043090 | -0.057493 | 0.060168 | 0.009638 | 0.137881 | NaN | 1.000000 |

Gambar 3. Hasil Uji Korelasi

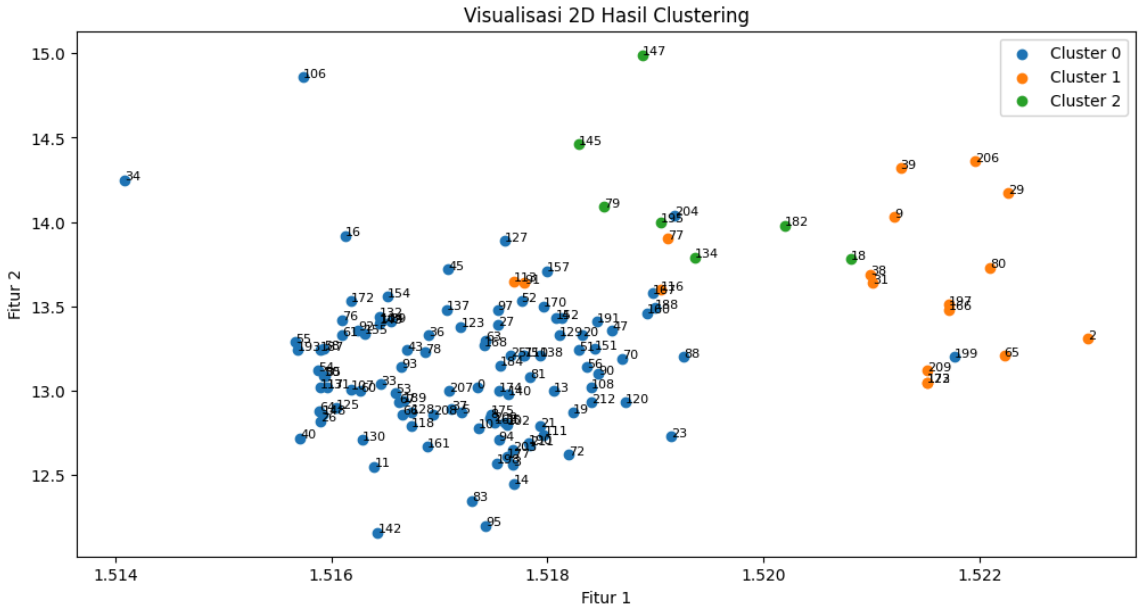
Berdasarkan analisis clustering dengan pendekatan beberapa jenis model clustering. **`KMeans, DBSCAN, Agglomerative Clustering, Gaussian Mixture Model (GMM), Mean Shift.`**

Dari hasil clustering diperoleh kesimpulan terdapat 3 kelompok formulasi berdasarkan **K Means, Agglomerative Clustering, Gaussian Mixture Model (GMM)** dan **Mean-shift** menunjukan ada 8 kelompok, sedangkan DBSCAN gagal dalam mengelompokan data diatas.

