# Project 1 : Cardiovascular Disease Prediction

## Perform data pre-processing operation

In [10]:
```
import pandas as pd
```

In [14]:
```
# Load the dataset
df1=pd.read_csv(r"C:\Users\bhava\Downloads\cardio_train.csv")
df1
```

Out[14]:

| | id;age;gender;height;weight;ap_hi;ap_lo;cholesterol;gluc;smoke;alco;active;cardio |
|---|---|
| 0 | 0;18393;2;168;62.0;110;80;1;1;0;0;1;0 |
| 1 | 1;20228;1;156;85.0;140;90;3;1;0;0;1;1 |
| 2 | 2;18857;1;165;64.0;130;70;3;1;0;0;0;1 |
| 3 | 3;17623;2;169;82.0;150;100;1;1;0;0;1;1 |
| 4 | 4;17474;1;156;56.0;100;60;1;1;0;0;0;0 |
| ... | ... |
| 69995 | 99993;19240;2;168;76.0;120;80;1;1;1;0;1;0 |
| 69996 | 99995;22601;1;158;126.0;140;90;2;2;0;0;1;1 |
| 69997 | 99996;19066;2;183;105.0;180;90;3;1;0;1;0;1 |
| 69998 | 99998;22431;1;163;72.0;135;80;1;2;0;0;0;1 |
| 69999 | 99999;20540;1;170;72.0;120;80;2;1;0;0;1;0 |

70000 rows × 1 columns

In [18]:
```
# Display the first few rows of the dataset and its summary info to understand its structure
df1_info = df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70000 entries, 0 to 69999
Data columns (total 1 columns):
 #   Column                                                                      Non-Null Count  Dtype 
---  ------                                                                      --------------  ----- 
 0   id;age;gender;height;weight;ap_hi;ap_lo;cholesterol;gluc;smoke;alco;active;cardio  70000 non-null  object
dtypes: object(1)
memory usage: 547.0+ KB
```

In [23]:
```
df1_head = df1.head()
df1_head
```

Out[23]:

| | id;age;gender;height;weight;ap_hi;ap_lo;cholesterol;gluc;smoke;alco;active;cardio |
|---|---|
| 0 | 0;18393;2;168;62.0;110;80;1;1;0;0;1;0 |
| 1 | 1;20228;1;156;85.0;140;90;3;1;0;0;1;1 |
| 2 | 2;18857;1;165;64.0;130;70;3;1;0;0;0;1 |
| 3 | 3;17623;2;169;82.0;150;100;1;1;0;0;1;1 |
| 4 | 4;17474;1;156;56.0;100;60;1;1;0;0;0;0 |

In [27]:
```
df1.tail()
```

Out[27]:

| | id;age;gender;height;weight;ap_hi;ap_lo;cholesterol;gluc;smoke;alco;active;cardio |
|---|---|
| 69995 | 99993;19240;2;168;76.0;120;80;1;1;1;0;1;0 |
| 69996 | 99995;22601;1;158;126.0;140;90;2;2;0;0;1;1 |
| 69997 | 99996;19066;2;183;105.0;180;90;3;1;0;1;0;1 |
| 69998 | 99998;22431;1;163;72.0;135;80;1;2;0;0;0;1 |
| 69999 | 99999;20540;1;170;72.0;120;80;2;1;0;0;1;0 |

In [29]:
```
# Displaying the updated info and first few rows
df1_info = df1.info()
df1_head = df1.head()

df1_info, df1_head
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70000 entries, 0 to 69999
Data columns (total 1 columns):
 #   Column                                                                      Non-Null Count  Dtype 
---  ------                                                                      --------------  ----- 
 0   id;age;gender;height;weight;ap_hi;ap_lo;cholesterol;gluc;smoke;alco;active;cardio  70000 non-null  object
dtypes: object(1)
memory usage: 547.0+ KB
```

