

BAB III

METODE PENELITIAN

Pada bagian metode penelitian ini, akan dijelaskan secara rinci mengenai langkah-langkah yang akan diambil dalam pelaksanaan penelitian, mulai dari perencanaan hingga pengumpulan dan analisis data.

3.1 Subjek dan Objek Penelitian

Subjek pada penelitian ini adalah petani dan pemangku kebijakan yang berkaitan erat dengan bidang pertanian dan siapa saja yang ingin melakukan kegiatan pertanian karena pada penelitian ini menghasilkan output berupa rekomendasi tanaman pangan. Lalu untuk objek penelitian ini adalah tanaman pangan yang akan ditanam oleh subjek berdasar hasil rekomendasi sistem yang telah dibuat dengan kecerdasan buatan yaitu klasifikasi dan regresi

3.2 Alat dan Bahan Penelitian

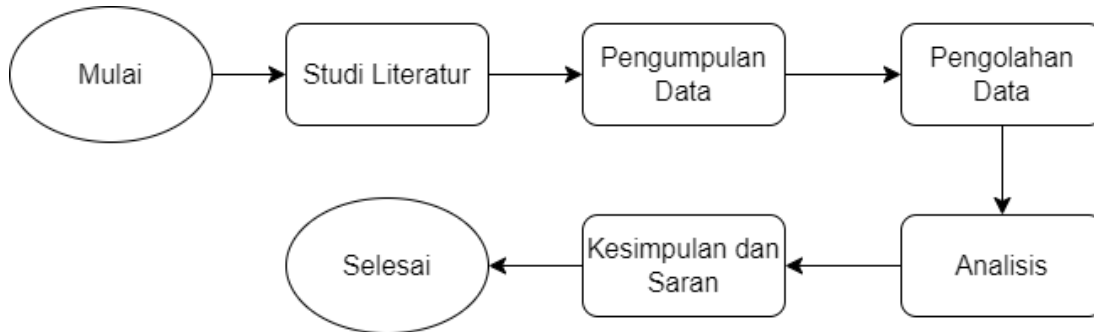
Untuk mendukung proses pembuatan sistem pada penelitian ini, maka dibutuhkan alat dan bahan guna menunjang pembuatan model kecerdasan buatan penataan potensi pertanian secara otomatis. Alat dan bahan yang digunakan dalam penelitian ini dijabarkan pada tabel berikut

Tabel 3. 1 Alat dan Bahan

No	Nama Alat & Bahan	Jumlah
1.	Google Colab	1
2.	Laptop	1
3.	Data Iklim Jawa Timur	252
4.	Data Pertanian Jawa Timur	252
5.	Perangkat lunak pemrograman <i>Python</i>	1
6.	Modul <i>Scikit-learn</i>	1
7.	<i>Pandas</i> dan <i>NumPy</i>	1

3.3 Diagram Alir Penelitian

Dalam suatu penelitian tentunya harus memiliki suatu tahapan proses agar penelitian dapat terarah dan mencapai tujuan. Tahapan proses tersebut bisa dilihat pada gambar 3.1



Gambar 3. 1 Diagram Alir Penelitian

Dalam diagram alir penelitian diatas, tahapan penelitian dilakukan sebagai berikut

3.3.1 Studi Literatur

Langkah pertama yang dilakukan pada penelitian ini adalah melakukan pencarian studi literatur yang terkait dengan penelitian ini, yaitu tentang metode klasifikasi dan juga regresi guna mendukung penelitian.

3.3.2 Pengumpulan Data

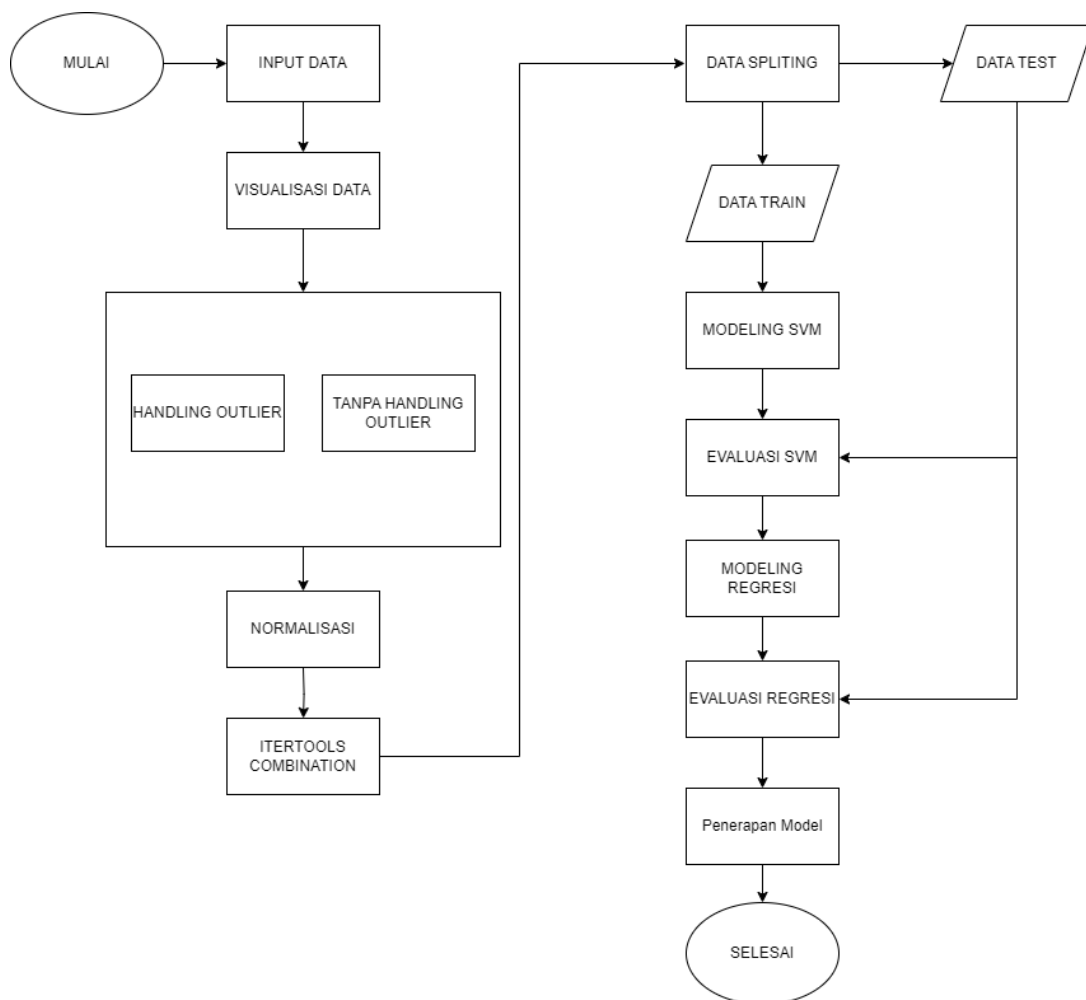
Pengumpulan data adalah proses mengumpulkan dan mengukur informasi tentang variabel-variabel yang diteliti, dengan cara sistematis yang memungkinkan penulis menjawab pertanyaan penelitian yang dinyatakan, menguji hipotesis, dan mengevaluasi hasil. Pengumpulan data pada penelitian ini merujuk pada beberapa sumber, seperti BPS. Wilayah yang digunakan untuk dijadikan *dataset* adalah 7 kecamatan, dimana setiap kecamatan memiliki beberapa parameter yang dapat dilihat pada tabel 3.2

