Assessment Test Solution

Candidate: Ramma Hayu Fitra Saleh

Update: 11 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. Uji normalitas digunakan untuk mengetahui apakah data memenuhi distribusi normal atau tidak. Berdasarkan uji normalitas pada data zat/formulasi a, b,c d,e,f,g,h diperoleh bahwa formulasi d dan formulasi h memenuhi distribusi normal sedangkan formulasi lainya tidak memenuhi distribusi normal berdasarkan hasil uji normalitas shapiro-wilk.

3. Dari hasil uji normalitas kita mendapatkan pandangan langkah lanjutan yang dapat kita lakukan yaitu melakukan uji korelasi. Uji korelasi dilakukan dengan

mempertimbangkan apakah data berdistribusi normal atau tidak. Jika data berdistribusi normal maka dapat dilakukan uji korelasi dengan pendekatan **Pearson** sedangkan data yang tidak berdistribusi normal akan dilakukan uji korelasi dengan pendekatan **Spearman/Kendall.** Hasil dari uji korelasi diperlihatkan pada Gambar 2.

Correlation Matrix:

	a	b	c	d	e	f	g	h	i
a	1.000000	0.031040	0.144156	-0.407326	-0.525733	-0.288001	0.703777	-0.000386	0.096181
b	0.031040	1.000000	-0.126451	0.156794	-0.265643	-0.584503	0.027205	0.326603	-0.217631
c	0.144156	-0.126451	1.000000	-0.481799	-0.336811	0.200742	-0.289119	-0.492262	0.095487
d	-0.491821	0.135910	-0.512420	1.000000	0.196513	0.153438	-0.280952	0.479404	-0.076313
e	-0.525733	-0.265643	-0.336811	-0.005524	1.000000	-0.000719	-0.221912	-0.102151	-0.071995
f	-0.288001	-0.584503	0.200742	0.325958	-0.000719	1.000000	-0.472703	-0.042618	0.091903
g	0.703777	0.027205	-0.289119	-0.259592	-0.221912	-0.472703	1.000000	-0.112841	0.111897
h	-0.181511	0.411111	-0.456107	0.479404	0.170212	-0.260406	-0.007770	1.000000	0.009680
i	0.096181	-0.217631	0.095487	-0.074402	-0.071995	0.091903	0.111897	-0.058692	1.000000

Gambar 2. Hasil Uji Korelasi

Analisa uji korelasi antar formulasi dapat dilihat pada Gambar 3.

Pasangan Variabel	Korelasi	Interpretasi
a - g	0.703777	Hubungan positif yang kuat
a - e	-0.525733	Hubungan negatif yang kuat
d - c	-0.481799	Hubungan negatif yang moderat
f - d	0.325958	Hubungan positif yang moderat
h - d	0.479404	Hubungan positif yang moderat
c - e	-0.336811	Hubungan negatif yang moderat
a - c	0.144156	Hubungan positif yang lemah
c - f	0.200742	Hubungan positif yang lemah
i - g	0.111897	Hubungan positif yang sangat lemah
b - h	0.411111	Hubungan positif yang moderat
e - h	0.170212	Hubungan positif yang lemah
d - e	0.196513	Hubungan positif yang lemah
a - b	0.03104	Hubungan positif yang sangat lemah
b - g	0.027205	Hubungan positif yang sangat lemah
h - f	-0.260406	Hubungan negatif yang lemah
a - f	-0.288001	Hubungan negatif yang lemah
c - g	-0.289119	Hubungan negatif yang lemah
d - g	-0.280952	Hubungan negatif yang lemah
f - g	-0.472703	Hubungan negatif yang moderat
i - f	0.091903	Hubungan positif yang sangat lemah
b - f	-0.584503	Hubungan negatif yang kuat
h - c	-0.456107	Hubungan negatif yang moderat
b - c	-0.126451	Hubungan negatif yang sangat lemah
d - a	-0.491821	Hubungan negatif yang kuat
i - a	0.096181	Hubungan positif yang sangat lemah
i - b	-0.217631	Hubungan negatif yang lemah
i - d	-0.074402	Hubungan negatif yang sangat lemah
i - e	-0.071995	Hubungan negatif yang sangat lemah

Gambar 3. Analisa uji korelasi antara formulasi

4. Langkah selanjutnya adalah melakukan uji ANOVA.
   Uji ANOVA hanya berlaku pada data yang dianggap berdistribusi normal dan memiliki varian homogen Berdasarkan uji normalitas formulasi b dan h dianggap berdistribusi normal. Namun, berdasarkan uji varian homogen dianggap sebagai formulasi yang tidak memenuhi varian homogen. Oleh karena itu kita dapat melakukan pendekatan uji Welch ANOVA dimana uji ini dapat digunakan untuk data yang dianggap berdistribusi normal.Namun, tidak homogen. Hasil dapat diperlihatkan pada tabel 1.

