Laporan Mata Kuliah Pembelajaran Mesin

DATA PREPARATION DARI SUMBER OPEN SOURCE

disusun untuk memenuhi
tugas mata kuliah Pemrosesan Mesin B

oleh :

Mila Lestari	(2208107010002)
Zahra Zafira	(2208107010040)
Pryta Rosela	(2208107010046)
Cut Sula Fhatia Rahma	(2208107010048)
Widya Nurul Sukma	(2208107010054)

JURUSAN INFORMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS SYIAH KUALA

2025

1. Pendahuluan

Polycystic Ovary Syndrome (PCOS) adalah salah satu gangguan endokrin yang paling umum terjadi pada wanita usia reproduksi. Kondisi ini ditandai oleh ketidakseimbangan hormon, siklus menstruasi yang tidak teratur, dan keberadaan kista di ovarium. Diagnosis dini sangat penting untuk mengelola kondisi ini secara efektif. Dalam laporan ini, kami melakukan analisis data PCOS menggunakan teknik machine learning untuk menemukan pola dalam data dan membangun model prediktif yang akurat.

2. Data Description

2.1 Nama Dataset dan Sumber

Dalam pembuatan tugas ini, kami menggunakan dataset dari kaggle https://www.kaggle.com/datasets/samikshadalvi/pcos-diagnosis-dataset. Dataset ini berisi informasi terkait **Sindrom Ovarium Polikistik (PCOS)**, yaitu gangguan hormonal umum yang memengaruhi perempuan usia reproduktif. Dataset ini memiliki 1000 entri, di mana setiap entri mewakili seorang pasien.

2.2 Deskripsi Singkat Dataset

Dataset ini berisi data kesehatan pasien yang mencakup variabel-variabel utama yang berkaitan dengan PCOS, seperti usia, BMI, kadar testosteron, dan jumlah folikel antral. Data ini dapat digunakan untuk eksplorasi, analisis, dan pengembangan model prediksi PCOS.

2.3 Jumlah Data

- **Jumlah sampel**: 1.000 pasien
- **Jumlah fitur**: 5 fitur utama yang terkait dengan PCOS
    - **Usia (tahun)**: Usia pasien, berkisar antara 18 hingga 45 tahun.
    - **BMI (kg/m²)**: Indeks Massa Tubuh (IMT), yaitu ukuran lemak tubuh berdasarkan tinggi dan berat badan, berkisar antara 18 hingga 35.
    - **Ketidakteraturan Menstruasi (biner)**: Indikator biner yang menunjukkan apakah pasien mengalami siklus menstruasi yang tidak teratur (0 = Tidak, 1 = Ya).
    - **Kadar Testosteron (ng/dL)**: Tingkat hormon testosteron dalam darah pasien, yang merupakan indikator penting dalam PCOS, berkisar antara 20 hingga 100 ng/dL.

- **Jumlah Folikel Antral**: Jumlah folikel antral yang terdeteksi melalui USG, berkisar antara 5 hingga 30**.** Fitur ini digunakan untuk menilai cadangan ovarium dan kemungkinan adanya PCOS.

- **Label**: PCOS Diagnosis (0 = Tidak PCOS, 1 = PCOS). Indikator biner yang menunjukkan apakah pasien didiagnosis mengidap PCOS (0 = Tidak, 1 = Ya).

2.4 Format Data

Dataset tersedia dalam format **CSV (Comma-Separated Values)** yang memudahkan pemrosesan menggunakan Python dan library seperti Pandas dan NumPy. Dataset ini memiliki 6 kolom dengan tipe data sebagai berikut :

