Students info

- Student number: s3804803
- Student name: Nguyen Bao Ngan

1. Preprocessing

In [1]:

```
import sklearn as sk
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

In [3]:

```
# Loading data and check if it's loaded properly
df = pd.read_csv("Paitients_Files_Test.csv", delimiter=",")
df.head(15)
```

Out[3]:

	ID	PRG	PL	PR	SK	TS	M11	BD2	Age	Insurance
0	ICU200609	1	109	38	18	120	23.1	0.407	26	1
1	ICU200610	1	108	88	19	0	27.1	0.400	24	1
2	ICU200611	6	96	0	0	0	23.7	0.190	28	1
3	ICU200612	1	124	74	36	0	27.8	0.100	30	1
4	ICU200613	7	150	78	29	126	35.2	0.692	54	0
5	ICU200614	4	183	0	0	0	28.4	0.212	36	1
6	ICU200615	1	124	60	32	0	35.8	0.514	21	1
7	ICU200616	1	181	78	42	293	40.0	1.258	22	1
8	ICU200617	1	92	62	25	41	19.5	0.482	25	0
9	ICU200618	0	152	82	39	272	41.5	0.270	27	0
10	ICU200619	1	111	62	13	182	24.0	0.138	23	1
11	ICU200620	3	106	54	21	158	30.9	0.292	24	1
12	ICU200621	3	174	58	22	194	32.9	0.593	36	1
13	ICU200622	7	168	88	42	321	38.2	0.787	40	1
14	ICU200623	6	105	80	28	0	32.5	0.878	26	1

1.1 Check data types, null values

- According to the below table, 10 features don't have any missing values, namely 169 out 169 is non-null.
- However, in the first 15 records, we see that there are some value = 0 ?, so we need look it up to see if 0 is valid value for those columns.

In [4]:

```
# Inspect data types of the dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 169 entries, 0 to 168
Data columns (total 10 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   ID         169 non-null    object 
 1   PRG        169 non-null    int64  
 2   PL         169 non-null    int64  
 3   PR         169 non-null    int64  
 4   SK         169 non-null    int64  
 5   TS         169 non-null    int64  
 6   M11        169 non-null    float64
 7   BD2        169 non-null    float64
 8   Age        169 non-null    int64  
 9   Insurance  169 non-null    int64  
dtypes: float64(2), int64(7), object(1)
memory usage: 13.3+ KB
```

In [ ]:

```
# drop 2 columns: ID and Insurance because they are not counted as patient attributes
%pip uninstall pandoc && conda install pandoc
```

```
Found existing installation: pandoc 2.1
Uninstalling pandoc-2.1:
  Would remove:
    /Users/lap11353-local/opt/anaconda3/envs/mlenv/lib/python3.9/site-packages/pandoc-2.1.dist-info/*
    /Users/lap11353-local/opt/anaconda3/envs/mlenv/lib/python3.9/site-packages/pandoc/*
Proceed (Y/n)?
```

In [ ]: