


Advanced Application of Artificial Intelligence and Digital Transformation



کاربرد پیشرفته هوش مصنوعی در تحول دیجیتال

5G


AI

Hasan Ghasemzadeh  
<http://wp.kntu.ac.ir/ghasemzadeh>

Adv. App. of AI and DT

K.N. Toosi University of Technology

Soft computing



داده‌ها

Adv. App. of AI and DT

2

K.N. TOOSI University of Technology

## آنچه در خصوص داده‌ها می‌بینیم

- تعریف و انواع داده
- پردازش داده
- مراحل پردازش داده
- مزایا و معایب
- کتابخانه‌های پردازش داده
- خواندن و ذخیره داده
- زمان، متن، انتقال و ادغام داده‌ها
- بررسی داده‌ها
- مثال و تمرین

Adv. App. of AI and DT

3

## تعریف و انواع داده

- **داده برچسب‌دار**
  - **داده بدون برچسب**
- داده‌های برچسب‌دار شامل یک برچسب یا متغیر هدف هستند که مدل سعی در پیش‌بینی آن متغیر دارد، در حالی که داده‌های بدون برچسب هیچ برچسب یا متغیر هدفی ندارند.

- **داده‌های عددی**
  - **داده‌های غیر عددی**
  - **داده‌های ترتیبی**
- داده‌های عددی مانند سن یا درآمد، داده‌هایی هستند که می‌توان آن‌ها را اندازه‌گیری و مرتب کرد. داده‌های غیر عددی مانند جنسیت یا نام میوه‌ها، گروه یا طبقه مقادیر را نشان می‌دهند. داده‌های ترتیبی به متغیری نامی اشاره دارد که دسته‌های آن در یک لیست دارای ترتیب قرار می‌گیرند. سایزهای لباس مانند کوچک، متوسط و بزرگ و یا اندازه‌گیری رضایت مشتری در مقیاسی از "کاملاً ناراضی" تا "کاملاً راضی" مثال‌هایی از این داده‌ها هستند.

Adv. App. of AI and DT

4

## تعریف و انواع داده

- داده‌های آموزش (Training Data)
- داده اعتبارسنجی (Validation Data)
- داده‌های تست (Test Data)

**داده‌های آموزش:** بخشی از داده است که برای آموزش مدل استفاده می‌شود که مدل هم داده‌های ورودی و هم داده‌های خروجی را می‌بیند و از آنها یاد می‌گیرد.

**داده اعتبارسنجی:** بخشی از داده‌ها که برای ارزیابی مکرر مدل استفاده می‌شود و همراه با مجموعه داده آموزش می‌تواند هایپرپارامترها (پارامترهایی که در ابتدا و پیش از شروع یادگیری مدل تنظیم می‌شوند) را بهبود دهد. این داده‌ها در حین آموزش مدل ایفای نقش می‌کند.

**داده‌های تست:** مجموعه داده‌های تست برای ارزیابی عملکرد مدل استفاده می‌شود.

باید داده‌ها را به طور تصادفی به زیرمجموعه‌های فوق تقسیم کنیم. داده‌های موجود در هر یک از این زیرمجموعه‌ها، باید نمایانگر مجموعه داده‌ها در حوزه مورد بررسی بوده و قادر به بازتولید الگوها و روابط موجود در مجموعه داده‌ها باشند

Adv. App. of AI and DT

5

## تعریف و انواع داده

**داده:** می‌تواند هر نکته، مقدار، متن، صوت یا تصویر پردازش نشده‌ای باشد که هنوز تفسیر و تحلیل نشده است.

مهم‌ترین بخش در تجزیه و تحلیل داده، یادگیری ماشین و هوش مصنوعی داده است. بدون داده، نمی‌توانیم هیچ مدلی را آموزش دهیم