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 6 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Age                        1000 non-null   int64
 1   BMI                        1000 non-null   float64
 2   Menstrual_Irregularity     1000 non-null   int64
 3   Testosterone_Level(ng/dL)  1000 non-null   float64
 4   Antral_Follicle_Count      1000 non-null   int64
 5   PCOS_Diagnosis             1000 non-null   int64
dtypes: float64(2), int64(4)
memory usage: 47.0 KB
```

Seluruh fitur dalam dataset memiliki **1000 entri tanpa missing values**, sehingga dapat langsung digunakan untuk analisis dan model prediktif.

3. Data Loading

3.1 Cara Memuat Data ke Lingkungan Pemrograman

Dataset dimuat ke dalam Python menggunakan Pandas, yang merupakan salah satu library paling umum untuk analisis data.Dataset ini dimuat ke dalam lingkungan pemrograman Python menggunakan **Pandas**, sebuah library yang umum digunakan untuk manipulasi dan analisis data. Selain itu, **kagglehub** digunakan untuk mengunduh dataset langsung dari Kaggle.

Proses pemuatan data dilakukan dengan langkah-langkah berikut:

1. **Mengunduh dataset** dari Kaggle menggunakan perintah `kagglehub.dataset_download()`.
2. **Menentukan path file CSV** dalam dataset yang telah diunduh.

Kode pemuatan data:

```
import kagglehub

# Download latest version
path = kagglehub.dataset_download("samikshadalvi/pcos-diagnosis-dataset")

print("Path to dataset files:", path)
```

```
Path to dataset files: /root/.cache/kagglehub/datasets/samikshadalvi/pcos-diagnosis-dataset/versions/1
```

```
path = os.path.join(path, "pcos_dataset.csv")
df = pd.read_csv(path)
```

3.2 Tantangan dalam Memuat Data

Dalam proses pemuatan dataset, terdapat beberapa aspek yang perlu diperhatikan untuk memastikan data siap digunakan dalam analisis dan pemodelan. Berikut beberapa tantangan yang dihadapi:

- **Menangani Data yang Hilang (Missing Values)**
  - Berdasarkan hasil pengecekan menggunakan df.isnull().sum(), tidak ditemukan nilai yang hilang dalam dataset ini. Namun, pada tahap eksplorasi data, penting untuk selalu melakukan pengecekan missing values karena :
    - Kehilangan data dapat menyebabkan bias dalam analisis dan mempengaruhi akurasi model.
    - Jika ditemukan missing values, dapat ditangani dengan metode seperti **penghapusan baris/kolom** yang memiliki nilai kosong atau **imputasi nilai** menggunakan mean, median, atau metode lainnya.

```
df.isnull().sum()
```

	0
Age	0
BMI	0
Menstrual_Irregularity	0
Testosterone_Level(ng/dL)	0
Antral_Follicle_Count	0
PCOS_Diagnosis	0

dtype: int64

- **Memastikan Konsistensi Format Data**
  - Dataset ini memiliki tipe data yang sesuai, yaitu **int64** dan **float64**, sehingga tidak diperlukan konversi tipe data.
  - Meskipun tidak ada masalah dalam dataset ini, dalam banyak kasus, tipe data numerik bisa tersimpan dalam format teks, yang memerlukan konversi sebelum digunakan dalam analisis.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 6 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Age                        1000 non-null   int64
 1   BMI                        1000 non-null   float64
 2   Menstrual_Irregularity     1000 non-null   int64
 3   Testosterone_Level(ng/dL)  1000 non-null   float64
 4   Antral_Follicle_Count      1000 non-null   int64
 5   PCOS_Diagnosis             1000 non-null   int64
dtypes: float64(2), int64(4)
memory usage: 47.0 KB
```

- **Deteksi dan Penanganan Data Duplikat**
  - Berdasarkan hasil pengecekan menggunakan df.duplicated().sum(), tidak ditemukan data duplikat dalam dataset.
  - Meskipun tidak ada duplikasi, pemeriksaan ini tetap diperlukan untuk mencegah kesalahan dalam analisis.

```
df.duplicated().sum()
```

```
0
```

- **Identifikasi dan Penanganan Outlier**
  - Dari hasil df.describe(), terlihat bahwa beberapa fitur seperti **Testosterone Level** dan **Antral Follicle Count** memiliki nilai maksimum yang cukup tinggi dibandingkan dengan nilai rata-ratanya.
  - Perlu dilakukan analisis lebih lanjut untuk memastikan apakah nilai ekstrim tersebut merupakan outlier atau bagian dari distribusi normal dataset.

```
df.describe()
```