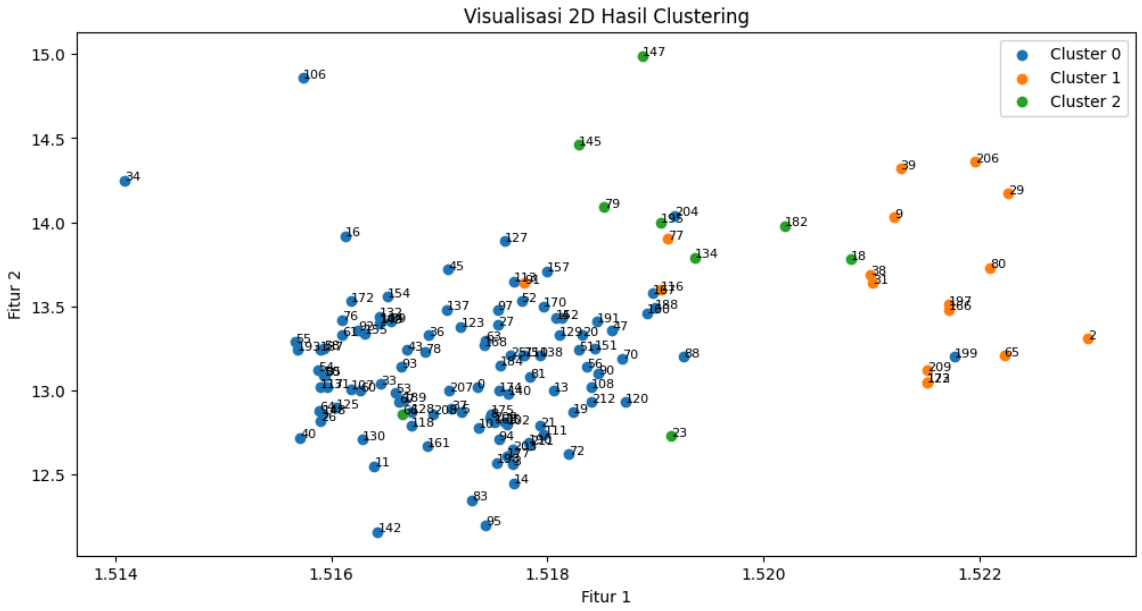
Uji Statistik	Statistik	P-value	Interpretasi
Levene's Test	179.6	9.66×10^{-32} (sangat kecil)	Varians antara data b dan h tidak homogen.
Welch's T-test	-3.22×10^{-15}	1.0 (sangat besar)	Tidak ada perbedaan signifikan antara rata-rata b dan h.

   Berdasarkan hasil yang diperlihatkan pada tabel 1. Disimpulkan bahwa formulasi b dan h adalah formulasi yang sama berdasarkan rata - rata nilai.
5. Untuk data yang dianggap tidak berdistribusi normal berdasarkan uji normalitas akan dilakukan pendekatan uji non-parametric. Uji non-parametric yang akan digunakan adalah uji **Kruskal-Wallis.** Hasil dari uji tersebut sebagai berikut:
   Test Statistic: 21.513 dan **P-value**: 0.00148 yang bermakna adalah hipotesis nol bahwa semua kelompok memiliki distribusi yang sama dapat ditolak dalam bahasa sederhana ada beberapa formulasi yang memiliki perbedaan. Untuk mengetahui formulasi mana yang memiliki perbedaan dapat dilakukan uji Post-Hoc. Berdasarkan uji Post-Hoc disimpulkan bahwa formulasi **a, c, h, dan i** memiliki perbedaan secara nilai rata - rata dan formulasi **b,d,e,f,g** memiliki kemiripan secara nilai rata - rata.