**اطلاعات:** داده‌ای است که دستکاری و تفسیر شده تا برای کاربران، نتایج ملموس و معناداری داشته باشد.

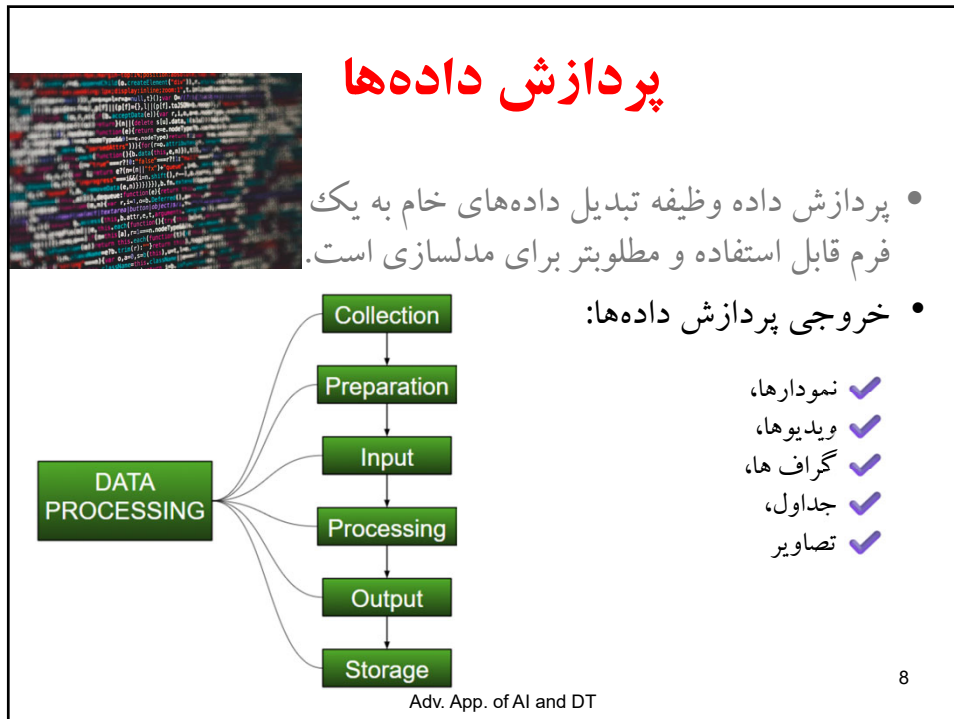
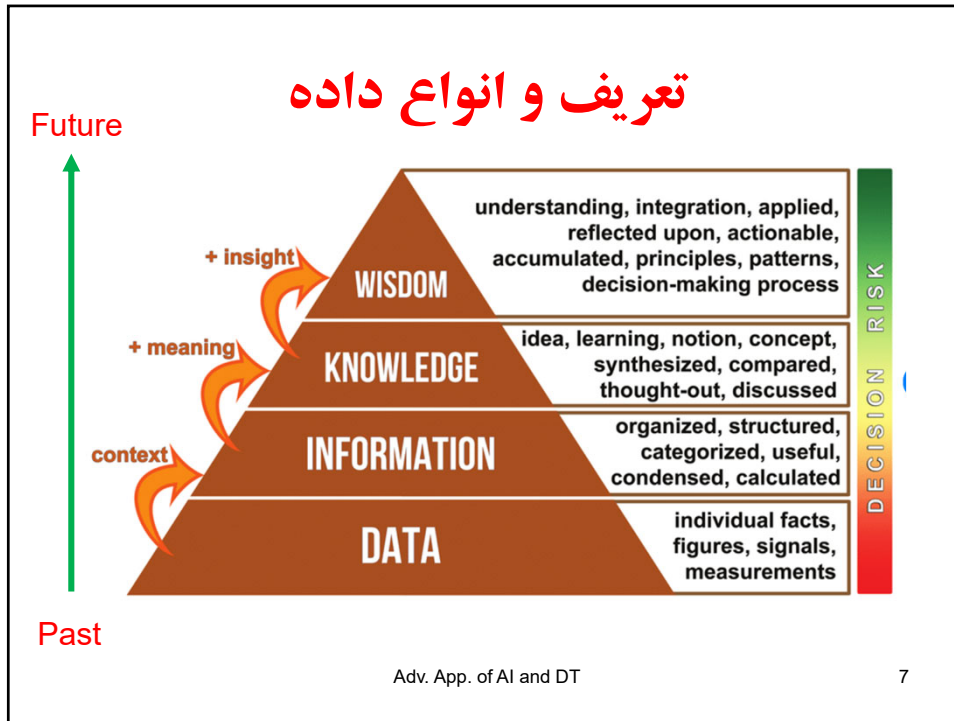
**دانش:** ترکیبی از اطلاعات استنتاجی، تجربیات، یادگیری و بینش‌ها است. دانش منجر به آگاهی بخشی یا ساخت مفاهیم برای یک فرد یا سازمان می‌شود.