	Age	BMI	Menstrual_Irregularity	Testosterone_Level(ng/dL)	Antral_Follicle_Count	PCOS_Diagnosis
count	1000.000000	1000.00000	1000.000000	1000.000000	1000.000000	1000.000000
mean	31.771000	26.38700	0.530000	60.159500	17.469000	0.199000
std	8.463462	4.93554	0.499349	23.160204	7.069301	0.399448
min	18.000000	18.10000	0.000000	20.000000	5.000000	0.000000
25%	24.000000	21.90000	0.000000	41.700000	12.000000	0.000000
50%	32.000000	26.40000	1.000000	60.000000	18.000000	0.000000
75%	39.000000	30.50000	1.000000	80.300000	23.250000	0.000000
max	45.000000	35.00000	1.000000	99.800000	29.000000	1.000000

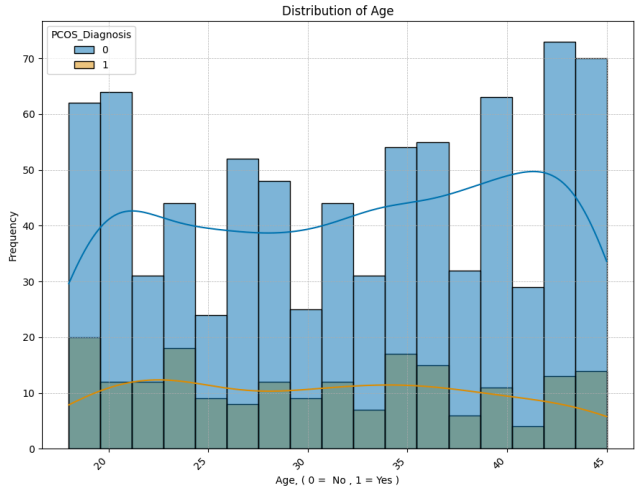
4. Data Understanding

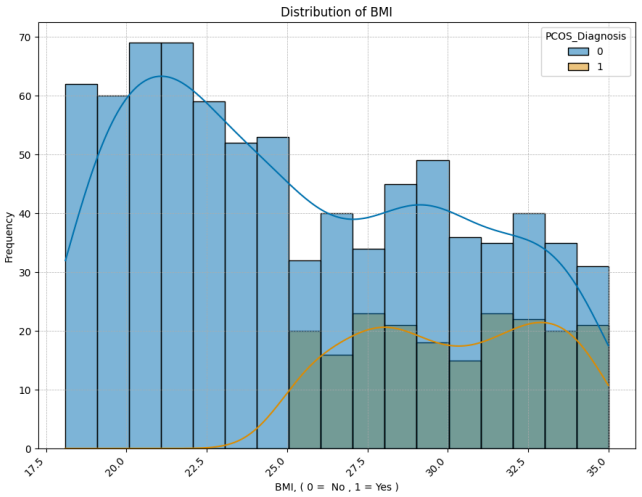
4.1 Statistik Dasar Dataset

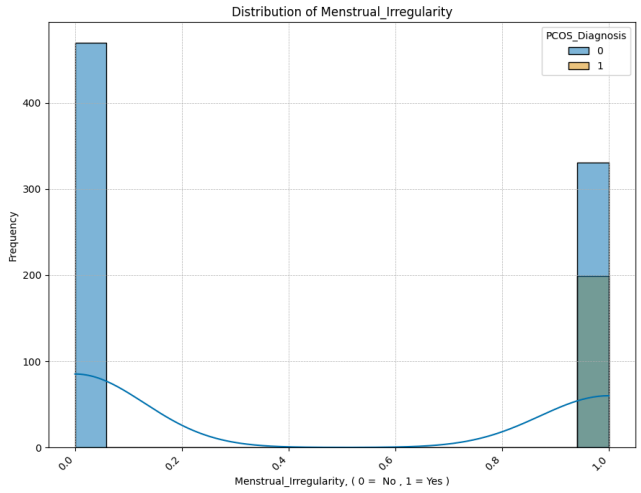
Langkah pertama dalam memahami dataset adalah melihat statistik dasar, seperti distribusi data dan korelasi antar fitur.