Tabel 3.2 Parameter Pada Dataset

No	Atribut	Inisialisasi	Keterangan
1.	Tahun	X1	Range tahun antara 2004 - 2021
2.	Daerah	X2	Pengambilan daerah berdasarkan data yang diambil
5.	Curah Hujan (mm)	X3	Data curah hujan didapatkan dari data citra satelit FLDAS diambil dari rentang waktu 2004 – 2021
6.	Kelembapan (%)	X4	Data kelembapan didapatkan dari data citra satelit FLDAS diambil dari rentang waktu 2004 – 2021
7.	Temperatur (°C)	X5	Data temperatur didapatkan dari data citra satelit FLDAS diambil dari rentang waktu 2004 – 2021
8.	Soil Temperatur (°C)	X6	Data Soil Temperatur didapatkan dari data citra satelit FLDAS diambil dari rentang waktu 2004 - 2021
7.	Soil Moisture (%)	X7	Data Soil Moisture didapatkan dari data citra satelit FLDAS diambil dari rentang waktu 2004 - 2021
8.	Produksi (Ton)	X8	Data hasil panen didapatkan dari BPS diambil dari rentang waktu 2004 – 2021
9.	Luas (Hektar)	X9	Data luas panen didapatkan dari BPS diambil dari rentang waktu 2004 – 2021
10.	Tanaman	Y1	Output dari Klasifikasi
11.	Produktivitas (ton/hektar)	Y2	Output dari regresi

3.3.3 Pengolahan Data

Setelah data terkumpul, maka langkah selanjutnya adalah melakukan pengolahan data untuk mendapatkan informasi yang bermanfaat. Dalam konteks kajian tentang jenis tanaman dan produksi tanaman, pengolahan data bertujuan untuk melakukan pemodelan klasifikasi jenis tanaman dan memprediksi produksi tanaman yang akan datang. Proses ini melibatkan beberapa tahapan, antara lain preprocessing data, pemilihan fitur, pemilihan model, dan evaluasi model.



Gambar 3.2 Diagram alir modeling klasifikasi dan regresi

Berdasarkan gambar diagram alir yang disajikan, proses modeling akan menggunakan dua skema, yaitu modeling tanpa data outlier dan modeling dengan menggunakan data outlier, lalu akan dibandingkan model mana yang memiliki akurasi terbaik.

1. Input Data

Proses penginputan data yang dilakukan dari hasil pencarian data pertanian yang bersumber dari BPS dan juga data iklim yang diambil dari citra satelit FLDAS. Data kemudian dijadikan satu dan disimpan dalam format Comma Separated Values (CSV). Data yang sudah diperoleh bisa dilihat pada tabel di bawah

Tabel 3. 3 Sampel Data

NO	TAHUN	DAERAH	CURAH HUJAN	KELEMBAPAN	TEMPREATUR	SOIL TEMPERATURE	SOIL MOIS	PRODUKSI	LUAS	PRODUKTIVITAS	TANAMAN
0	2004	SURABAYA	316.958649	83.825798	27.187769	28.646875	41.700382	6304	1268	4.97	PADI
1	2005	SURABAYA	154.184189	79.881996	27.18844	29.536764	41.12318	6907	1387	4.98	PADI
2	2006	SURABAYA	365.939545	85.098007	26.650049	27.781427	41.837715	7297	1422	5.13	PADI
3	2007	SURABAYA	190.664093	78.235806	28.168116	29.536764	41.174019	8996	1602	5.62	PADI
4	2008	SURABAYA	242.604889	80.93589	26.8901	28.832514	41.393124	8495	1505	5.64	PADI
...
121	2017	MAGETAN	385.950714	85.628159	24.626459	23.806452	43.043476	26033.99875	1125	32.31	UBI JALAR
122	2018	MAGETAN	412.503418	84.734398	24.656092	23.669641	43.184547	26033.99875	1124	35.64	UBI JALAR
123	2019	MAGETAN	409.979156	83.707924	25.310572	24.388544	42.657726	26033.99875	1161	37.1	UBI JALAR
124	2020	MAGETAN	257.67099	81.39325	25.718409	25.080347	40.74062	26033.99875	1200	29.26	UBI JALAR
125	2021	MAGETAN	511.505402	87.200127	24.523523	23.720179	43.559612	26033.99875	1123	35.3	UBI JALAR

2. Visualisasi Data

Visualisasi data merupakan salah satu tahapan penting dalam metode penelitian. Dalam tahap ini, data diubah menjadi bentuk grafik atau plot yang mudah dipahami. Tujuan dari visualisasi data adalah untuk membantu memahami pola, hubungan, dan tren yang mungkin sulit terlihat hanya dengan melihat data mentah [32].

Penelitian ini akan menggunakan berbagai jenis visualisasi data untuk menganalisis hubungan antara variabel dalam dataset yang telah dikumpulkan. Salah satu jenis visualisasi yang akan digunakan adalah *box plot*, yang berguna untuk menampilkan distribusi data dan mendeteksi adanya *outlier* pada setiap variabel [33]. Nantinya data yang mengandung *outlier* akan dilakukan *handling outlier*.

Penelitian ini juga akan menggunakan scatter plot untuk mengetahui apakah hubungan antara variabel-variabel dalam dataset bersifat linear atau tidak. *Scatter plot* dapat membantu peneliti untuk memvisualisasikan hubungan antara dua variabel dan menentukan apakah hubungan tersebut bersifat positif atau negatif. Selain itu, scatter

plot juga dapat membantu dalam mengidentifikasi *outlier* dan pola lain yang tidak terlihat dengan visualisasi data yang lain [34].

Terakhir, penelitian ini juga akan menggunakan *heatmap* untuk mengetahui hubungan antar variabel dalam dataset. *Heatmap* dapat membantu peneliti untuk memvisualisasikan pola hubungan antar variabel secara lebih jelas dan mudah dibaca. Dengan *heatmap*, peneliti dapat mengidentifikasi korelasi antara variabel dan menemukan pola hubungan yang kompleks di antara variabel [35].

3. Handle Outlier

Skema yang akan dilakukan pada penelitian ini akan menggunakan dua skenario modeling, yaitu modeling dengan menggunakan data yang sudah dibersihkan *outlier* dan juga data yang belum dibersihkan *outlier*. Hal ini bermaksud untuk membandingkan akurasi mana yang lebih bagus antara yang sudah dibersihkan dan yang belum dibersihkan dengan *outlier*.