**خود:** به معنی استفاده درست از دانش برای مدیریت کردن و حل مشکلات تعریف شده داخل سیستم‌های اطلاعاتی می‌باشد.



Adv. App. of AI and DT

6



## مراحل پردازش داده‌ها

### ۱- جمع آوری داده‌ها:

این فرآیند به معنای گردآوری داده از منابع مختلف مثل حسگرها، پایگاه‌های داده یا سیستم‌های دیگر است. داده‌ها ممکن است ساختاردار یا بدون ساختار باشند و ممکن است به اشکال مختلفی نظیر متن، تصاویر یا صدا تولید شوند.

### ۲- آماده سازی

در این مرحله، داده‌ها تمیز می‌شوند، فیلتر می‌شوند و به گونه‌ای تغییر شکل می‌یابند که برای تحلیل بعدی مناسب باشند. این مرحله شامل حذف مقادیر گم‌شده، مقیاس‌دهی یا نرمال‌سازی داده‌ها یا تبدیل آن به فرمت دیگر می‌شود.

9

Adv. App. of AI and DT

## مراحل پردازش داده‌ها

### ۳- ورودی داده‌ها

در این مرحله داده‌ها به فرمت مناسب برای ورودی به برنامه یادگیری ماشین آماده می‌شوند. داده‌های توصیفی، تصویری و ویدئویی بایستی برای ماشین قابل فهم شوند.

### ۴- تجزیه، تحلیل و تفسیر داده

داده‌ها با استفاده از تکنیک‌های مختلفی نظیر تحلیل آماری، الگوریتم‌های یادگیری ماشین یا تجسم داده‌ها تحلیل می‌شوند. هدف این مرحله استخراج اطلاعات یا دانش از داده است. این مرحله شامل تفسیر نتایج تحلیل داده و استنتاج‌هایی بر اساس دانش به دست آمده است. این مرحله ممکن است شامل ارائه نتایج به شیوه‌ای واضح و مختصر مثل گزارش‌ها، داشبوردها یا تجسم‌های دیگر باشد.

10

Adv. App. of AI and DT

## مراحل پردازش داده‌ها

### ۵- خروجی (تجسم و گزارش دهی داده)

نتایج تحلیل داده به صورتی ارائه می‌شوند که به راحتی قابل فهم برای کاربر باشند این مرحله شامل ایجاد تجسم‌ها، گزارش‌ها یا داشبوردهایی باشد که نتایج کلیدی و روندهای داده را نشان می‌دهند.

### ۶- ذخیره و مدیریت داده

داده‌ها باید به گونه‌ای ذخیره و مدیریت شوند که امن و به راحتی قابل دسترسی باشند. این امر ممکن است شامل ذخیره داده در پایگاه‌داده، ذخیره در فضای ابری یا سیستم‌های دیگر باشد. این مرحله می‌تواند شامل استفاده از راهکارهای پشتیبان‌گیری و بازیابی برای محافظت در برابر از دست رفتن داده‌ها باشد.