Out[29]:
```
(None,
   id;age;gender;height;weight;ap_hi;ap_lo;cholesterol;gluc;smoke;alco;active;cardio
  0             0;18393;2;168;62.0;110;80;1;1;0;0;1;0
  1             1;20228;1;156;85.0;140;90;3;1;0;0;1;1
  2             2;18857;1;165;64.0;130;70;3;1;0;0;0;1
  3            3;17623;2;169;82.0;150;100;1;1;0;0;1;1
  4             4;17474;1;156;56.0;100;60;1;1;0;0;0;0                          )
```

In [35]:
```
# Reloading the dataset with the correct delimiter `;`
df1= pd.read_csv(r"C:\Users\bhava\Downloads\cardio_train.csv", delimiter=';')

# Displaying the updated info and first few rows
df1_info = df1.info()
df1_head = df1.head()

df1_info, df1_head
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70000 entries, 0 to 69999
Data columns (total 13 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   id           70000 non-null  int64  
 1   age          70000 non-null  int64  
 2   gender       70000 non-null  int64  
 3   height       70000 non-null  int64  
 4   weight       70000 non-null  float64
 5   ap_hi        70000 non-null  int64  
 6   ap_lo        70000 non-null  int64  
 7   cholesterol  70000 non-null  int64  
 8   gluc         70000 non-null  int64  
 9   smoke        70000 non-null  int64  
 10  alco         70000 non-null  int64  
 11  active       70000 non-null  int64  
 12  cardio       70000 non-null  int64  
dtypes: float64(1), int64(12)
memory usage: 6.9 MB
```

Out[35]:
```
(None,
     id    age  gender  height  weight  ap_hi  ap_lo  cholesterol  gluc  smoke  \
  0   0  18393       2     168    62.0    110     80            1     1      0   
  1   1  20228       1     156    85.0    140     90            3     1      0   
  2   2  18857       1     165    64.0    130     70            3     1      0   
  3   3  17623       2     169    82.0    150    100            1     1      0   
  4   4  17474       1     156    56.0    100     60            1     1      0   

     alco  active  cardio  
  0     0       1       0  
  1     0       1       1  
  2     0       0       1  
  3     0       1       1  
  4     0       0       0  )
```

In [45]:
```
# Reordering the columns based on their actual order in the dataset for clarity
correct_column_order = ['id', 'age', 'gender', 'height', 'weight', 'ap_hi', 'ap_lo', 'cholesterol', 'gluc', 'sm
df2 = df1[correct_column_order]
df2
```

Out[45]:

| | id | age | gender | height | weight | ap_hi | ap_lo | cholesterol | gluc | smoke | alco | active | cardio |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 18393 | 2 | 168 | 62.0 | 110 | 80 | 1 | 1 | 0 | 0 | 1 | 0 |
| **1** | 1 | 20228 | 1 | 156 | 85.0 | 140 | 90 | 3 | 1 | 0 | 0 | 1 | 1 |
| **2** | 2 | 18857 | 1 | 165 | 64.0 | 130 | 70 | 3 | 1 | 0 | 0 | 0 | 1 |
| **3** | 3 | 17623 | 2 | 169 | 82.0 | 150 | 100 | 1 | 1 | 0 | 0 | 1 | 1 |
| **4** | 4 | 17474 | 1 | 156 | 56.0 | 100 | 60 | 1 | 1 | 0 | 0 | 0 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **69995** | 99993 | 19240 | 2 | 168 | 76.0 | 120 | 80 | 1 | 1 | 1 | 0 | 1 | 0 |
| **69996** | 99995 | 22601 | 1 | 158 | 126.0 | 140 | 90 | 2 | 2 | 0 | 0 | 1 | 1 |
| **69997** | 99996 | 19066 | 2 | 183 | 105.0 | 180 | 90 | 3 | 1 | 0 | 1 | 0 | 1 |
| **69998** | 99998 | 22431 | 1 | 163 | 72.0 | 135 | 80 | 1 | 2 | 0 | 0 | 0 | 1 |
| **69999** | 99999 | 20540 | 1 | 170 | 72.0 | 120 | 80 | 2 | 1 | 0 | 0 | 1 | 0 |