Data yang sudah dikumpulkan nantinya akan dilakukan pengecekan, seperti data null dan menangani data yang memiliki outlier. Nantinya jika data memiliki outlier, maka akan dilakukan penanganan outlier menggunakan metode *Interquartile Range* (IQR). Metode IQR adalah salah satu teknik yang umum digunakan untuk menangani outlier pada data numerik [36]. Teknik ini menggunakan nilai IQR untuk mengidentifikasi nilai-nilai ekstrim dalam data. IQR adalah selisih antara kuartil atas dan kuartil bawah dalam distribusi data. Persamaan perhitungan IQR adalah sebagai berikut:

$$IQR = Q3 - Q1 \quad (3.1)$$

Di mana Q1 adalah nilai kuartil pertama (25% terbawah) dan Q3 adalah nilai kuartil ketiga (25% tertinggi) dalam distribusi data. IQR mengukur jarak antara kuartil atas dan kuartil bawah dan digunakan untuk mengidentifikasi nilai-nilai ekstrim (outlier) dalam data numerik.

4. Normalisasi

Normalisasi berfungsi untuk merubah nilai-nilai dalam sebuah dataset ke dalam skala yang seragam untuk meningkatkan kinerja model dan akurasi prediksi. Normalisasi biasanya dilakukan pada data numerik yang berbeda-beda skala atau rentang nilainya [37]. Proses normalisasi pada penelitian ini menggunakan *min-max scaling* dengan persamaan

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (3.1)$$

Di mana :

X_{scaled} Nilai skala yang dihasilkan setelah melakukan normalisasi pada variabel x

X = Nilai awal dari variabel x

X_{min} = Nilai minimum dari variabel x

X_{max} = Nilai maksimum dari variabel x

Setiap nilai dalam variabel akan ditransformasi menjadi nilai antara 0 dan 1, sehingga mempermudah interpretasi hasil dan penggunaan pada model machine learning. Data yang sudah dinormalisasi bisa dilihat pada tabel di bawah.

Tabel 3. 4 Tabel Normalisasi Data

NO	TAHUN	DAERAH	CURAH HUJAN	KELEMBAPAN	TEMPREATUR	SOIL TEMPERATURE	SOIL MOIS	PRODUKSI	LUAS	PRODUKTIVITAS	TANAMAN
0	0	0.833333	0.515202	0.531441	0.738532	0.859473	0.772782	0.241956	0.4442	0.125701	0.666667
1	0.058824	0.833333	0.175028	0.156503	0.738691	1	0.702241	0.265123	0.486035	0.125974	0.666667
2	0.117647	0.833333	0.617565	0.65239	0.611297	0.722806	0.789565	0.280108	0.498339	0.130055	0.666667
3	0.176471	0.833333	0.251265	0	0.9705	1	0.708454	0.345385	0.561618	0.143387	0.666667
4	0.235294	0.833333	0.359814	0.256697	0.668098	0.888788	0.735231	0.326136	0.527518	0.143931	0.666667
...
121	0.764706	0.333333	0.659385	0.702792	0.132477	0.095097	0.936922	1	0.393929	0.869571	0.833333
122	0.823529	0.333333	0.714876	0.617822	0.139489	0.073493	0.954163	1	0.393577	0.960174	0.833333
123	0.882353	0.333333	0.709601	0.520235	0.294351	0.187018	0.88978	1	0.406585	0.999898	0.833333
124	0.941176	0.333333	0.3913	0.300179	0.390853	0.296264	0.655488	1	0.420295	0.786586	0.833333
125	1	0.333333	0.921775	0.852239	0.108121	0.081473	1	1	0.393226	0.950923	0.833333

Pada kolom label tanaman diubah kedalam bentuk kategorikal sehingga kategori tanaman akan berbentuk sebagai berikut:

Tabel 3. 5 *Label Encoding*

TANAMAN	LABEL
JAGUNG	0
KACANG HIJAU	1
KACANG TANAH	2
KEDELAI	3
PADI	4
UBI JALAR	5
UBI KAYU	6

5. Itertools Combination

Penggunaan *itertools combinations* dalam pembuatan model machine learning bertujuan untuk melakukan *feature selection* atau seleksi fitur yang paling relevan dalam set data. Fitur atau variabel pada dataset bisa sangat banyak, tetapi tidak semua fitur tersebut berkontribusi signifikan terhadap hasil prediksi. Dengan menggunakan teknik *itertools combinations*, model dapat memilih kombinasi fitur yang paling relevan dan memberikan kontribusi yang signifikan terhadap kinerja model. Hal ini membantu dalam membangun model yang lebih efisien dan akurat dengan mengurangi overfitting dan mempercepat waktu komputasi [38].

Dalam penerapan teknik *itertools combinations*, penulis menggunakan bantuan modul bernama *itertools.combinations* digunakan untuk membuat semua kombinasi yang mungkin dari elemen-elemen dalam iterable (seperti list atau set) dengan panjang tertentu. Lebih spesifik, fungsi ini menghasilkan iterator yang mengembalikan semua kombinasi yang mungkin dari elemen-elemen tersebut dalam urutan yang tidak menentu, tanpa mengulangi elemen yang sama. Maka fitur yang sudah diproses dengan menggunakan *itertools combinations* akan terlihat seperti tabel berikut.

Tabel 3. 6 Fitur Combination

NO	FITUR
1	TAHUN
2	CURAH HUJAN
3	KELEMBAPAN
4	TEMPREATUR
5	SOIL TEMPERATURE
6	SOIL MOIS
...	...
250	TAHUN,CURAH HUJAN,KELEMBAPAN,TEMPREATUR,SOIL MOIS,LUAS,TANAMAN
251	TAHUN,CURAH HUJAN,KELEMBAPAN,SOIL TEMPERATURE,SOIL MOIS,LUAS,TANAMAN
252	TAHUN,CURAH HUJAN,TEMPREATUR,SOIL TEMPERATURE,SOIL MOIS,LUAS,TANAMAN
253	TAHUN,KELEMBAPAN,TEMPREATUR,SOIL TEMPERATURE,SOIL MOIS,LUAS,TANAMAN
254	CURAH HUJAN,KELEMBAPAN,TEMPREATUR,SOIL TEMPERATURE,SOIL MOIS,LUAS,TANAMAN
255	TAHUN,CURAH HUJAN,KELEMBAPAN,TEMPREATUR,SOIL TEMPERATURE,SOIL MOIS,LUAS,TANAMAN

Nantinya fitur yang sudah diperoleh akan dilakukan modeling dan akan dilakukan perbandingan untuk mencari model akurasi yang terbaik, baik untuk model klasifikasi maupun model regresi

6. Data Splitting

Data yang sudah dinormalisasi akan dipecah menggunakan metode K-fold Cross Validation dengan parameter K yang ditetapkan sebesar 5. Dataset yang telah dinormalisasi akan dibagi menjadi 5 subset yang seimbang. Salah satu subset akan dijadikan data uji, sementara 4 subset lainnya akan digunakan sebagai data latih. Proses ini akan diulang sebanyak 5 kali, dengan setiap subset bergantian menjadi data uji seperti pada gambar berikut.