11

Adv. App. of AI and DT

## معایب و مزایای پردازش داده‌ها

### مزایای پردازش داده

- ✓ افزایش کارایی مدل: با تمیز کردن و تبدیل داده‌ها، مدل یادگیری ماشین بهتر عمل می‌کند.
- ✓ نمایش مناسب‌تر داده: با پردازش، داده‌ها به شکلی تبدیل می‌شوند که روابط و الگوهای موجود در آنها بهتر به نمایش درآیند، و این باعث می‌شود ماشین بهتر و آسان‌تر یاد بگیرد.
- ✓ افزایش دقت: با اطمینان از صحت و یکپارچگی داده‌ها، دقت مدل یادگیری بهبود می‌یابد.

12

Adv. App. of AI and DT

## معایب و مزایای پردازش داده‌ها

### معایب پردازش داده

- ✓ زمان‌بر: بخصوص در مجموعه‌های داده بزرگ، پردازش می‌تواند طول بکشد.
- ✓ خطر خطا: در فرآیند پردازش، ممکن است اطلاعاتی از بین بروند یا خطاهای جدیدی ایجاد شود.
- ✓ درک ناقص از داده: گاهی داده‌های تبدیل شده، تمام جنبه‌ها و روابط موجود در داده اصلی را نمایان نمی‌کنند.

Adv. App. of AI and DT

13

## برنامه نویسی

برای زبان برنامه‌نویسی پایتون از نرم افزار `visual studio (vs) code`، استفاده می‌کنیم.



نرم افزار سبک و راحتی برای استفاده است زبان‌های مختلف نظیر `Python`، `Java`، `C++`، `CSS`، `Go` و `Dockerfile` را پشتیبانی می‌کند.

Adv. App. of AI and DT

14

## کتابخانه‌های پردازش داده‌ها

برای پردازش داده در یادگیری ماشین در زبان برنامه‌نویسی پایتون، ابزارها و کتابخانه‌های متعددی نظیر Pandas وجود دارد.

**Pandas (Panel Data and "Python Data Analysis)**

پانداس یک کتابخانه نرم‌افزاری نوشته شده برای زبان برنامه‌نویسی پایتون برای دستکاری، پاکسازی و تجزیه و تحلیل داده‌ها است.

```
C:>pip install pandas
```

در خط فرمان

```
import pandas as pd
```

در برنامه

Adv. App. of AI and DT

15

## کتابخانه‌های پردازش داده‌ها

**NumPy (Numerical Python)**: نام پای برای محاسبات عددی متنوعی در پایتون به کار می‌رود. سرعت محاسبات به کمک آرایه‌ها بالاست و علاوه بر آن، توابع این پکیج در ساخت پکیج‌های محاسباتی دیگر مورد استفاده قرار گرفته است.

**SciPy**: سای پای برای انجام محاسبات علمی و مهندسی مورد استفاده قرار می‌گیرد. کتابخانه‌ی SciPy بر مبنای کتابخانه‌ی NumPy است و امکان کار با آرایه‌های n بعدی را فراهم می‌کند.

```
import numpy as np
import scipy
```

Adv. App. of AI and DT

16

## کتابخانه‌های پردازش داده‌ها

sklearn (SciKit-learn) : یک کتابخانه متن‌باز برای داده‌کاوی در زبان برنامه‌نویسی پایتون است.

Matplotlib: مت‌پلات‌لیب برای نمودارها و گرافیک‌های مختلف در پایتون است.

Matplotlib.pyplot: مجموعه‌ای از توابع است که شبیه نرم‌افزار متلب کار می‌کند.

Seaborn (Numerical Python): سی‌بورن برای تجسم و رسم نمودار داده‌ها در پایتون است که توسعه داده شده مت‌پلات‌لیب است.