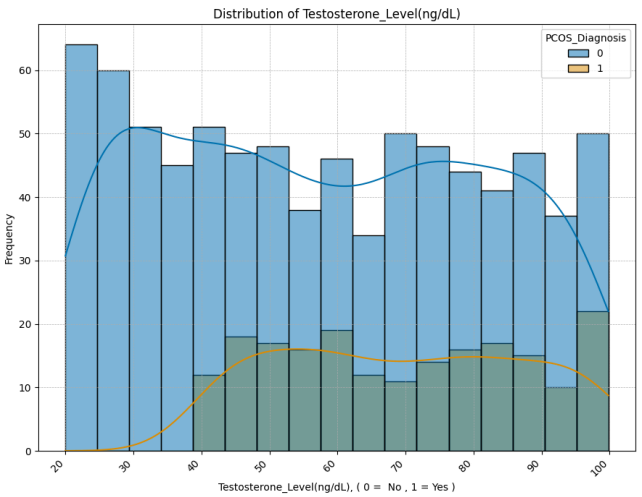
Kode eksplorasi data:

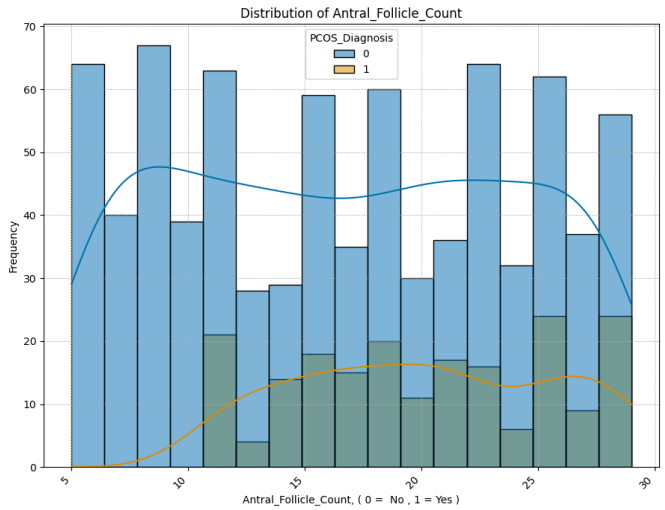
```
for i in df.columns:
  plt.figure(figsize=(9,7))
  sns.histplot(x=i,kde=True,hue='PCOS_Diagnosis',palette='colorblind',data=df,bins=17)

  plt.xlabel(i+', ( 0 =  No , 1 = Yes )')
  plt.ylabel("Frequency")
  plt.title("Distribution of {}".format(i))
  plt.xticks(rotation=45, ha='right')
  plt.tight_layout()
  plt.grid(linestyle = '--', linewidth = 0.5)
  plt.show()
  print("\n")
```