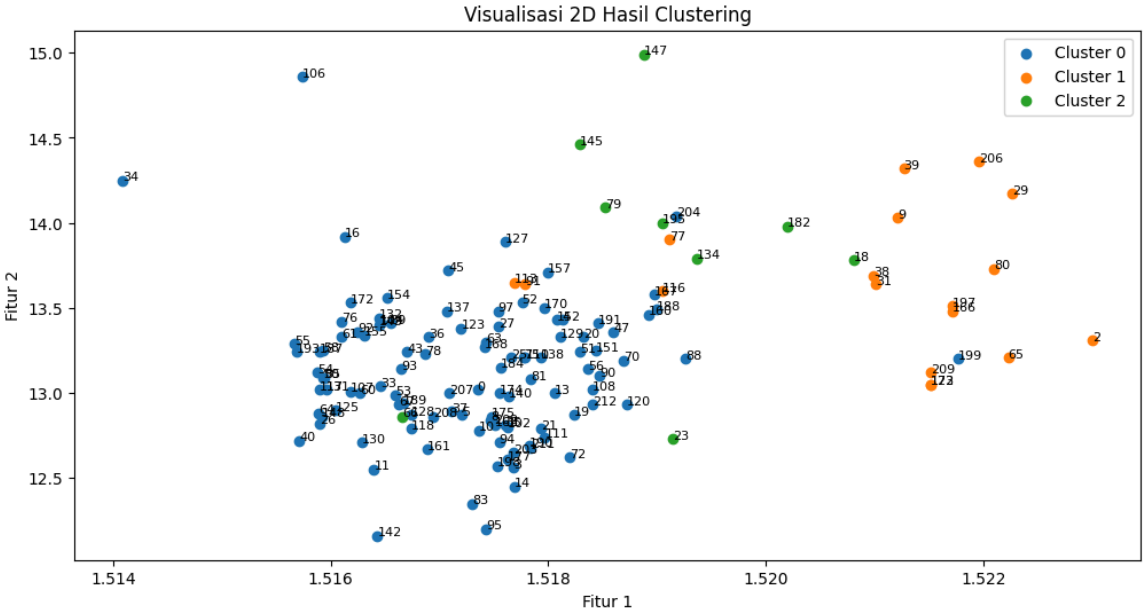
Hasil plot masing - masing cluster dimulai dari **`K Means, Agglomerative Clustering, Gaussian Mixture Model (GMM), Mean Shift, DBSCAN`** `diperlihatkan pada Gambar 4,5,6,7,8`



