Tugas 2

Latihan Data Cleaning, Data Preprocessing, Modelling Data dan Evaluation Data

Dosen Pengampu : Adhy Rizaldy, S.Kom., M.Kom

Nama	: Fajrul Hidayat Amini
NIM	: 60900123019
Kelas	: A

JURUSAN SISTEM INFORMASI

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI ALAUDDIN MAKASSAR

2025

A. Import Data

dataset diabetes.csv dimuat ke dalam program menggunakan pandas. Tujuan tahap ini adalah untuk memastikan data berhasil terbaca dan memahami struktur awalnya—termasuk jumlah baris, kolom, dan apakah ada nilai nol atau kosong yang bisa mengganggu analisis.

```
 Tahap 1: Import Dataset
Jumlah baris dan kolom: (768, 9)

5 baris pertama:
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  DiabetesPedigreeFunction  Age  Outcome
0            6      148             72             35        0  33.6                     0.627   50        1
1            1       85             66             29        0  26.6                     0.351   31        0
2            8      183             64              0        0  23.3                     0.672   32        1
3            1       89             66             23       94  28.1                     0.167   21        0
4            0      137             40             35      168  43.1                     2.288   33        1

Jumlah nilai nol pada tiap kolom:
Pregnancies                 111
Glucose                       5
BloodPressure                35
SkinThickness               227
Insulin                     374
BMI                          11
DiabetesPedigreeFunction      0
Age                           0
Outcome                     500
dtype: int64
```

B. Data Cleaning

Tahap cleaning dilakukan karena beberapa kolom, seperti **Glucose**, **BloodPressure**, **SkinThickness**, **Insulin**, dan **BMI**, memiliki nilai **0**. Nilai nol ini kemudian **diganti dengan nilai median** dari masing-masing kolom agar distribusi data tetap seimbang.

```
Tahap 2: Data Cleaning
Nilai nol setelah cleaning:
Pregnancies                 111
Glucose                       0
BloodPressure                 0
SkinThickness                 0
Insulin                       0
BMI                           0
DiabetesPedigreeFunction      0
Age                           0
Outcome                     500
dtype: int64

Statistik deskriptif setelah cleaning:
       Pregnancies     Glucose  BloodPressure  ...  DiabetesPedigreeFunction         Age     Outcome
count   768.000000  768.000000     768.000000  ...                768.000000  768.000000  768.000000
mean      3.845052  121.656250      72.386719  ...                  0.471876   33.240885    0.348958
std       3.369578   30.438286      12.096642  ...                  0.331329   11.760232    0.476951
min       0.000000   44.000000      24.000000  ...                  0.078000   21.000000    0.000000
25%       1.000000   99.750000      64.000000  ...                  0.243750   24.000000    0.000000
50%       3.000000  117.000000      72.000000  ...                  0.372500   29.000000    0.000000
75%       6.000000  140.250000      80.000000  ...                  0.626250   41.000000    1.000000
max      17.000000  199.000000     122.000000  ...                  2.420000   81.000000    1.000000
```

C. Data Preprocessing

Deteksi Outlier (IQR) digunakan untuk melihat apakah ada nilai-nilai ekstrem di setiap kolom.

Standardisasi (Z-Score) agar semua fitur dinormalisasi agar memiliki rata-rata mendekati 0 dan standar deviasi mendekati 1.

```
Tahap 3: Deteksi Outlier (IQR)
Pregnancies: 4 outlier terdeteksi
Glucose: 0 outlier terdeteksi
BloodPressure: 14 outlier terdeteksi
SkinThickness: 35 outlier terdeteksi
Insulin: 49 outlier terdeteksi
BMI: 8 outlier terdeteksi
DiabetesPedigreeFunction: 29 outlier terdeteksi
Age: 9 outlier terdeteksi

Tahap 4: Pemisahan Fitur dan Target
Jumlah fitur: 8
Jumlah target (Outcome=1): 268 | (Outcome=0): 500

Tahap 5: Standardisasi (Z-Score)
Rata-rata fitur setelah standarisasi (mendekati 0): 0.0
Standar deviasi fitur setelah standarisasi (mendekati 1): 1.0
```

data juga dibagi menjadi tiga bagian yaitu, 60% untuk training, 20% untuk validation, 20% untuk testing.

```
=== Tahap 6: Pembagian Data ===
Training set: 460 data
Validation set: 154 data
Testing set: 154 data
```

D. Modelling Data

Dua model digunakan untuk perbandingan:

Random Forest Classifier, yaitu model berbasis kumpulan pohon keputusan (ensemble learning) yang cenderung lebih akurat pada data kompleks.

```
Training Model - Random Forest
Model Random Forest selesai dilatih.

[Evaluasi Random Forest - Validation Set]
Accuracy: 0.7208
Precision: 0.6222
Recall: 0.5185
F1-Score: 0.5657
ROC-AUC: 0.6743
Confusion Matrix:
 [[83 17]
 [26 28]]
```

Logistic Regression, yaitu model linear yang lebih sederhana dan mudah ditafsirkan.

```
Training Model - Logistic Regression
Model Logistic Regression selesai dilatih.

[Evaluasi Logistic Regression - Validation Set]
Accuracy: 0.7532
Precision: 0.66
Recall: 0.6111
F1-Score: 0.6346
ROC-AUC: 0.7206
Confusion Matrix:
 [[83 17]
 [21 33]]
```

E. Evaluasi Data

Tahap ini menilai seberapa baik model dalam memprediksi seseorang terkena diabetes. Metrik yang digunakan meliputi Accuracy, Precision, Recall, F1-Score, dan ROC-AUC.

```
Tahap 8: Evaluasi Akhir (Test Set)

[Random Forest - Test Set]
Accuracy: 0.7208
Precision: 0.5926
Recall: 0.6038
F1-Score: 0.5981
ROC-AUC: 0.693
Confusion Matrix:
 [[79 22]
 [21 32]]

[Logistic Regression - Test Set]
Accuracy: 0.7532
Precision: 0.6923
Recall: 0.5094
F1-Score: 0.587
ROC-AUC: 0.6953
Confusion Matrix:
 [[89 12]
 [26 27]]
PS D:\Kuliah\Semester 5\Data Mining\Teori\Tugas>
```

-