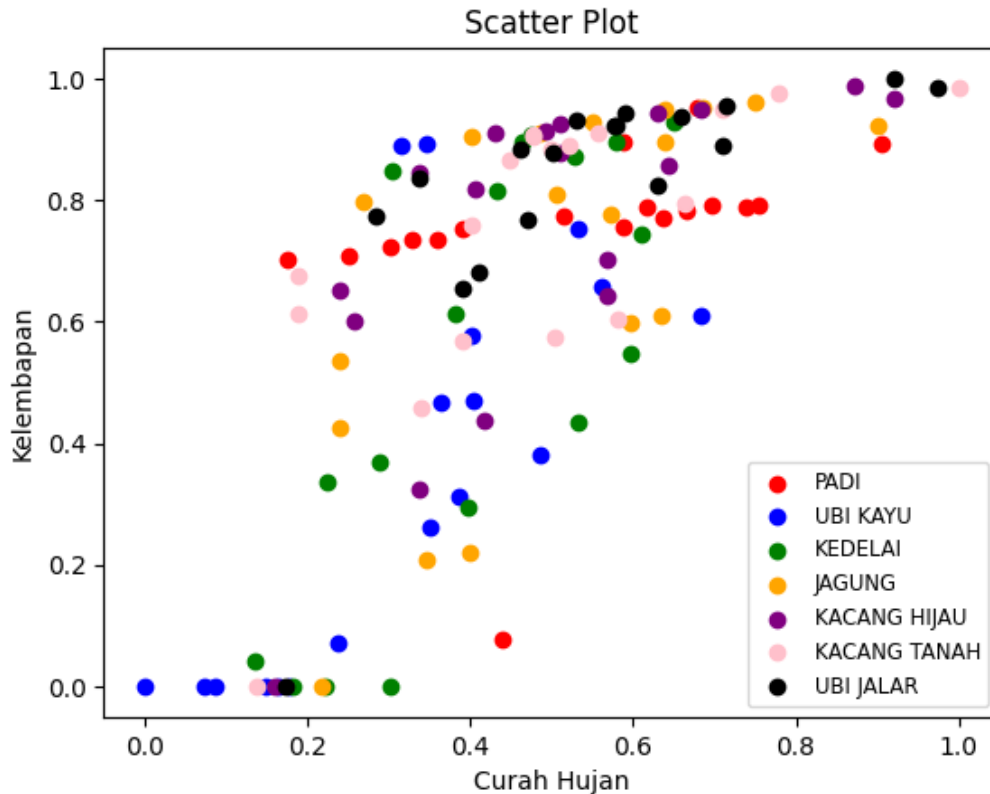
Gambar 3. 3 Skema Kfold

Model kemudian diuji pada setiap subset dan dihitung kinerja rata-ratanya. Teknik ini berguna untuk menghindari overfitting dan memastikan bahwa model dapat bekerja dengan baik pada data yang belum pernah dilihat sebelumnya. Sehingga nantinya dapat mengukur kinerja model secara objektif dan memastikan bahwa model yang dibuat mampu menggeneralisasi dengan baik terhadap data yang baru.

7. Model Klasifikasi

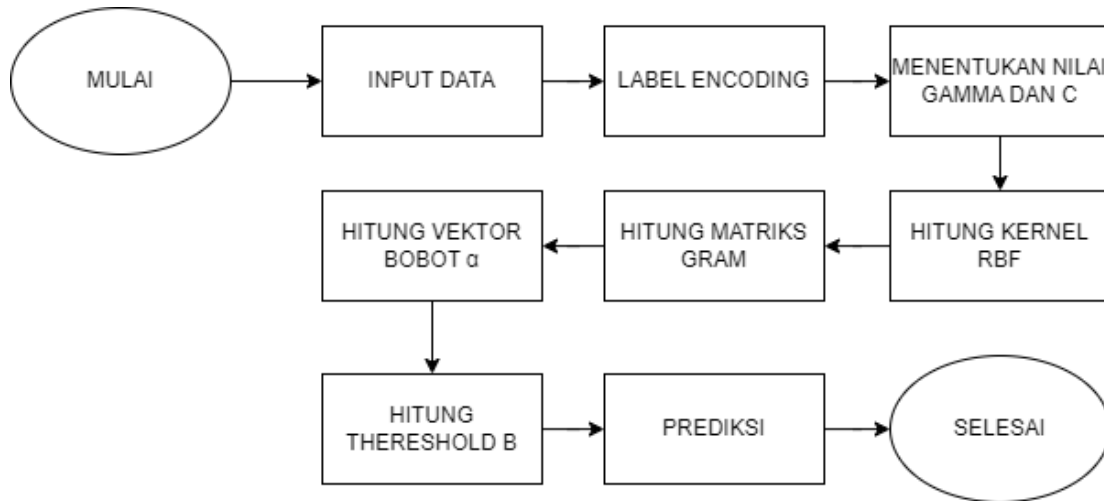
Langkah selanjutnya adalah modeling klasifikasi dengan menggunakan algoritma SVM. Modeling klasifikasi akan menggunakan dua skema, yaitu skema data dengan outlier, dan data tanpa outlier. Sebelum membangun model, perlu diketahui apakah data yang dimiliki bersifat linear atau non-linear. Cara paling umum untuk mengetahui apakah data linear atau non-linear adalah dengan melakukan visualisasi data dalam bentuk *scatter plot*. Jika titik-titik data dapat dikelompokkan dalam bentuk pola garis lurus atau linear, maka data tersebut dikatakan linear. Namun, jika tidak

terdapat pola garis lurus pada data tersebut, melainkan memiliki pola yang lebih kompleks seperti kurva atau lingkaran, maka data tersebut dikatakan non-linear.



Gambar 3. 4 Visualisasi Data dalam *Scatter Plot*

Dari data yang sudah diperoleh dan divisualisasikan dengan variabel kelembapan dan curah hujan, dapat terlihat persebaran data dalam *scatter plot*. Namun, jika diperhatikan dengan cermat, data tidak memiliki pola atau garis yang jelas dan membentuk bentuk yang kompleks dan tidak teratur. Hal ini menunjukkan bahwa data bersifat non linear. Pada kasus seperti ini, algoritma SVM dengan kernel non-linear seperti RBF dapat digunakan untuk melakukan analisis dan pemodelan. Kernel RBF memiliki keunggulan dalam menghasilkan kecocokan terbaik antara data yang ada dan model yang dibuat, bahkan untuk data yang bersifat non-linear sekalipun. Diagram alir perhitungan SVM dapat dilihat pada gambar dibawah.



Gambar 3. 5 Diagram Alir SVM

Pemodelan klasifikasi akan menggunakan variabel x yang berasal dari tahap *itertools combination* sebagai input, yang akan menghasilkan output variabel y_1 . Tujuan dari pemodelan ini adalah untuk memprediksi nilai y_1 berdasarkan nilai x yang diberikan, dimana x adalah data iklim, dan y_1 adalah jenis tanaman. Kernel RBF akan menghitung jarak antara dua titik dalam ruang fitur dengan menggunakan fungsi radial basis. Pada saat dilatih, SVM akan menentukan nilai vektor bobot dan nilai bias yang optimal untuk meminimalkan error pada data latih. Setelah dilatih, model SVM dapat digunakan untuk memprediksi label pada data uji dengan menggunakan nilai bobot dan bias yang telah ditentukan tersebut [40]. Contoh perhitungan SVM dengan menggunakan kernel RBF dengan beberapa sampel data yang digunakan untuk modeling klasifikasi.