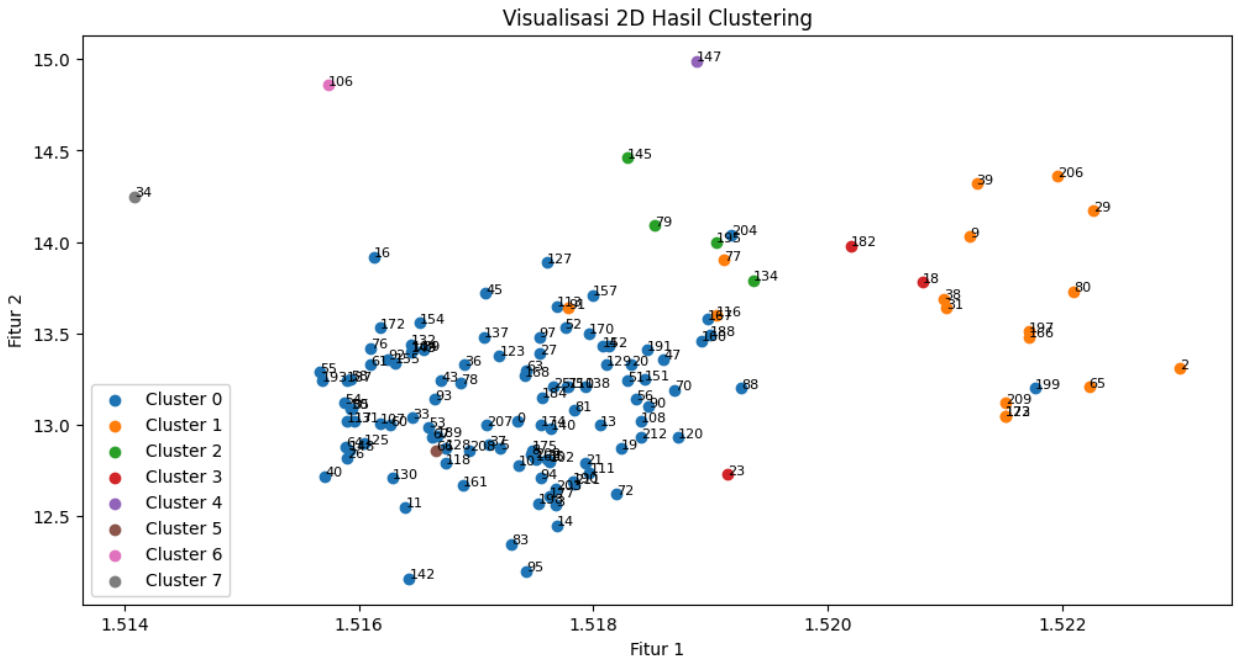

Gambar4. Hasil Model K-Means Clustering



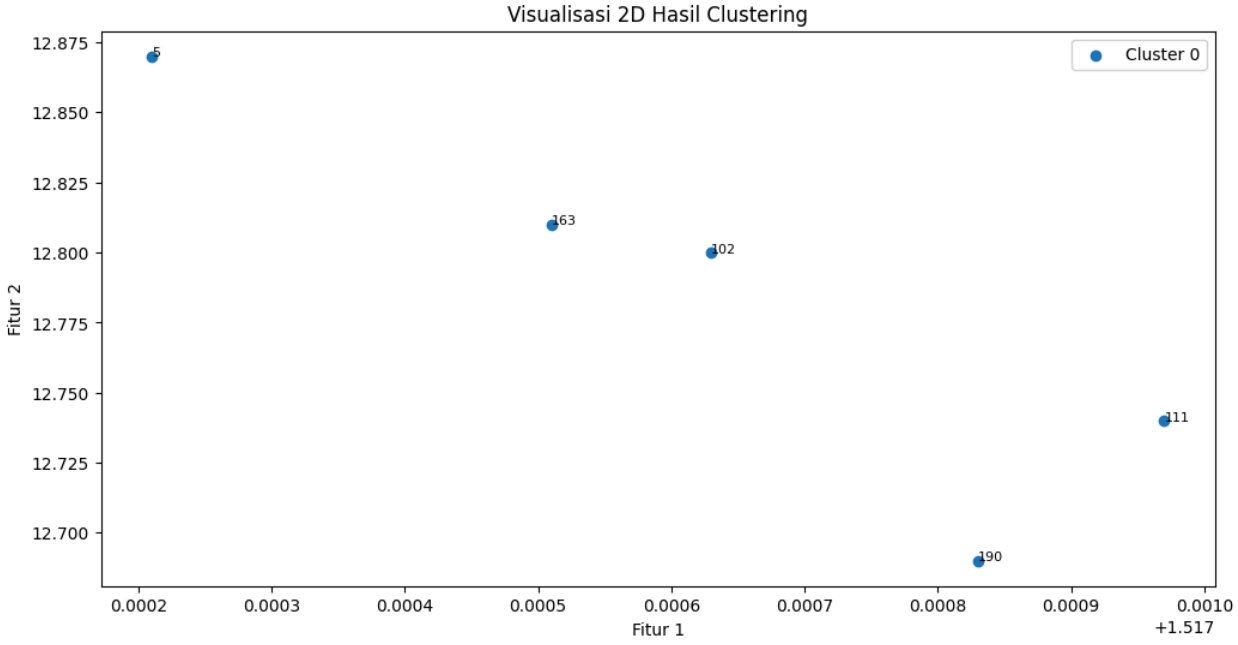

Gambar 5. Hasil Model Agglomerative Clustering




Gambar 6. Hasil Model Agglomerative Clustering




Gambar 6. Hasil Model Mean Shift




Gambar 7. Hasil Model DBSCAN

Berdasarkan hasil Evaluasi Silhouette Score masing-masing clustering diperoleh score pada rentang 0.3 hingga 0.5 yang menyatakan bahwa clustering cukup baik, dengan sebagian besar titik data berada dalam cluster yang sesuai. (catatan: kecuali DBSCAN yang gagal dalam mengelopokkan )

## 2. Solusi Pertanyaan Nomor 2

Dari Pernyataan terdapat penjelasan bahwa hasil FFB ditentukan oleh pembungaan pohon kelapa sawit dan terkait faktor eksternal. Faktor eksternal yang disajikan dalam data adalah kelembaban tanah, suhu rata-rata, minimum, maksimum, curah hujan, hari kerja, luas panen.