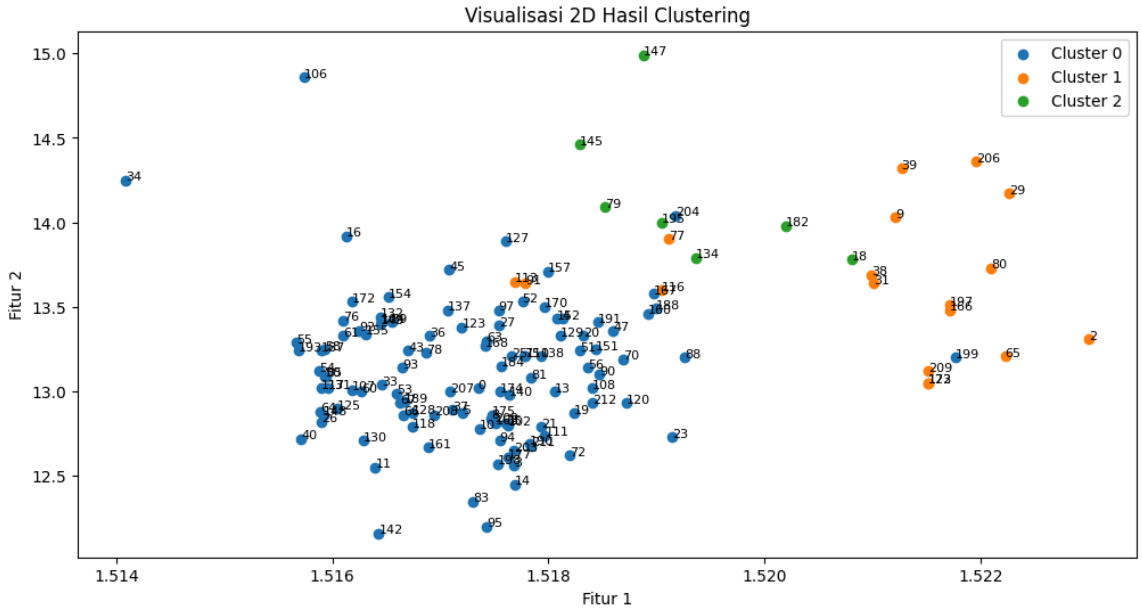
Berdasarkan analisis clustering dengan pendekatan beberapa jenis model clustering. **K Means, DBSCAN, Agglomerative Clustering, Gaussian Mixture Model (GMM), Mean Shift.**

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.

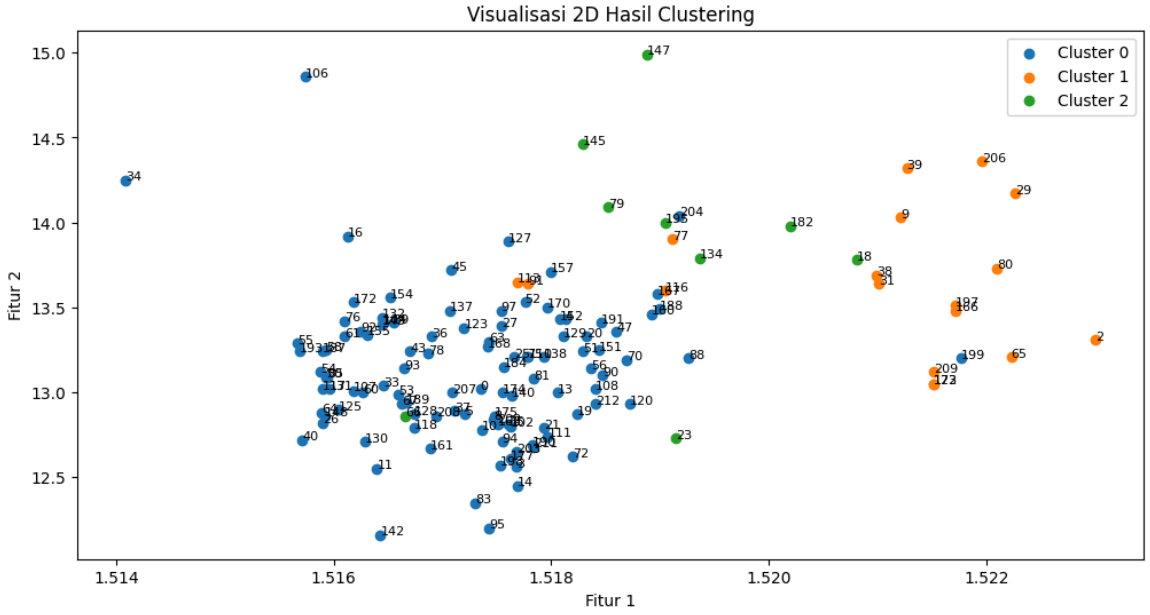
Hasil plot masing - masing cluster dimulai dari **K Means, Agglomerative Clustering, Gaussian Mixture Model (GMM), Mean Shift, DBSCAN** diperlihatkan pada Gambar 4,5,6,7,8