Tabel 3. 7 Sampel data klasifikasi

NO	TAHUN	KELEMBAPAN	TEMPREATUR	SOIL TEMPERATURE	SOIL MOIS	LUAS	LABEL TANAMAN
0	0	0.531441	0.738532	0.859473	0.772782	0.4442	4
10	0.588235	0.633332	0.541376	0.74494	0.733762	0.380218	4
20	0.117647	0.759502	0.587128	0.46469	0.46582	0.050816	6
30	0.705882	0	0.953914	1	0	0.018474	6

Ketika melakukan klasifikasi dengan SVM, kita ingin memisahkan dua kelas data yang berbeda. Dalam SVM, kelas-kelas ini diwakili oleh label positif dan label negatif. Oleh karena itu, label tanaman pada Tabel 3.6 perlu diubah agar sesuai dengan

format label positif dan negatif yang diperlukan oleh SVM. Setelah perubahan label dilakukan, data dapat disiapkan untuk diproses melalui algoritma SVM.

Tabel 3. 8 Label Encoding Kelas Tanaman

LABEL TANAMAN	LABEL ENCODING
4	1
4	1
6	-1
6	-1

Setelah itu, dilakukan persiapan parameter dengan nilai $\gamma = 1$ dan $C = 1$. γ merupakan parameter yang mengontrol bentuk dari *decision boundary*. Nilai γ yang besar akan menghasilkan *decision boundary* yang lebih kompleks dan fleksibel, sehingga lebih mungkin *overfitting* pada data *training*. C merupakan parameter yang mengontrol *trade-off* antara mendukung kelancaran *decision boundary* dan mendukung *margin* yang lebar. Nilai C yang besar akan menekankan pentingnya *margin* yang lebar dan memprioritaskan pilihan vektor pendukung yang lebih sedikit. Selanjutnya dilakukan perhitungan menggunakan kernel RBF dari persamaan 2.1 sebagai berikut:

$$\begin{aligned} K(1,1) &= \exp(1(0^2 + 0.259^2 + 0.313^2 + 0.366^2 + 0.314^2 + 0.087^2)) \\ &= 1.0000 \end{aligned}$$

$$\begin{aligned} K(1,3) &= \exp(1(0.118^2 + 0.228^2 + 0.152^2 + 0.243^2 + 0.307^2 + 0.464^2)) \\ &= 0.0821 \end{aligned}$$

$$\begin{aligned} K(1,4) &= \exp(1(0.706^2 + 0.531^2 + 0.784^2 + 0.720^2 + 0.773^2 + 0.172^2)) \\ &= 0.1653 \end{aligned}$$

$$\begin{aligned} K(2,2) &= \exp(1(0^2 + 0.011^2 + 0.227^2 + 0.019^2 + 0.009^2 + 0.225^2)) \\ &= 1.0000 \end{aligned}$$

$$K(2,3) = \exp(1(0.470^2 + 0.130^2 + 0.052^2 + 0.347^2 + 0.341^2 + 0.146^2))$$

$$= 0.6035$$

$$K(2,4) = \exp(1 (0.118^2 + 0.634^2 + 0.052^2 + 0.256^2 + 0.231^2 + 0.367^2)) \\ = 0.9287$$

$$K(3,3) = \exp(1 (0^2 + 0.322^2 + 0.174^2 + 0.229^2 + 0.237^2 + 0.235^2)) \\ = 1.0000$$

$$K(3,4) = \exp(1 (0.706^2 + 0.228^2 + 0.363^2 + 0.067^2 + 0.466^2 + 0.293^2)) \\ = 0.3121$$

$$K(4,4) = \exp(1 (0^2 + 0.531^2 + 0.061^2 + 0^2 + 0.531^2 + 0.408^2)) \\ = 1.000$$

$$K(1,2) = \exp(1(0.588^2 + 0.102^2 + 0.197^2 + 0.195^2 + 0.039^2 + 0.172^2)) \\ = 0.7903$$

Dari perhitungan diatas maka diperoleh tabel matriks dibawah sebagai berikut

Tabel 3. 9 Tabel Matriks K

K(I,J)	K(1,1)	K(1,2)	K(1,3)	K(1,4)
K(1,1)	1	0,7903	0.0821	0.1653
K(1,2)	0.7903	1	0.6035	0.9287
K(1,3)	0.0821	0.6035	1	0.3121
K(1,4)	0.1653	0.9287	0.3121	1

Selanjutnya dilakukan perhitungan matriks Gram dari matriks kernel K diatas. Persamaan dari matriks Gram adalah sebagai berikut.

$$G_{i,j} = \sum \alpha_i \times \alpha_j \times K_{i,j}. \quad (3.1)$$

Terlebih dahulu, nilai α dari solusi persamaan optimasi SVM harus ditentukan. Untuk menentukan nilai α maka digunakan algoritma optimasi SMO (Sequential Minimal Optimization). Keunggulan dari optimasi SMO adalah dapat meningkatkan kecepatan perhitungan untuk mendapatkan nilai α yang optimal. Namun, dalam penelitian ini,

perhitungan nilai α dilakukan dengan menggunakan bahasa pemrograman Python, sehingga diperoleh nilai α sebagai berikut

$$\alpha = [0.87042982, 0.80949852, 1.06707394, 0.61285441]$$

Maka, kita dapat menghitung matriks Gram G Dengan mengganti nilai α_i dan K_{ij} ke dalam persamaan tersebut, kita dapat menghitung nilai G_{ij} pada setiap pasang i dan j sehingga diperoleh matriks gram G sebagai berikut

$$G = \begin{bmatrix} 1.0000 & 0.7903 & 0.0821 & 0.1653 \\ 0.7903 & 1.0000 & 0.6035 & 0.9287 \\ 0.0821 & 0.6035 & 1.0000 & 0.3121 \\ 0.1653 & 1.0000 & 0.3121 & 1.0000 \end{bmatrix}$$

Setelah memperoleh matriks Gram, selanjutnya menghitung vektor bobot α , dan memecahkan persamaan sistem linier:

$$G\alpha = y \quad (3.2)$$

di mana y adalah vektor target dengan nilai 1 untuk kelas yang dipilih dan -1 untuk kelas lainnya. Dalam contoh ini memiliki 2 kelas, dan perlu memecahkan persamaan sistem linier ini dua kali dengan menggunakan vektor target yang berbeda untuk masing-masing kelas. Untuk menyelesaikan persamaan ini maka digunakan metode eliminasi *Gauss*. Pertama, tambahkan vektor target untuk kelas pertama dan kelas kedua ke dalam matriks G :

$$G = \begin{bmatrix} 1.0000 & 0.7903 & 0.0821 & 0.1653 & 1 \\ 0.7903 & 1.0000 & 0.6035 & 0.9287 & 1 \\ 0.0821 & 0.6035 & 1.0000 & 0.3121 & -1 \\ 0.1653 & 1.0000 & 0.3121 & 1.0000 & -1 \end{bmatrix}$$