Pembungaan merupakan faktor kritis dalam siklus hidup kelapa sawit. Faktor eksternal yang berpengaruh terhadap pembungaan adalah curah hujan, suhu dan kelembaban suhu dan jumlah hari kerja. Dari pemahaman ini, diajukan empat hipotesis awal yaitu sebagai berikut:

**Hipotesis 1:** Terdapat korelasi positif antara curah hujan yang optimal dengan jumlah bunga yang terbentuk.
**Hipotesis 2:** rata - rata suhu atau minimum suhu mendukung pembungaan yang baik.
**Hipotesis 3:** Kelembaban tanah yang cukup mendukung pertumbuhan bunga.
**Hipotesis 4:** jumlah hari kerja berarti jumlah perhatian terhadap proses pembungaan

Berikut ini adalah langkah-langkah untuk menguji empat hipotesis yang telah disusun.

1. Penarikan data asli. Data asli dapat dilihat pada Gambar 1. Data yang ditampilkan adalah lima baris teratas.

| | Date | SoilMoisture | Average_Temp | Min_Temp | Max_Temp | Precipitation | Working_days | HA_Harvested | FFB_Yield |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 2008-01-01 | 616.4 | 25.306452 | 21.3 | 32.2 | 184.4 | 25 | 777778.3951 | 1.62 |
| 1 | 2008-02-01 | 568.9 | 26.165517 | 20.9 | 35.1 | 140.2 | 23 | 767988.2759 | 1.45 |
| 2 | 2008-03-01 | 577.6 | 25.448387 | 21.3 | 32.9 | 280.4 | 25 | 783951.9231 | 1.56 |
| 3 | 2008-04-01 | 581.1 | 26.903333 | 20.6 | 34.8 | 173.3 | 25 | 788987.0504 | 1.39 |
| 4 | 2008-05-01 | 545.4 | 27.241935 | 20.9 | 35.0 | 140.6 | 25 | 813659.7222 | 1.44 |