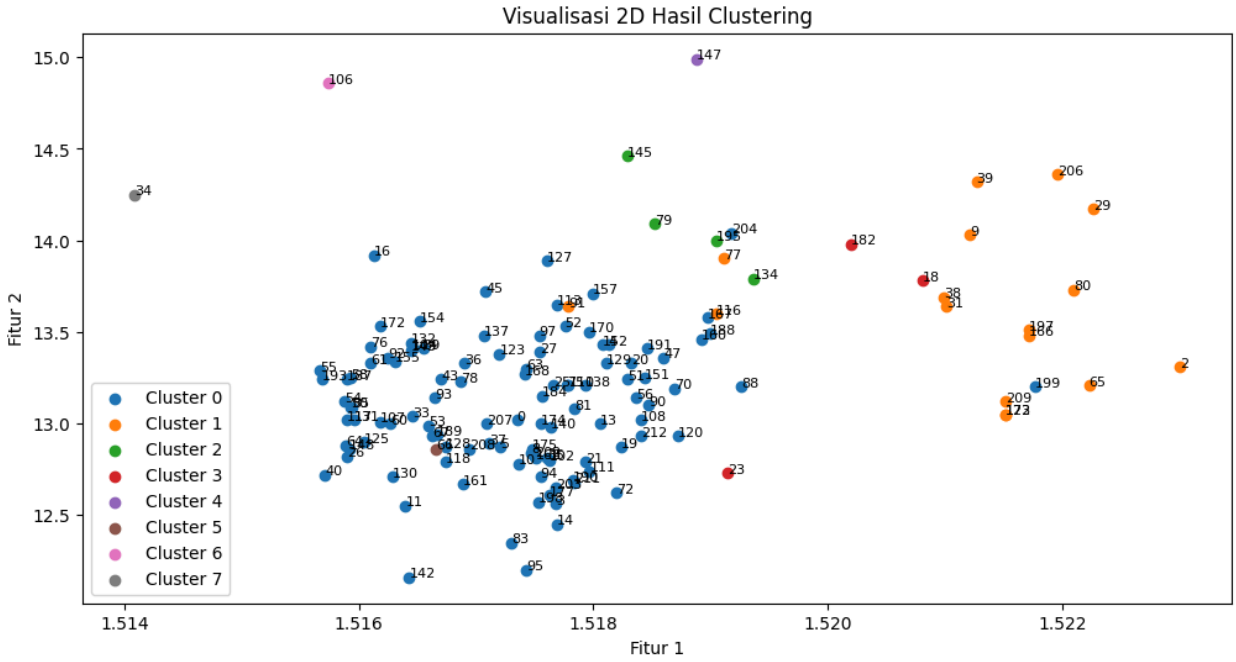
Gambar4. Hasil Model K-Means Clustering



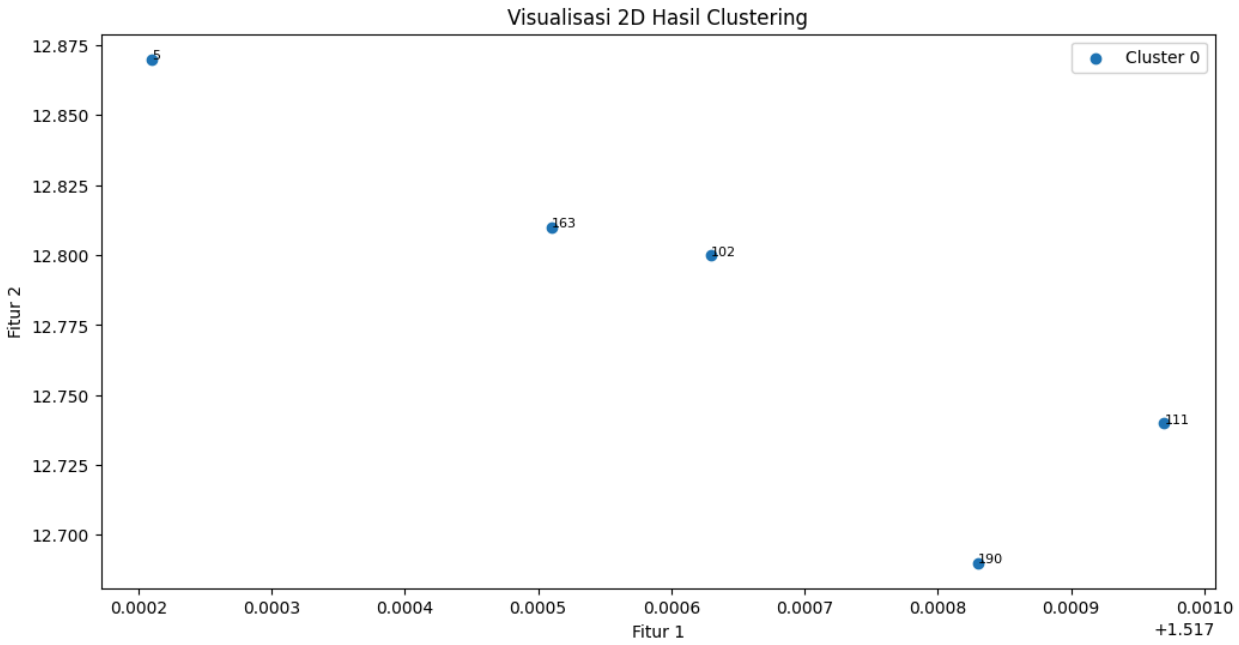
Gambar 5. Hasil Model Agglomerative Clustering