Kemudian, gunakan metode eliminasi *Gauss* untuk menghasilkan matriks segitiga atas:

$$G' = \begin{bmatrix} 1.0000 & 0.7903 & 0.0821 & 0.1653 & 1 \\ 0 & 0.2109 & 0.5868 & 0.6477 & 0.2109 \\ 0 & 0 & 0.9846 & 0.1536 & -1.0294 \\ 0 & 0 & 0 & 0.5785 & -1.2764 \end{bmatrix}$$

Dengan matriks segitiga atas ini, maka dapat dengan mudah menghitung vektor bobot α dengan metode *back substitution*:

$$\begin{aligned}
\alpha_4 &= -1.2764 - 0.5785 \\
&= -2.2042 \\
\alpha_3 &= \frac{(0.2109 - 0.1536(-2.2042) + (-1.0294)(0))}{0.9846} \\
&= 0.3747 \\
\alpha_2 &= \frac{(0.6477 - 0.5868(0.3747) - 0.2109(-2.2042) + (0)(0))}{0.2109} \\
&= 4.0184 \\
\alpha_1 &= \frac{(1 - 0.7903(4.0184) - 0.0821(-2.2042) - 1(-2.2042))}{1} \\
&= 4.4751
\end{aligned}$$

Setelah nilai vektor bobot diketahui selanjutnya mencari nilai eigenvalue dari matriks Gram yang sudah diubah menjadi matriks segitiga atas. Nilai *eigenvalue* tersebut akan digunakan untuk menentukan nilai *threshold* b dengan mengikuti aturan yang telah ditentukan sebelumnya. Dari hasil perhitungan sebelumnya, telah didapatkan matriks segitiga atas sebagai berikut:

$$G' = \begin{bmatrix} 1.0000 & 0.7903 & 0.0821 & 0.1653 & 1 \\ 0 & 0.2109 & 0.5868 & 0.6477 & 0.2109 \\ 0 & 0 & 0.9846 & 0.1536 & -1.0294 \\ 0 & 0 & 0 & 0.5785 & -1.2764 \end{bmatrix}$$

Nilai eigenvalue dari persamaan segitiga atas adalah elemen-elemen diagonal dari matriks tersebut. Sehingga, nilai *eigenvalue* dari matriks G' adalah:

$$\lambda_1 = 1.0000$$

$$\lambda_2 = 0.2109$$

$$\lambda_3 = 0.9846$$

$$\lambda_4 = 0.5785$$

nilai *threshold* b dapat dihitung dengan menggunakan persamaan

$$b = \frac{(\lambda_{max} + \lambda_{min})}{2} \quad (3.3)$$

Dengan λ_{max} adalah *eigenvalue* terbesar dan λ_{min} adalah *eigenvalue* terkecil.

$$\begin{aligned}
b &= \frac{(1.0000 + 0.2109)}{2} \\
&= 0.60545
\end{aligned}$$

Sehingga nilai *threshold* b yang didapatkan adalah 0.60545.

Untuk melakukan prediksi dengan data uji, pertama-tama perlu dihitung nilai kernel RBF dari data uji. Kemudian baru dapat dihitung nilai fungsi diskriminan untuk memprediksi label data uji. Sebagai contoh diberikan data uji $x_1 = [0.5, 0.3, 0.2, 0.4]$. Hitung nilai kernel RBF data latih yang digunakan untuk melakukan klasifikasi pada model:

$$\begin{aligned} K(x_1, x_1) &= \exp(-1 \times 0^2) \\ &= 1 \end{aligned}$$

$$\begin{aligned} K(x_2, x_1) &= \exp(-1 \times (0.5 - 0.3)^2) \\ &= 0.7788 \end{aligned}$$

$$\begin{aligned} K(x_1, x_2) &= \exp(-1 \times (0.2 - 0.5)^2) \\ &= 0.0909 \end{aligned}$$

$$\begin{aligned} K(x_1, x_2) &= \exp(-1 \times (0.4 - 0.5)^2) \\ &= 0.6065 \end{aligned}$$

Selanjutnya dilakukan perhitungan nilai fungsi diskriminan:

$$f(x_1) = \sum (\alpha_i y_i K(x_i, x_1)) + b \quad (3.5)$$

Secara lebih rinci, dalam persamaan ini:

- $f(x_1)$ adalah nilai output dari model SVM untuk input x_1 .
- \sum adalah simbol sigma yang menunjukkan operasi penjumlahan pada seluruh nilai yang ada di dalam kurung.
- α_i adalah koefisien alpha
- y_i adalah target output
- $K(x_i, x_1)$ adalah kernel function
- b adalah nilai offset yang juga dihitung saat pelatihan model SVM.

Sehingga dapat diperoleh perhitungan

$$\begin{aligned}
 f(x_1) &= (-2.2042 \times 1 \times 1) + (0.3747 \times 1 \times 0.7788) \\
 &\quad + (4.0184 \times -1 \times 0.0909) + (4.4751 \times -1 \times 0.6065) \\
 &\quad + 0.6054 \\
 &= -4.4044
 \end{aligned}$$

maka diperoleh nilai $f(x_1)$ adalah -4.4044. Jika $f(x_1) > 0$, maka labelnya adalah kelas 1. Jika $f(x_1) < 0$, maka labelnya adalah kelas -1. Dari perhitungan tadi diperoleh $f(x_1) > 0$, sehingga label data uji adalah kelas 0. Jika model sudah bisa memprediksi maka bisa dilakukan evaluasi untuk mengukur performa model.

5. Evaluasi Model Klasifikasi

Evaluasi model klasifikasi yang sudah dilakukan modeling menggunakan *confusion matrix*. *Confusion matrix* adalah sebuah tabel yang digunakan untuk mengevaluasi performa model dalam melakukan klasifikasi pada suatu data. *Confusion matrix* memuat empat nilai penting, yaitu *True Positive (TP)*, *False Positive (FP)*, *True Negative (TN)*, dan *False Negative (FN)* [41]. *Confusion matrix* digunakan untuk menghitung berbagai metrik evaluasi model, seperti akurasi, presisi, *recall*, dan *F1 score*. Metrik-metrik ini dapat membantu dalam mengevaluasi performa model dan mengidentifikasi area yang perlu diperbaiki atau dioptimalkan. Contoh perhitungan manual untuk evaluasi performa model klasifikasi.

Tabel 3. 10 *Confusion Matrix*

CLASS	PREDIKSI 1	PREDIKSI 2	PREDIKSI 3	PREDIKSI 4	PREDIKSI 5	PREDIKSI 6	PREDIKSI 7
AKTUAL 1	10	1	0	0	0	0	0
AKTUAL 2	2	8	1	1	1	0	0
AKTUAL 3	0	0	10	0	0	0	0
AKTUAL 4	0	1	0	9	0	0	0
AKTUAL 5	0	2	0	0	8	0	0
AKTUAL 6	0	0	0	0	0	10	0
AKTUAL 7	0	0	0	0	0	0	10

Setelah didapatkan confusion matriks, maka bisa dilakukan penghitungan performa model sebagai berikut:

$$\begin{aligned} \text{Akurasi} &= (\text{TP}_1 + \text{TP}_2 + \text{TP}_3 + \text{TP}_4 + \text{TP}_5 + \text{TP}_6 + \text{TP}_7) / \text{Total} \\ &= (10 + 8 + 10 + 9 + 8 + 10 + 10) / (70) \\ &= 0.914 \text{ atau } 91.4\% \end{aligned}$$

$$\begin{aligned} \text{Presisi}_1 &= \text{TP}_1 / (\text{TP}_1 + \text{FP}_1) \\ &= 10 / (10 + 3 + 2 + 1 + 1 + 1 + 1) \\ &= 0.625 \text{ atau } 62.5\% \end{aligned}$$

$$\begin{aligned} \text{Recall}_1 &= \text{TP}_1 / (\text{TP}_1 + \text{FN}_1) \\ &= 10 / (10 + 1 + 0 + 0 + 0 + 0 + 0) \\ &= 0.909 \text{ atau } 90.9\% \end{aligned}$$

$$\begin{aligned} F1 \text{ score}_1 &= 2 \times \text{Presisi}_1 \times \text{Recall}_1 / (\text{Presisi}_1 + \text{Recall}_1) \\ &= 2 \times 0.625 \times 0.909 / (0.625 + 0.909) \\ &= 0.742 \text{ atau } 74.2\% \end{aligned}$$

Perhitungan untuk kelas 2, 3, 4, 5, 6, dan 7 juga dapat dilakukan dengan cara yang sama. Dengan demikian, dapat dilihat bahwa model tersebut memiliki akurasi yang cukup baik dan juga cukup baik dalam mengklasifikasikan setiap kelas.

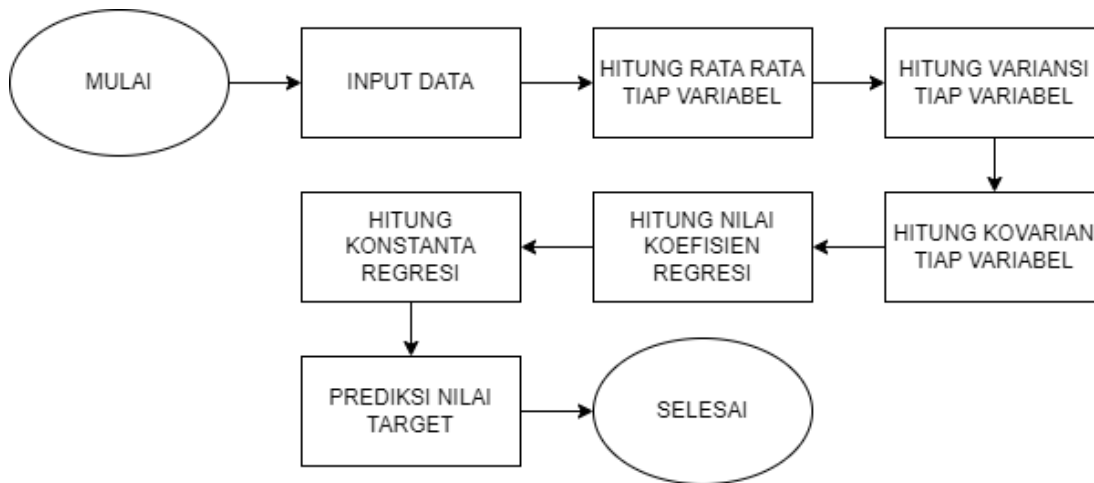
Evaluasi model regresi dilakukan untuk mengukur seberapa baik model dapat memprediksi nilai output yang kontinu atau numerik [42]. Pengukuran model regresi pada penelitian ini menggunakan *Mean Absolute Error* (MAE). MAE menghitung rata-rata dari selisih absolut antara nilai prediksi dan nilai sebenarnya. Semakin kecil nilai MAE, semakin baik model dalam memprediksi nilai output. Contoh perhitungan untuk mengukur evaluasi dari model regresi yang sudah dilatih.

6. Model Regresi

Setelah melakukan proses klasifikasi dan mendapatkan variabel Y_1 , maka langkah selanjutnya adalah mencari nilai produktivitas dengan menggunakan regresi linear. Proses modeling akan menggunakan dua skema, yaitu modeling dengan menggunakan data outlier, dan data tanpa outlier. Nantinya, nilai variabel Y_1 juga akan digunakan untuk melakukan pencarian produktivitas.

Regresi linear dapat digunakan untuk memprediksi produktivitas pertanian dengan mengidentifikasi variabel-variabel yang berpengaruh terhadap produktivitas

pertanian. Proses modeling regresi linear akan menggunakan variabel X dari *itertools combination* sebagai input yang akan menghasilkan output regresi berupa Y. Contoh perhitungan produktivitas dengan menggunakan regresi linear berdasarkan diagram alir pada gambar 3.5



Gambar 3. 6 Diagram Alir Perhitungan Regresi

Dalam diagram alir di atas, langkah pertama adalah memasukkan data. Contoh sampel data yang digunakan mencakup tiga variabel yaitu Curah Hujan (X_1), Tanaman (X_2), dan Produktivitas (Y), yang dapat dilihat pada Tabel 3.9 di bawah.

Tabel 3. 11 Sampel data regresi

NO	CURAH HUJAN	TANAMAN	PRODUKTIVITAS
1	388.8277283	4	5.53
2	280.910675	4	6.49
3	431.8197632	4	6.51
4	141.8188171	6	23.48
5	70.43282318	6	18.13
6	244.5579987	6	17.35

Kemudian dilakukan penghitungan rata-rata dari masing-masing variabel CURAH HUJAN (x_1), TANAMAN (x_2), serta nilai PRODUKTIVITAS (y)

$$\bar{x}_1 = \frac{(388.8 + 280.9 + 431.8 + 141.8 + 10.4 + 244.5)}{6}$$

$$\begin{aligned}
 &= 249.2 \\
 \bar{x}_2 &= \frac{(4 + 4 + 4 + 6 + 6 + 6)}{6} \\
 &= 5 \\
 \bar{y} &= \frac{(5.53 + 6.49 + 6.51 + 23.48 + 18.13 + 17.35)}{6} \\
 &= 12.9
 \end{aligned}$$

Setelah didapatkan rata rata dari variabel tersebut selanjutnya adalah menghitung variansi dari variabel x1 dan x2. Persamaan untuk menghitung variansi adalah sebagai berikut

$$var = \Sigma(x_i - \bar{x})^2 / (n - 1) \quad (3.6)$$

Di mana:

