

Amine Ait Laamim

Importation des bibliothèques

```
In [57]: import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import BaggingClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix, classification_report
import joblib
```

Chargement des données

```
In [58]: df = pd.read_csv("synthetic_heart_disease_dataset.csv")
```

Exploration des données

```
In [59]: df.shape
```

```
Out[59]: (50000, 21)
```

```
In [60]: df.describe()
```

```
Out[60]:
```

	Age	Weight	Height	BMI	Hypertension	Diabetes	Hyperlipidemia	Family_History	Previous_Heart_Attack	Systolic_BP	Diastolic_BP	Heart_Rate	Blood_Sugar_Fasting	Cholesterol_Total	Heart_Disease
count	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000	50000.000000
mean	54.46406	84.547520	174.460000	28.984284	0.299620	0.199260	0.251660	0.400500	0.059280	139.295950	89.528800	84.449560	124.493020	224.556360	0.6436460
std	14.43809	12.03157	14.420379	6.367494	0.458096	0.399448	0.433971	0.490005	0.299041	23.083544	17.258063	14.491325	31.691507	43.157467	0.498668
min	30.00000	50.00000	150.00000	18.00000	0.00000	0.00000	0.00000	0.00000	0.00000	100.00000	60.00000	60.00000	70.00000	150.00000	0.00000
25%	42.00000	79.00000	174.02357	23.50000	0.00000	0.00000	0.00000	0.00000	0.00000	119.00000	75.00000	72.00000	97.00000	187.00000	0.00000
50%	54.00000	85.00000	174.00000	29.00000	0.00000	0.00000	0.00000	0.00000	0.00000	139.00000	90.00000	85.00000	125.00000	225.00000	0.00000
75%	67.00000	102.00000	187.00000	34.50000	1.00000	0.00000	1.00000	1.00000	0.00000	159.00000	104.00000	97.00000	152.00000	262.00000	1.00000
max	79.00000	119.00000	199.00000	40.00000	1.00000	1.00000	1.00000	1.00000	1.00000	179.00000	119.00000	109.00000	179.00000	299.00000	1.00000

```
In [61]: df.info()
```

```
Out[61]:
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype  
 --- 
  0   Age             50000 non-null   int64  
  1   Gender          50000 non-null   object 
  2   Height          50000 non-null   int64  
  3   Weight          50000 non-null   int64  
  4   BMI             50000 non-null   float64 
  5   Smoking          50000 non-null   object 
  6   Alcohol_Contake 20893 non-null  object 
  7   Physical_Activity 50000 non-null  object 
  8   Diet             50000 non-null   object 
  9   Stress_Level    50000 non-null   object 
  10  Hypertension     50000 non-null   int64  
  11  Diabetes         50000 non-null   int64  
  12  Hyperlipidemia  50000 non-null   int64  
  13  Family_History  50000 non-null   int64  
  14  Previous_Heart_Attack 50000 non-null   int64  
  15  Systolic_BP      50000 non-null   int64  
  16  Diastolic_BP    50000 non-null   int64  
  17  Heart_Rate       50000 non-null   int64  
  18  Blood_Sugar_Fasting 50000 non-null   int64  
  19  Cholesterol_Total 50000 non-null   int64  
  20  Heart_Disease    50000 non-null   int64  
dtypes: float64(13), int64(14), object(2)
memory usage: 8.8+ MB
```

```
In [62]: df.head()
```

```
Out[62]:
```

	Age	Gender	Weight	Height	BMI	Smoking	Alcohol_Contake	Physical_Activity	Diet	Stress_Level	...	Diabetes	Hyperlipidemia	Family_History	Previous_Heart_Attack	Systolic_BP	Diastolic_BP	Heart_Rate	Blood_Sugar_Fasting	Cholesterol_Total	Heart_Disease	
0	48	Male	78	157	26.4	Never	NaN	Sedentary	Healthy	Medium	...	0	1	1	1	0	104	99	71	165	200	0
1	35	Female	73	163	33.0	Never	Low	Active	Average	High	...	0	1	1	0	0	111	72	60	145	206	0
2	79	Female	88	182	32.3	Never	NaN	Moderate	Average	Medium	...	0	0	1	0	0	116	102	78	148	208	0
3	75	Male	106	171	37.4	Never	Moderate	Moderate	Average	Low	...	0	1	0	0	0	171	92	109	105	290	1
4	34	Female	65	191	18.5	Current	NaN	Sedentary	Healthy	Low	...	1	0	0	0	0	164	67	108	116	220	1