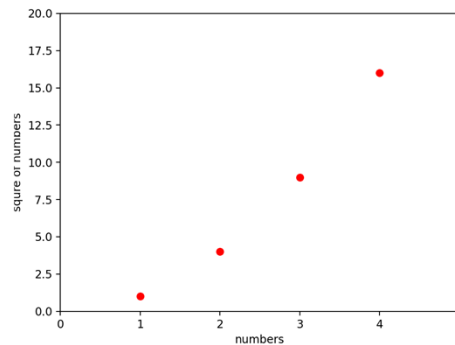
Adv. App. of AI and DT

17

## کتابخانه‌های پردازش داده‌ها

با کتابخانه‌های ارایه شده تمرین انجام دهید  
مثال ۱:

```
import matplotlib.pyplot as plt
plt.plot([1, 2, 3, 4], [1, 4, 9, 16], 'ro')
plt.axis((0, 5, 0, 20))
plt.xlabel('numbers')
plt.ylabel('square of numbers')
plt.show()
```



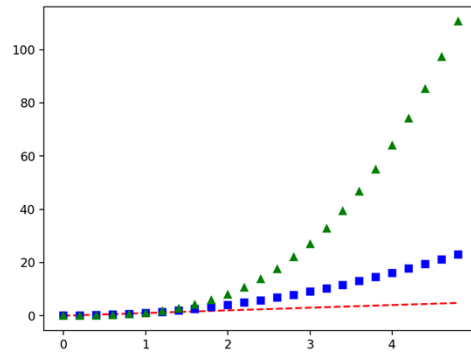
18

Adv. App. of AI and DT

## کتابخانه‌های پردازش داده‌ها

مثال ۲:

```
import matplotlib.pyplot as plt
import numpy as np
# evenly sampled time at 200ms intervals
t = np.arange(0., 5., 0.2)
# red dashes, blue squares and green triangles
plt.plot(t, t, 'r--', t, t**2, 'bs', t, t**3, 'g^')
plt.show()
```



Adv. App. of AI and DT

19

## کتابخانه‌های پردازش داده‌ها

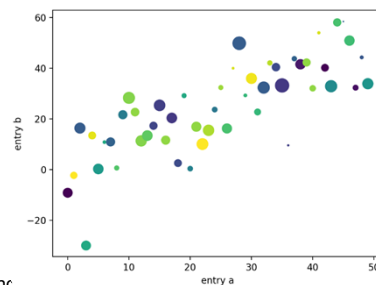
مثال ۳:

```
import matplotlib.pyplot as plt
import numpy as np
data = {'a': np.arange(50),
        'c': np.random.randint(0, 50, 50),
        'd': np.random.randn(50)}
data['b'] = data['a'] + 10 * np.random.randn(50)
data['d'] = np.abs(data['d']) * 100
```

```
plt.scatter('a', 'b', c='c', s='d', data=data)
plt.xlabel('entry a')
plt.ylabel('entry b')
plt.show()
```

Table of data

color Size



Adv. App. of AI and DT

## خواندن فایل داده‌ها

```
import pandas as pd

# Load data from a CSV file
df = pd.read_csv('data.csv')

# Load data from an Excel file
df = pd.read_excel('data.xlsx')

# Load data from a JSON file
df = pd.read_json('data.json')

# Load data from a SQL database
import sqlite3
conn = sqlite3.connect('database.db')
df = pd.read_sql_query("SELECT * FROM table_name",
conn)
```

خواندن انواع فایل داده

Adv. App. of AI and DT

21

## ذخیره فایل داده‌ها

```
# Save DataFrame to a CSV file
df.to_csv('processed_data.csv', index=False)

# Save DataFrame to an Excel file
df.to_excel('processed_data.xlsx', index=False)
```

ذخیره انواع فایل داده

پارامتر `index=False` باعث می‌شود که اندیس‌های DataFrame در فایل CSV ذخیره نشوند.