Gambar 6. Gaussian Mixture Model (GMM)



Gambar 7. Hasil Model Mean Shift



Gambar 8. 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 mengelompokkan)

Kajian lebih mendalam terkait hubungan antara formulasi dari hasil uji korelasi, uji parametric/non-parametric dan clustering perlu dievaluasi secara bersamaan untuk mendapatkan pemahaman yang lebih lengkap tentang struktur dan hubungan dalam data.

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,minimum suhu dan suhu maksimum 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. Setelah melakukan preprocessing data, uji kualitatif, uji korelasi maka data ter-reduksi, diperlihatkan pada Gambar 1.

	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 1. Data setelah ter-reduksi

Informasi Max Temp dihapus karena dianggap memiliki korelasi terhadap Average Temp. HA Harvested dihapus karena dianggap tidak mempengaruhi pembungaan.

Data diatas dianalisa dengan menggunakan pendekatan regresi linear berganda. Hasil sebagai berikut:

Model Summary:

Variable	Coefficient	Standard Error	t-Value	p-Value
const	1.6126	0.0232	69.5362	0.0000
x1	-0.0893	0.0324	-2.7598	0.0067
x2	-0.0074	0.0301	-0.2471	0.8052
x3	-0.0537	0.0278	-1.9363	0.0553
x4	0.1649	0.0314	5.2446	0.0000
x5	0.0109	0.0235	0.4645	0.6431
R-squared	0.2102			
Adjusted R-squared	0.1759			

Kesimpulan sementara yang dapat diambil dari informasi Tabel 1 sebagai berikut:
Keterangan : x1, x2, x3, x4 dan x5 berturut turut adalah SoilMoisture Average_Temp, Min_Temp,Precipitation Working_days

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.

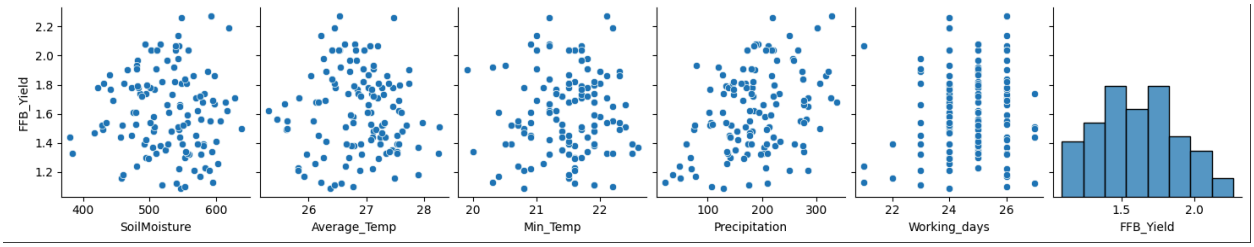
2. **Hipotesis 2 diterima:** Jika melihat hasil dari p-value Average_Temp hipotesis ditolak.Namun jika melihat hasil dari p-value Min Temp hipotesis diterima. Dalam artian Min Temp memberikan pengaruh terhadap pembungaan yang berlanjut pada hasil FFB Yield.

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 pembungaan. 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