70000 rows × 13 columns

In [47]:
```
df2.head()
```

Out[47]:

| | id | age | gender | height | weight | ap_hi | ap_lo | cholesterol | gluc | smoke | alco | active | cardio |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 18393 | 2 | 168 | 62.0 | 110 | 80 | 1 | 1 | 0 | 0 | 1 | 0 |
| **1** | 1 | 20228 | 1 | 156 | 85.0 | 140 | 90 | 3 | 1 | 0 | 0 | 1 | 1 |
| **2** | 2 | 18857 | 1 | 165 | 64.0 | 130 | 70 | 3 | 1 | 0 | 0 | 0 | 1 |
| **3** | 3 | 17623 | 2 | 169 | 82.0 | 150 | 100 | 1 | 1 | 0 | 0 | 1 | 1 |
| **4** | 4 | 17474 | 1 | 156 | 56.0 | 100 | 60 | 1 | 1 | 0 | 0 | 0 | 0 |

In [49]:
```
df2.tail()
```

Out[49]:

| | id | age | gender | height | weight | ap_hi | ap_lo | cholesterol | gluc | smoke | alco | active | cardio |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **69995** | 99993 | 19240 | 2 | 168 | 76.0 | 120 | 80 | 1 | 1 | 1 | 0 | 1 | 0 |
| **69996** | 99995 | 22601 | 1 | 158 | 126.0 | 140 | 90 | 2 | 2 | 0 | 0 | 1 | 1 |
| **69997** | 99996 | 19066 | 2 | 183 | 105.0 | 180 | 90 | 3 | 1 | 0 | 1 | 0 | 1 |
| **69998** | 99998 | 22431 | 1 | 163 | 72.0 | 135 | 80 | 1 | 2 | 0 | 0 | 0 | 1 |
| **69999** | 99999 | 20540 | 1 | 170 | 72.0 | 120 | 80 | 2 | 1 | 0 | 0 | 1 | 0 |

In [53]:
```
# Converting age from days to years
df2['age'] = (df2['age'] / 365).round(1)
df2
```

Out[53]:

| | id | age | gender | height | weight | ap_hi | ap_lo | cholesterol | gluc | smoke | alco | active | cardio |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 0.1 | 2 | 168 | 62.0 | 110 | 80 | 1 | 1 | 0 | 0 | 1 | 0 |
| **1** | 1 | 0.2 | 1 | 156 | 85.0 | 140 | 90 | 3 | 1 | 0 | 0 | 1 | 1 |
| **2** | 2 | 0.1 | 1 | 165 | 64.0 | 130 | 70 | 3 | 1 | 0 | 0 | 0 | 1 |
| **3** | 3 | 0.1 | 2 | 169 | 82.0 | 150 | 100 | 1 | 1 | 0 | 0 | 1 | 1 |
| **4** | 4 | 0.1 | 1 | 156 | 56.0 | 100 | 60 | 1 | 1 | 0 | 0 | 0 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **69995** | 99993 | 0.1 | 2 | 168 | 76.0 | 120 | 80 | 1 | 1 | 1 | 0 | 1 | 0 |
| **69996** | 99995 | 0.2 | 1 | 158 | 126.0 | 140 | 90 | 2 | 2 | 0 | 0 | 1 | 1 |
| **69997** | 99996 | 0.1 | 2 | 183 | 105.0 | 180 | 90 | 3 | 1 | 0 | 1 | 0 | 1 |
| **69998** | 99998 | 0.2 | 1 | 163 | 72.0 | 135 | 80 | 1 | 2 | 0 | 0 | 0 | 1 |
| **69999** | 99999 | 0.2 | 1 | 170 | 72.0 | 120 | 80 | 2 | 1 | 0 | 0 | 1 | 0 |

70000 rows × 13 columns

In [59]:
```
# Checking for any missing values
missing_values = df2.isnull().sum()
missing_values
```

Out[59]:
```
id             0
age            0
gender         0
height         0
weight         0
ap_hi          0
ap_lo          0
cholesterol    0
gluc           0
smoke          0
alco           0
active         0
cardio         0
dtype: int64
```