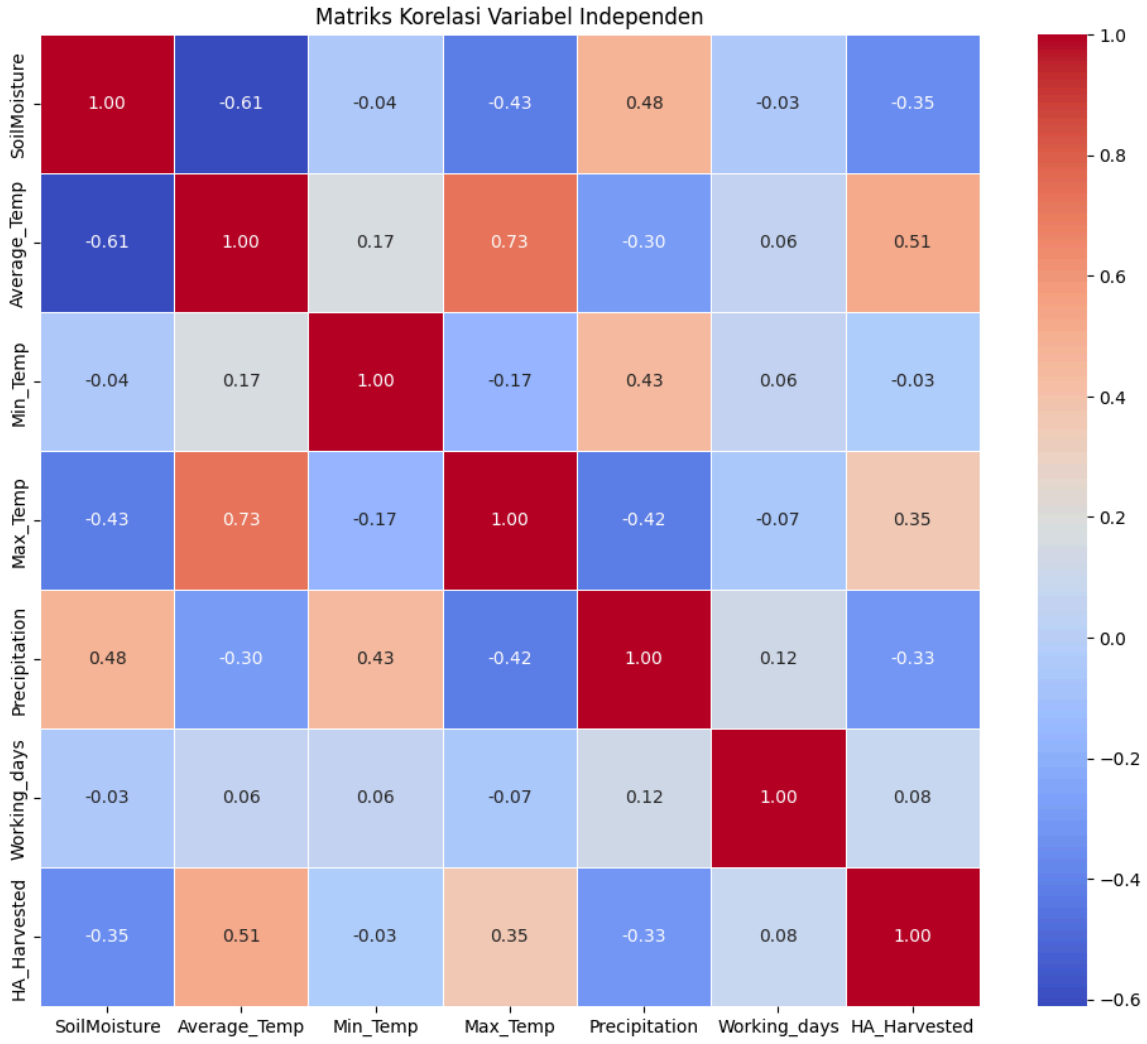
Gambar 1. Penampakan lima baris data asli teratas

2. Berdasarkan faktor eksternal yang memberikan pengaruh terhadap pembungaan maka data yang digunakan adalah data faktor eksternal berpengaruh. Data tersebut dapat diperlihatkan pada Gambar 2.

| | SoilMoisture | Average_Temp | Min_Temp | Max_Temp | Precipitation | Working_days | FFB_Yield |
|---|---|---|---|---|---|---|---|
| 0 | 616.4 | 25.306452 | 21.3 | 32.2 | 184.4 | 25.0 | 1.62 |
| 1 | 568.9 | 26.165517 | 20.9 | 35.1 | 140.2 | 23.0 | 1.45 |
| 2 | 577.6 | 25.448387 | 21.3 | 32.9 | 280.4 | 25.0 | 1.56 |
| 3 | 581.1 | 26.903333 | 20.6 | 34.8 | 173.3 | 25.0 | 1.39 |
| 4 | 545.4 | 27.241935 | 20.9 | 35.0 | 140.6 | 25.0 | 1.44 |

Gambar 2. Penampakan data eksternal berpengaruh

3. Kemudian dilakukan uji korelasi antara variabel independen.Namun,data di pre-processing dulu sebelum uji korelasi dan uji multikolinier, Visualisasi data uji korelasi dapat diperlihatkan pada gambar 3. Secara singkat terlihat adanya korelasi cukup signifikan antara suhu rata-rata dengan suhu maksimum. Sehingga data suhu maksimum kita eliminasi.