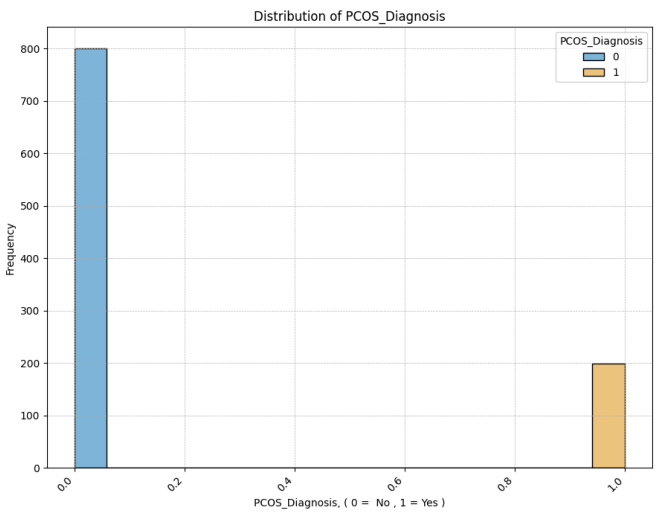






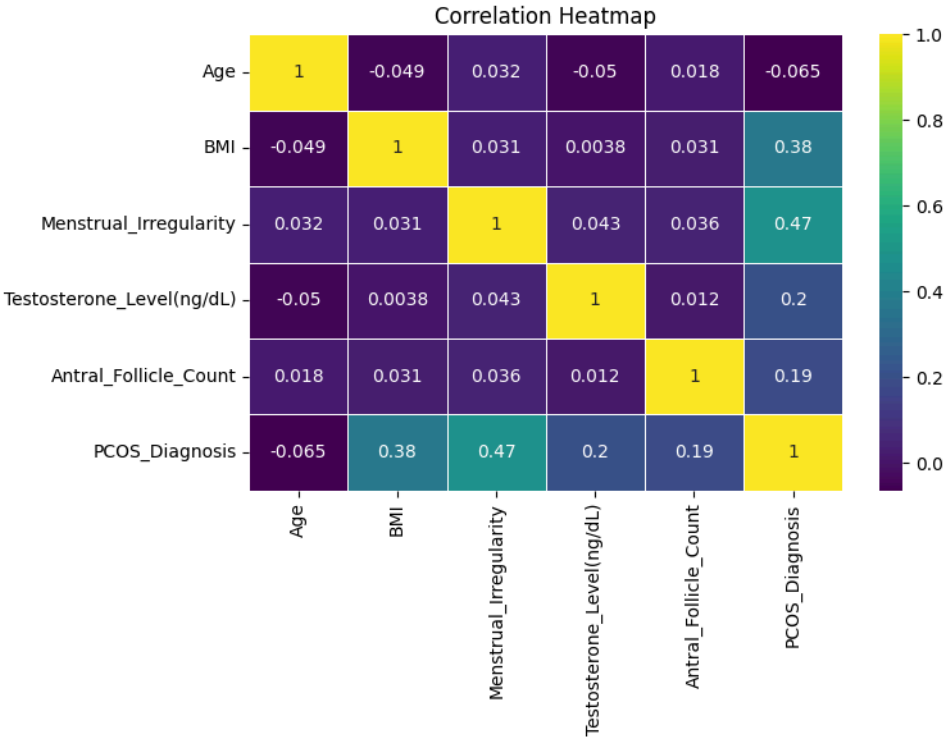






Kode korelasi antar fitur:

```
plt.figure(figsize=(8,6))
sns.heatmap(df.corr(),annot=True,cmap='viridis',linewidths=0.5)
plt.title("Correlation Heatmap")
plt.tight_layout()
plt.show()
```



Berdasarkan HeatMap berikut adalah beberapa korelasi penting antara fitur-fitur dalam dataset :

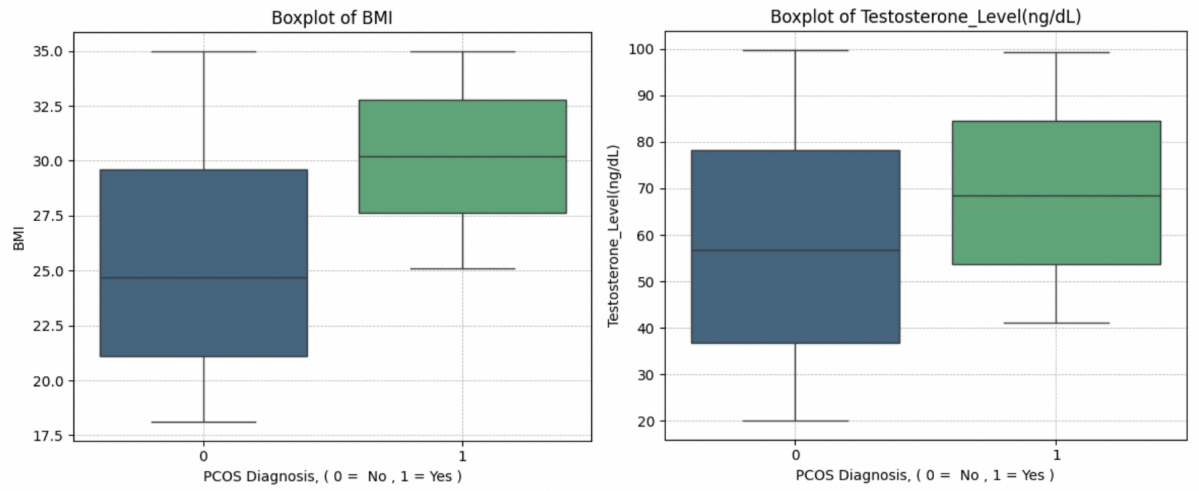
- **PCOS Diagnosis** memiliki korelasi positif dengan **Menstrual Irregularity (0.47)**, menunjukkan bahwa individu dengan PCOS cenderung mengalami ketidakteraturan menstruasi.
- **PCOS Diagnosis** juga berkorelasi positif dengan **BMI (0.38)** dan **Testosterone Level (0.20)**, yang mengindikasikan bahwa individu dengan BMI lebih tinggi dan kadar testosteron lebih tinggi lebih mungkin didiagnosis dengan PCOS.
- **Antral Follicle Count** memiliki korelasi positif dengan **PCOS Diagnosis (0.19)**, meskipun tidak terlalu kuat.
- Korelasi antar fitur lainnya relatif rendah, menunjukkan bahwa variabel-variabel dalam dataset tidak terlalu redundant.

4.2 Visualisasi Data

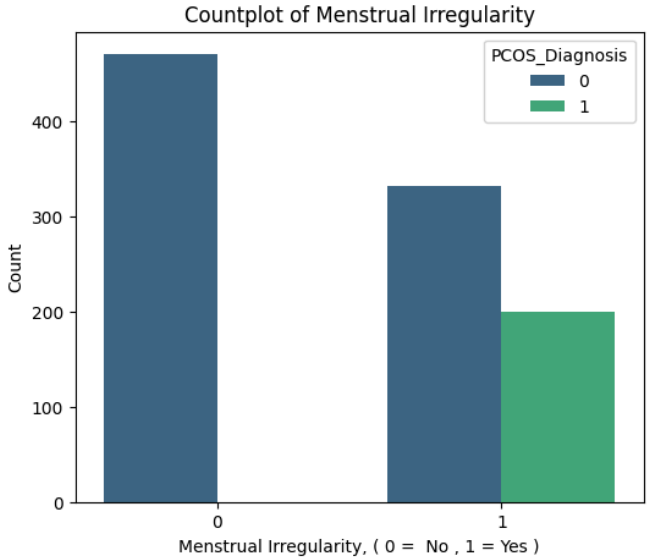
Visualisasi data digunakan untuk mendapatkan wawasan awal tentang pola dalam dataset. Beberapa metode yang digunakan termasuk histogram, boxplot, dan scatter plot.

Kode visualisasi data:

```
key_features = ['BMI','Testosterone_Level(ng/dL)']
for i in key_features:
  plt.figure(figsize=(6,5))
  sns.boxplot(x='PCOS_Diagnosis',y=i,data=df,palette='viridis')
  plt.xlabel('PCOS Diagnosis, ( 0 =  No , 1 = Yes )')
  plt.ylabel(i)
  plt.title("Boxplot of {}".format(i))
  plt.grid(linestyle = '--', linewidth = 0.5)
  plt.tight_layout()
  plt.show()
  print("\n")
```




```
plt.figure(figsize=(6,5))
sns.countplot(x='Menstrual_Irregularity',hue='PCOS_Diagnosis',data=df,palette='viridis')
plt.xlabel('Menstrual Irregularity, ( 0 =  No , 1 = Yes )')
plt.ylabel('Count')
plt.title("Countplot of Menstrual Irregularity")
```



4.3 Pola dan Insight

- **Distribusi Usia** hampir normal dengan rentang usia yang mencakup usia reproduksi wanita.
- **BMI menunjukkan adanya kecenderungan peningkatan risiko PCOS**, karena terdapat korelasi positif antara BMI dan diagnosis PCOS.
- **Menstrual Irregularity memiliki hubungan yang kuat dengan PCOS**, yang mendukung penelitian sebelumnya bahwa ketidakteraturan menstruasi adalah gejala utama PCOS.
- **Tingkat testosteron lebih tinggi pada individu dengan PCOS**, yang merupakan indikasi hormon androgen yang meningkat pada penderita kondisi ini.

5. Data Preparation

5.1 Penanganan Missing Values

Setelah dilakukan pemeriksaan terhadap dataset, tidak ditemukan missing values dalam dataset ini. Hal ini menunjukkan bahwa semua fitur memiliki nilai yang lengkap, sehingga tidak memerlukan imputasi atau penanganan lebih lanjut terkait missing values.

5.2 Encoding Variabel Kategorikal

Pada dataset ini, semua fitur bersifat numerik, sehingga tidak diperlukan encoding variabel kategorikal.

5.3 Normalisasi Data

Normalisasi atau standardisasi diperlukan untuk memastikan bahwa setiap fitur memiliki skala yang seragam, terutama untuk algoritma yang sensitif terhadap skala data. Dalam kasus ini, dilakukan standardisasi menggunakan **StandardScaler** dari `sklearn.preprocessing` agar data memiliki distribusi dengan mean 0 dan standar deviasi 1.

Kode normalisasi data:

```
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

5.4 Feature Selection

Feature selection bertujuan untuk mengurangi dimensi data dengan memilih fitur yang paling relevan terhadap target `PCOS_Diagnosis`. Berdasarkan heatmap korelasi, fitur **Menstrual_Irregularity** dan **Testosterone_Level(ng/dL)** memiliki korelasi yang cukup tinggi dengan diagnosis PCOS. Oleh karena itu, dipilih fitur-fitur berikut untuk analisis lebih lanjut:

- Age
- BMI
- Menstrual_Irregularity
- Testosterone_Level(ng/dL)
- Antral_Follicle_Count

Kode seleksi fitur:

```
X = df[['Age', 'BMI', 'Menstrual_Irregularity', 'Testosterone_Level(ng/dL)', 'Antral_Follicle_Count']]
y= df['PCOS_Diagnosis']
```

5.5 Alasan di Balik Keputusan Preprocessing

- **Missing Values** : Tidak ada missing values yang perlu ditangani.
- **Encoding** : Tidak diperlukan karena semua data bersifat numerik.
- **Standardisasi** : Dilakukan agar fitur memiliki skala yang sama, meningkatkan performa model.
- **Feature Selection** : Dilakukan untuk mengurangi dimensi data dan memilih fitur yang memiliki hubungan kuat dengan target.

6. Kesimpulan

Dalam proses eksplorasi dan persiapan data untuk analisis **PCOS Diagnosis**, beberapa langkah utama telah dilakukan untuk memastikan kualitas data yang optimal. Berikut adalah rangkuman dari setiap tahap yang telah dilakukan:

1. **Eksplorasi Data**
    - Statistik deskriptif menunjukkan distribusi data yang bervariasi pada fitur utama seperti **Age, BMI, Menstrual Irregularity, Testosterone Level, dan Antral Follicle Count**.
    - Tidak ditemukan **missing values**, sehingga tidak diperlukan proses imputasi.
    - Beberapa fitur memiliki rentang nilai yang cukup luas, sehingga normalisasi diperlukan agar skala data lebih seragam.
2. **Preprocessing Data**

- **Feature Selection :** Dipilih fitur yang memiliki hubungan kuat dengan diagnosis PCOS berdasarkan analisis korelasi yaitu `'Age', 'BMI', 'Menstrual_Irregularity', 'Testosterone_Level(ng/dL)', 'Antral_Follicle_Count'`
- **Standardisasi Data :** Dilakukan menggunakan **StandardScaler** untuk memastikan skala yang seragam di semua fitur yang digunakan.
- **Encoding :** Tidak diperlukan karena semua variabel dalam dataset berbentuk numerik.

3. **Insight dari Data**
   - **Menstrual Irregularity** dan **Testosterone Level** memiliki hubungan yang cukup kuat dengan **PCOS Diagnosis**, menunjukkan bahwa keduanya dapat menjadi indikator utama dalam identifikasi kondisi ini.
   - **BMI** juga menunjukkan distribusi yang mengarah ke kategori overweight, yang dapat dikaitkan dengan risiko PCOS.
   - Sebagian besar individu dalam dataset tidak terdiagnosis PCOS, sehingga terdapat ketidakseimbangan kelas yang perlu diperhatikan dalam pemodelan lebih lanjut.

Dataset telah diproses dengan baik melalui eksplorasi awal, normalisasi, dan seleksi fitur yang relevan. Langkah-langkah ini akan meningkatkan akurasi dan efektivitas model yang akan dibangun dalam tahap berikutnya