Σ = simbol sigma, yang artinya menjumlahkan

x_i = nilai pada sampel

\bar{x} = rata-rata sampel

n = jumlah elemen pada sampel

Maka untuk penerapan perhitungannya adalah sebagai berikut

$$\begin{aligned}
 var_{x_1} &= ((388.8277283 - 249.2)^2 + (280.910675 - 249.2)^2 \\
 &\quad + (431.8197632 - 249.2)^2 + (141.8188171 - 249.2)^2 \\
 &\quad + (70.43282318 - 249.2)^2 + (244.5579987 - 249.2)^2) / 5 \\
 &= 1059.21491264 \\
 var_{x_2} &= ((4 - 5)^2 + (4 - 5)^2 + (4 - 5)^2 + (6 - 5)^2 + (6 - 5)^2 \\
 &\quad + (6 - 5)^2) / 5 \\
 &= 0.6
 \end{aligned}$$

Langkah selanjutnya adalah menghitung Hitung kovarians antara x_1 dan y dengan persamaan:

$$cov_{(x,y)} = \Sigma[(x - \bar{x}) * (y - \bar{y})] / (n - 1) \quad (3.7)$$

Di mana:

Σ = Simbol sigma yang menunjukkan operasi penjumlahan
dari suatu deret bilangan

X = Nilai pada variabel x

\bar{x} = Nilai rata-rata dari variabel x

y = Nilai pada variabel y

\bar{y} = Nilai rata-rata dari variabel y

n = Jumlah total data

Maka bisa dilakukan perhitungan dari persamaan diatas sebagai berikut:

$$\begin{aligned} cov_{(x1,y)} &= ((388.8277283 - 252.0750966)(5.53 - 11.155) \\ &\quad + (280.910675 - 252.0750966)(6.49 - 11.155) \\ &\quad + (431.8197632 - 252.0750966)(6.51 - 11.155) \\ &\quad + (141.8188171 - 252.0750966)(23.48 - 11.155) \\ &\quad + (70.43282318 - 252.0750966)(18.13 - 11.155) \\ &\quad + (244.5579987 - 252.0750966)(17.35 - 11.155))/5 \\ &= 5623.085015 \end{aligned}$$

$$\begin{aligned} cov_{(x2,y)} &= ((4 - 4.67)(5.53 - 11.155) + (4 - 4.67)(6.49 - 11.155) \\ &\quad + (4 - 4.67)(6.51 - 11.155) + (6 - 4.67)(23.48 - 11.155) \\ &\quad + (6 - 4.67)(18.13 - 11.155) + (6 - 4.67)(17.35 - 11.155))/5 \\ &= 8.646666667 \end{aligned}$$

Berikutnya adalah menentukan nilai koefisien regresi β_1 dan β_2 dengan persamaan sebagai berikut.

$$\beta_1 = cov_{(x1,y)} / var_{(x1)} \quad (3.8)$$

Kovarians adalah ukuran statistik yang mengukur seberapa dekat hubungan antara dua variabel. Dalam kasus ini, $cov_{(x1,y)}$ adalah kovarians antara x_1 dan y , yang menggambarkan seberapa erat hubungan antara variabel independen x_1 dan variabel dependen y . Sehingga bisa dihitung kovarians

$$\begin{aligned}\beta_1 &= 5623.085015 / 1059.21 \\ &= 0.1911320058\end{aligned}$$

$$\begin{aligned}\beta_2 &= 8.646666667 / 0.6 \\ &= 18.18181818\end{aligned}$$

Selanjutnya menghitung konstanta regresi β_0 dengan persamaan:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2 \quad (3.9)$$

dengan \bar{y} , \bar{x}_1 , dan \bar{x}_2 adalah nilai rata-rata dari kolom y , x_1 , dan x_2 , masing-masing.

$$\begin{aligned}\beta_0 &= 11.155 - 0.1911320058 \times 252.0750966 - 18.18181818 \times 4.67 \\ &= -63.25452719\end{aligned}$$

Dengan demikian, model regresi linier dari sampel data tersebut adalah:

$$y = -37.25221428 + 0.1911320058 \times x_1 + 18.18181818 \times x_2$$

Jika data uji memiliki nilai x_1 dan x_2 , maka dapat digunakan persamaan tersebut untuk memprediksi nilai y yang sesuai. Misalkan, jika $x_1 = 300$ dan $x_2 = 5$, maka:

$$\begin{aligned}y &= -37.25221428 + 0.1911320058 \times 300 + 18.18181818 \times 5 \\ &= 15.11827645\end{aligned}$$

Jadi, model ini memprediksi bahwa nilai y akan sekitar 15.12 jika $x_1 = 300$ dan $x_2 = 5$.

7. Evaluasi Model Regresi

Evaluasi model regresi dilakukan untuk mengukur seberapa baik model dapat memprediksi nilai output yang kontinu atau numerik [42]. Pengukuran model regresi pada penelitian ini menggunakan *Mean Absolute Error* (MAE). MAE menghitung rata-rata dari selisih absolut antara nilai prediksi dan nilai sebenarnya. Semakin kecil nilai MAE, semakin baik model dalam memprediksi nilai output. Contoh perhitungan untuk mengukur evaluasi dari model regresi yang sudah dilatih.

Tabel 3. 12 Data Aktual dan Prediksi

NO	PREDIKSI	AKTUAL
1	6.8	7.5
2	9.9	10.2
3	12.6	12.1
4	13.9	13.7
5	14.5	14.2

Maka perhitungan MAE adalah sebagai berikut:

$$MAE = \frac{1}{n} \sum_{i=1}^n |x_i - x| \quad (3.10)$$

Dimana:

- n adalah jumlah sampel yang diprediksi
- x_i adalah nilai prediksi untuk sampel ke-i
- x adalah nilai sebenarnya dari target

$$\begin{aligned} MAE &= (1/5) \times (|7.5 - 6.8| + |10.2 - 9.9| + |12.1 - 12.6| + |13.7 - \\ &13.9| + |14.2 - 14.5|) \\ &= 0.66 \end{aligned}$$

Dengan demikian, MAE dari model regresi tersebut adalah 0.66. Semakin kecil nilai MAE, semakin baik model dalam memprediksi nilai output

3.3.4 Analisis

Setelah selesai melakukan proses pengolahan data, langkah berikutnya adalah melakukan analisis terhadap hasil pengolahan tersebut. Tujuannya adalah untuk mengevaluasi kinerja dan akurasi dari hasil pengolahan tersebut. Hal ini penting untuk menentukan apakah data yang dihasilkan dapat dipercaya dan digunakan untuk keperluan selanjutnya atau tidak.

3.3.5 Kesimpulan dan Saran

Setelah melakukan analisis, hasil yang diperoleh kemudian dapat digunakan untuk menyimpulkan jawaban atas permasalahan yang telah diangkat dalam penelitian. Dengan mengacu pada hasil analisis tersebut, kesimpulan dapat dirumuskan secara ringkas dan padat untuk memberikan jawaban yang tepat dan jelas terhadap permasalahan yang telah diidentifikasi dalam penelitian..