Gambar 3. Visualisasi korelasi variabel independen

Sedangkan hasil uji multikolinear diperlihatkan pada gambar 4.

| | Variable | VIF |
|---|---|---|
| 0 | const | 5974.672667 |
| 1 | SoilMoisture | 1.992389 |
| 2 | Average_Temp | 3.451316 |
| 3 | Min_Temp | 1.496976 |
| 4 | Max_Temp | 2.772080 |
| 5 | Precipitation | 1.803813 |
| 6 | Working_days | 1.053250 |

Gambar 4. Hasil Uji Multicollinear

Dari kedua uji disimpulkan sementara bahwa ada satu data yakni max temp memiliki korelasi dengan Average Temp (Threshold > 0.7) dan setiap variabel independen tidak terdiagnosa multicollinear (< 5 )

Sehingga data yang akan dianalisa adalah data yang diperlihatkan pada Gambar 5.

| | SoilMoisture | Average_Temp | Min_Temp | Precipitation | Working_days | FFB_Yield |
|---|---|---|---|---|---|---|
| 0 | 616.4 | 25.306452 | 21.3 | 184.4 | 25.0 | 1.62 |
| 1 | 568.9 | 26.165517 | 20.9 | 140.2 | 23.0 | 1.45 |
| 2 | 577.6 | 25.448387 | 21.3 | 280.4 | 25.0 | 1.56 |
| 3 | 581.1 | 26.903333 | 20.6 | 173.3 | 25.0 | 1.39 |
| 4 | 545.4 | 27.241935 | 20.9 | 140.6 | 25.0 | 1.44 |

Gambar 5. Data setelah melalui uji korelasi dan uji multikolinear

Setelah melalui beberapa tahapan, maka data siap untuk dilakukan dengan pendekatan linear regresi berganda. Hasilnya diperlihatkan pada Tabel 1:

Tabel 1: Hasil pemrosesan data dengan pendekatan model linear regresi berganda

| Variable | Coefficient | Standard Error | t-Value | p-Value |
|---|---|---|---|---|
| const | 1.6150 | 0.0230 | 70.2245 | 0.0000 |
| x1 | -0.0898 | 0.0321 | -2.7938 | 0.0061 |
| x2 | -0.0070 | 0.0297 | -0.2342 | 0.8153 |
| x3 | -0.0565 | 0.0267 | -2.1124 | 0.0368 |
| x4 | 0.1649 | 0.0305 | 5.4168 | 0.0000 |
| x5 | 0.0111 | 0.0234 | 0.4742 | 0.6363 |
| R-squared | 0.2169 | | | |
| Adjusted R-squared | 0.1831 | | | |

Kesimpulan sementara yang dapat diambil dari informasi Tabel 1 sebagai berikut:

1. **Hipotesis 1 diterima**: Curah hujan yang optimal memiliki korelasi positif yang signifikan dengan jumlah bunga yang terbentuk. Dengan alasan karena koefisien dari curah hujan (Precipitation) adalah positif dan p-value-nya sangat kecil. Ini berarti ada hubungan positif yang signifikan antara curah hujan dan jumlah bunga yang terbentuk.

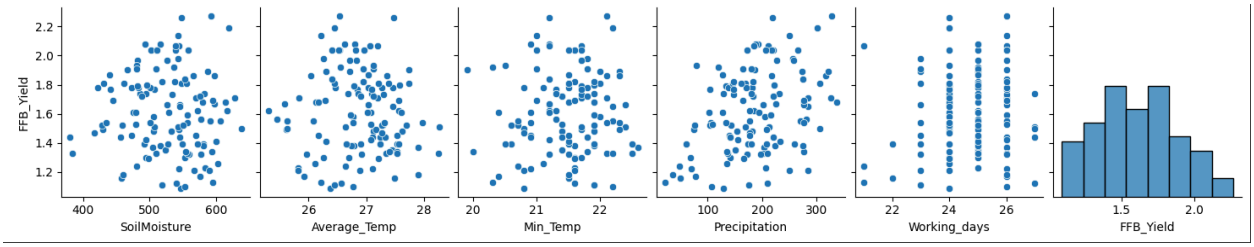
2. **Hipotesis 2 diterima:** dalam konteks pengaruh suhu minimum terhadap pembungaan, meskipun suhu rata-rata tidak berkontribusi secara signifikan.