Adv. App. of AI and DT

22

## زمان در داده‌ها

زمان در فایل داده

```
# Convert column to datetime
df['date_column'] = pd.to_datetime(df['date_column'])

# Convert column to datetime
df['date_column'] = pd.to_datetime(df['date_column'])

# Extract year, month, day, etc.
df['year'] = df['date_column'].dt.year
df['month'] = df['date_column'].dt.month
df['day'] = df['date_column'].dt.day

# Filter data based on date range
df_filtered = df[(df['date_column'] >= '2023-01-01') &
(df['date_column'] <= '2023-12-31')]
```

Adv. App. of AI and DT

23

## متن در داده‌ها

کار کردن با داده های متنی

```
# Convert text to lowercase
df['text_column'] = df['text_column'].str.lower()

# Remove punctuation
df['text_column'] = df['text_column'].str.replace(r'[^\w\s]', '')

# Tokenize text
df['tokens'] = df['text_column'].str.split()

# Count word frequency
word_counts =
df['text_column'].str.split(expand=True).stack().value_counts()
```

Adv. App. of AI and DT

24

## انتقال داده‌ها

```
# Apply a function to a column
df['new_column'] = df['column_name'].apply(lambda x: x * 2)

# Filter rows based on a condition
df_filtered = df[df['column_name'] > 10]

# Group by a column and aggregate
df_grouped = df.groupby('category_column').agg({'numeric_column':
'mean'})

# Sort DataFrame by a column
df_sorted = df.sort_values(by='column_name', ascending=False)

# Pivot table
df_pivot = df.pivot_table(index='row_column',
columns='column_column', values='value_column', aggfunc='mean')
```

در دستور فوق اگر داده های تکراری وجود داشته باشد تجمیع آنها را با میانگین گیری در نظر می گیرد

Adv. App. of AI and DT

25

## ادغام داده‌ها

```
# Merge two DataFrames on a common column
df_merged = pd.merge(df1, df2, on='common_column', how='inner')
```

الحاق (چسباندن) قائم داده ها

```
# Concatenate DataFrames vertically
df_concat = pd.concat([df1, df2], axis=0)
```

الحاق (چسباندن) افقی داده ها

```
# Concatenate DataFrames horizontally
df_concat = pd.concat([df1, df2], axis=1)
```

Adv. App. of AI and DT

26

## بررسی داده‌ها

✓ خلاصه دستورهای بررسی داده‌ها

```
# Check the shape of the DataFrame (rows, columns)
print(df.shape)

# Display the first 5 rows of the DataFrame
print(df.head())

# Display the last 5 rows of the DataFrame
print(df.tail())

# Get a summary of the DataFrame
print(df.info())

# Get statistical summary of numerical columns
print(df.describe())

# Check for missing values
print(df.isnull().sum())
```

Adv. App. of AI and DT

27

## فراخوانی کتابخانه‌ها

فراخوانی کتابخانه‌ها

```
# importing libraries
import pandas as pd
import scipy
import numpy as np
from sklearn.preprocessing import MinMaxScaler
import seaborn as sns
import matplotlib.pyplot as plt
```

Adv. App. of AI and DT

28

## بررسی داده‌ها

2 data process diabet.py      مثال ۴: بررسی فایل آزمایش‌های افراد (قند حاملگی)  
خواندن یک فایل

نوع فایل  
CSV (Comma Separated Values)

```
# Load the dataset
df = pd.read_csv('diabetes.csv')
print(df.head())
```

✓ بررسی داده‌ها - کلیات شکل داده‌ها

```
# Check the shape of the DataFrame (rows, columns)
print(df.shape)
```

(768, 9)