5 rows × 21 columns

```
In [63]: df.columns
```

```
Out[63]:
```

```
Index(['Age', 'Gender', 'Weight', 'Height', 'BMI', 'Smoking', 'Alcohol_Contake', 'Physical_Activity', 'Diet', 'Stress_Level', 'Hypertension', 'Diabetes', 'Hyperlipidemia', 'Family_History', 'Previous_Heart_Attack', 'Systolic_BP', 'Diastolic_BP', 'Heart_Rate', 'Blood_Sugar_Fasting', 'Cholesterol_Total', 'Heart_Disease'], dtype='object')
```

```
In [64]: df['Heart_Disease'].value_counts()
```

```
Out[64]:
```

Heart_Disease	count
0	26627
1	23373

```
Name: count, dtype: int64
```

```
In [65]: df.isna().sum()
```

```
Out[65]:
```

Age	Gender	Weight	Height	BMI	Smoking	Alcohol_Contake	Physical_Activity	Diet	Stress_Level	...	Diabetes	Hyperlipidemia	Family_History	Previous_Heart_Attack	Systolic_BP	Diastolic_BP	Heart_Rate	Blood_Sugar_Fasting	Cholesterol_Total	Heart_Disease		
0	48	Male	78	157	26.4	Never	NaN	Sedentary	Healthy	Medium	...	0	1	1	1	0	104	99	71	165	200	0
1	35	Female	73	163	33.0	Never	Low	Active	Average	High	...	0	1	1	0	0	111	72	60	145	206	0
2	79	Female	88	182	32.3	Never	NaN	Moderate	Average	Medium	...	0	0	1	0	0	116	102	78	148	208	0
3	75	Male	106	171	37.4	Never	Moderate	Moderate	Average	Low	...	0	1	0	0	0	171	92	109	105	290	1
4	34	Female	65	191	18.5	Current	NaN	Sedentary	Healthy	Low	...	1	0	0	0	0	164	67	108	116	220	1

5 rows × 21 columns

```
In [66]: df['Alcohol_Contake'].isna()
```

```
Out[66]:
```

Alcohol_Contake	count
0	48
1	Gender
2	Height
3	BMI
4	Smoking
5	Alcohol_Contake
6	Physical_Activity
7	Diet
8	Stress_Level
9	Diabetes
10	Hyperlipidemia
11	Family_History
12	Previous_Heart_Attack
13	Systolic_BP
14	Diastolic_BP
15	Heart_Rate
16	Blood_Sugar_Fasting
17	Cholesterol_Total
18	Heart_Disease

```
dtypes: int64(14), float64(7), object(1)
```

```
memory usage: 8.4+ MB
```

```
In [67]: df[df['Alcohol_Contake'].isna()]
```

```
Out[67]:
```

Gender	Weight	Height	BMI	Smoking	Alcohol_Contake	Physical_Activity	Diet	Stress_Level	...	Diabetes	Hyperlipidemia	Family_History	Previous_Heart_Attack	Systolic_BP	Diastolic_BP	Heart_Rate	Blood_Sugar_Fasting	Cholesterol_Total	Heart_Disease			
0	48	Male	78	157	26.4	Never	NaN	Sedentary	Healthy	Medium	...	0	1	1	1	0	104	99	71	165	200	0
1	35	Female	73	163	33.0	Never	Low	Active	Average	High	...	0	1	1	0	0	111	72	60	145	206	0
2	79	Female	88	182	32.3	Never	NaN	Moderate	Average	Medium	...	0	0	1	0	0	116	102	78	148	208	0
3	75	Male	106	171	37.4	Never	Moderate	Moderate	Average	Low	...	0	1	0	0	0	171	92	109	105	290	1
4	34	Female	65	191	18.5	Current	NaN	Sedentary	Healthy	Low	...	1	0	0	0	0	164	67	108	116	220	1

5 rows × 21 columns

```
In [68]: df.isna().sum()
```

```
Out[68]:
```

Age	Gender	Weight	Height	BMI	Smoking	Alcohol_Contake	Physical_Activity	Diet	Stress_Level	...	Diabetes	Hyperlipidemia	Family_History	Previous_Heart_Attack	Systolic_BP	Diastolic_BP	Heart_Rate	Blood_Sugar_Fasting	Cholesterol_Total	Heart_Disease		
0	48	Male	78	157	26.4	Never	NaN	Sedentary	Healthy	Medium	...	0	1	1	1	0	104	99	71	165	200	0
1	35	Female	73	163	33.0	Never	Low	Active	Average	High	...	0	1	1	0	0	111	72	60	145	206	0
2	79	Female	88	182	32.3	Never	NaN	Moderate	A													