3. **Hipotesis 3 diterima:** Hipotesis ini diterima karena nilai p untuk SoilMoisture kurang dari 0.05, yang menunjukkan bahwa kelembaban tanah memiliki pengaruh signifikan terhadap pertumbuhan bunga. Namun, koefisien negatif menunjukkan bahwa kelembaban tanah yang lebih tinggi dapat mengurangi pertumbuhan bunga, mungkin karena kelembaban berlebih dapat menghambat pertumbuhan pada titik tertentu.

4. **Hipotesis 4 ditolak:** Nilai p lebih dari 0.05 berarti jumlah hari kerja tidak berpengaruh signifikan terhadap jumlah bunga. Jadi, tidak ada bukti bahwa jumlah hari kerja mempengaruhi proses pembungaan.

**Tambahan:**
Jika memperhatikan nilai R-squared dan Adjusted R-squared dapat disimpulkan bahwa. Model belum terlalu baik dalam melakukan analisis hubungan faktor eksternal dan pembungaan. Ada kemungkinan bahwa model tidak mendapatkan faktor eksternal yang sangat signifikan terhadap pembungaan yang akhirnya mempengaruhi nilai FFB. Dengan kata lain, model mungkin perlu ditingkatkan dengan menambahkan variabel baru atau menggunakan metode yang berbeda untuk mendapatkan pemahaman yang lebih baik mengenai faktor-faktor yang mempengaruhi pembungaan.

**Analisa Tambahan:**
Berdasarkan plotting hubungan antara variabel dependen dengan variabel independen yang diperlihatkan pada Gambar 6.




Gambar 6. hubungan antara variabel dependen dengan variabel independen

Tidak memiliki hubungan linear yang bagus dalam artian datanya pola acak. Sehingga penulis mencoba untuk melakukan pendekatan Non-Linear. Hasilnya sebagai berikut:

Berikut adalah ringkasan hasil model diperlihatkan pada Gambar 7:

| Model | MSE | $R^2$ | Kesimpulan |
|---|---|---|---|
| Polinomial Regression | 0.128 | -0.735 | Tidak efektif, kemungkinan overfitting. |
| Decision Tree Regressor | 0.146 | -0.975 | Kurang baik, mungkin overfitting atau terlalu rumit. |
| Random Forest Regressor | 0.075 | -0.021 | Terbaik di antara model, tetapi masih kurang memadai. |

Gambar 7. ringkasan hasil model

Disini penulisan menyimpulkan bahwa ada beberapa faktor eksternal yang dapat ditinjau sebagai pengaruh pembungaan yang berefek pada FFB yield. Namun, dengan mempertimbangkan confidential informasi atau varian mungkin perlu ditambahkan untuk memberikan informasi lebih sehingga model mengenali pola hubungan variabel independen terhadap variabel dependen lebih baik.

**3. Solusi Pertanyaan Nomor 3**

A. jawaban untuk pertanyaan 1 What is the probability of the word “data” occurring in each line ?

Line 1 probability: 0.0345
Line 2 probability: 0.0541
Line 3 probability: 0.0370
Line 4 probability: 0.0238
Line 5 probability: 0.0385
Line 6 probability: 0.0667
Line 7 probability: 0.0952
Line 8 probability: 0.0000
Line 9 probability: 0.1250
Line 10 probability: 0.2143

Line 11 probability: 0.0588
Line 12 probability: 0.0333

B. jawaban untuk pertanyaan 2 What is the distribution of distinct word counts across all the lines ?

Distinct word count: 26, Frequency: 4
Distinct word count: 29, Frequency: 1
Distinct word count: 40, Frequency: 1
Distinct word count: 15, Frequency: 1
Distinct word count: 33, Frequency: 1
Distinct word count: 8, Frequency: 1
Distinct word count: 11, Frequency: 1
Distinct word count: 17, Frequency: 1
Distinct word count: 27, Frequency: 1

C. jawaban untuk pertanyaan 3 What is the probability of the word “analytics” occurring after the word “data” ?

probability : 0.3333333333333333 data_analytic_count : 6 data_count : 18