Adv. App. of AI and DT

29

## بررسی داده‌ها

```
# Display the first 5 rows of the DataFrame
print(df.head())
```

✓ بررسی داده‌ها - کلیات

	Pregnancies	Glucose	BloodPressure	SkinThickness	...	BMI		
	DiabetesPedigreeFunction	Age	Outcome					
0	6	NaN	72	35 ...	33.6	0.627	50	1
1	1	85.0	66	29 ...	26.6	0.351	31	0
2	8	183.0	64	0 ...	23.3	0.672	32	1
3	1	89.0	66	23 ...	NaN	0.167	21	0
4	0	137.0	40	35 ...	NaN	2.288	33	1

[5 rows x 9 columns]

Adv. App. of AI and DT

30

## بررسی داده‌ها

```
# Display the last 5 rows of the DataFrame
print(df.tail())
```

✓ بررسی داده‌ها - کلیات

```
Pregnancies  Glucose  BloodPressure  SkinThickness  ...  BMI
DiabetesPedigreeFunction  Age  Outcome
763      10  101.0      76      48 ... 32.9      0.171  63      0
764       2  122.0      70      27 ... 36.8      0.340  27      0
765       5  121.0      72      23 ... 26.2      0.245  30      0
766       1  126.0      60       0 ... 30.1      0.349  47      1
767       1   93.0      70      31 ... 30.4      0.315  23      0
```

[5 rows x 9 columns]

Adv. App. of AI and DT

31

## بررسی داده‌ها

```
# Get a summary of the DataFrame
print(df.info())
```

✓ بررسی داده‌ها - کلیات

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Pregnancies           768 non-null   int64
1   Glucose               767 non-null   float64
2   BloodPressure         768 non-null   int64
3   SkinThickness         768 non-null   int64
4   Insulin               768 non-null   int64
5   BMI                   766 non-null   float64
6   DiabetesPedigreeFunction 768 non-null   float64
7   Age                   768 non-null   int64
8   Outcome               768 non-null   int64
dtypes: float64(3), int64(6)
memory usage: 54.1 KB
```

Adv. App. of AI and DT

32

## بررسی داده‌ها

✓ بررسی داده‌ها: تجزیه و تحلیل آماری

```
# Get statistical summary of numerical columns
print(df.describe())
```

	Pregnancies	Glucose	BloodPressure ...	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	767.000000	768.000000 ...	768.000000	768.000000	768.000000
mean	3.845052	120.859192	69.105469 ...	0.471876	33.240885	0.348958
std	3.369578	31.978468	19.355807 ...	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000 ...	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000 ...	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000 ...	0.372500	29.000000	0.000000
75%	6.000000	140.000000	80.000000 ...	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000 ...	2.420000	81.000000	1.000000

[8 rows x 9 columns]

Adv. App. of AI and DT

33

## بررسی داده‌ها

✓ بررسی داده‌ها - مقادیر گمشده

```
# Check for missing values
print(df.isnull().sum())
```

Pregnancies	0
Glucose	1
BloodPressure	0
SkinThickness	0
Insulin	0
BMI	2
DiabetesPedigreeFunction	0
Age	0
Outcome	0
dtype:	int64

Adv. App. of AI and DT

34

## تمرین برنامه نویسی

تمرین دوم: یک برنامه به زبان پایتون بنویسید که یک فایل داده را بگیرد و عملیات زیر را انجام دهد

Depth (m)	Temperature ( c )
0.0	4.3
2.7	4.3
5.4	4.3
8.2	4.4
10.9	4.4
13.7	4.4
16.4	4.5
19.2	4.5
21.9	4.6
24.7	4.6
27.4	4.7
30.2	4.7
32.9	4.8
35.7	4.9

۱- تعیین تعداد داده‌ها و مدیریت فیلهای خالی

۲- تعیین ماکزیمم و مینیمم داده‌ها

۴- تعیین میانگین و میانه داده‌ها

۵- مرتب نمودن صعودی داده‌ها

۶- نرمال کردن داده‌ها

تمرین در کوئرا تا هفته بعد حل شود