H4212

Yi ZHANG Fatma LARIBI Korantin AIMÉ Arthur UNGRIA

30 Avril, 2023

Lien vers les datasets

[128]:
```
path_mat = "/content/content/student-mat.csv"
path_por = "/content/content/student-por.csv"
```

0.1 Imports

Afin de l'analyse du modèle, nous introduisons le cadre de shap. SHAP peut être utilisé pour expliquer les causes des prédictions individuelles, ainsi que le comportement du modèle dans son ensemble.

[129]:
```
pip install shap
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Requirement already satisfied: shap in /usr/local/lib/python3.10/dist-packages
(0.41.0)
Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages
(from shap) (1.22.4)
Requirement already satisfied: scipy in /usr/local/lib/python3.10/dist-packages
(from shap) (1.10.1)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.10/dist-
packages (from shap) (1.2.2)
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages
(from shap) (1.5.3)
Requirement already satisfied: tqdm>4.25.0 in /usr/local/lib/python3.10/dist-
packages (from shap) (4.65.0)
Requirement already satisfied: packaging>20.9 in /usr/local/lib/python3.10/dist-
packages (from shap) (23.1)
Requirement already satisfied: slicer==0.0.7 in /usr/local/lib/python3.10/dist-
packages (from shap) (0.0.7)
Requirement already satisfied: numba in /usr/local/lib/python3.10/dist-packages
(from shap) (0.56.4)
Requirement already satisfied: cloudpickle in /usr/local/lib/python3.10/dist-
packages (from shap) (2.2.1)
Requirement already satisfied: llvmlite<0.40,>=0.39.0dev0 in
/usr/local/lib/python3.10/dist-packages (from numba->shap) (0.39.1)
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/dist-
packages (from numba->shap) (67.7.2)
```

```
Requirement already satisfied: python-dateutil>=2.8.1 in
/usr/local/lib/python3.10/dist-packages (from pandas->shap) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-
packages (from pandas->shap) (2022.7.1)
Requirement already satisfied: joblib>=1.1.1 in /usr/local/lib/python3.10/dist-
packages (from scikit-learn->shap) (1.2.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/usr/local/lib/python3.10/dist-packages (from scikit-learn->shap) (3.1.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-
packages (from python-dateutil>=2.8.1->pandas->shap) (1.16.0)
```

[130]:
```
import numpy as np
import pandas as pd
import math
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import KFold, cross_val_score, train_test_split,␣
  ↪GridSearchCV
from sklearn.linear_model import LinearRegression, Ridge
from sklearn.preprocessing import StandardScaler, PolynomialFeatures
from sklearn.pipeline import make_pipeline
from sklearn.metrics import mean_squared_error, r2_score
import seaborn as sns
import shap

pd.set_option('display.max_columns', None)
```

0.2 Lecture des données

[131]:
```
data_mat = pd.read_csv(path_mat,sep=';')
data_por = pd.read_csv(path_por,sep=';')
```

0.3 Visualisation et exploration des données

Explorer la forme des ensembles de données

[133]:
```
data_mat.shape
```

[133]: (395, 33)

[134]:
```
data_por.shape
```

[134]: (649, 33)

[135]:
```
data_mat.columns
```

[135]:
```
Index(['school', 'sex', 'age', 'address', 'famsize', 'Pstatus', 'Medu', 'Fedu',
       'Mjob', 'Fjob', 'reason', 'guardian', 'traveltime', 'studytime',
       'failures', 'schoolsup', 'famsup', 'paid', 'activities', 'nursery',
       'higher', 'internet', 'romantic', 'famrel', 'freetime', 'goout', 'Dalc',
       'Walc', 'health', 'absences', 'G1', 'G2', 'G3'],
      dtype='object')
```

[136]:
```
data_por.columns
```

[136]:
```
Index(['school', 'sex', 'age', 'address', 'famsize', 'Pstatus', 'Medu', 'Fedu',
       'Mjob', 'Fjob', 'reason', 'guardian', 'traveltime', 'studytime',
       'failures', 'schoolsup', 'famsup', 'paid', 'activities', 'nursery',
       'higher', 'internet', 'romantic', 'famrel', 'freetime', 'goout', 'Dalc',
       'Walc', 'health', 'absences', 'G1', 'G2', 'G3'],
      dtype='object')
```

On remarque que les colonnes des 2 datasets sont les mêmes. En effet, on possède 2 datasets de la même structure : une pour les notes de Mathématiques et l'autre pour les notes de Portugais. On explore encore les 2 datasets en utilisant 'head' pour les 5 premiers lignes, et 'tail' pour les 5 derniers lignes.

[137]:
```
data_mat.head()
```

[137]:
```
  school sex  age address famsize Pstatus  Medu  Fedu     Mjob      Fjob  \
0     GP   F   18       U     GT3       A     4     4  at_home   teacher
1     GP   F   17       U     GT3       T     1     1  at_home     other
2     GP   F   15       U     LE3       T     1     1  at_home     other
3     GP   F   15       U     GT3       T     4     2   health  services
4     GP   F   16       U     GT3       T     3     3    other     other

   reason guardian  traveltime  studytime  failures schoolsup famsup paid  \
0  course   mother           2          2         0       yes     no   no
1  course   father           1          2         0        no    yes   no
2   other   mother           1          2         3       yes     no  yes
3    home   mother           1          3         0        no    yes  yes
4    home   father           1          2         0        no    yes  yes

  activities nursery higher internet romantic  famrel  freetime  goout  Dalc  \
0         no     yes    yes       no       no       4         3      4     1
1         no      no    yes      yes       no       5         3      3     1
2         no     yes    yes      yes       no       4         3      2     2
3        yes     yes    yes      yes      yes       3         2      2     1
4         no     yes    yes       no       no       4         3      2     1

   Walc  health  absences  G1  G2  G3
0     1       3         6   5   6   6
1     1       3         4   5   5   6
```

```
2     3        3         10    7    8   10
3     1        5          2   15   14   15
4     2        5          4    6   10   10
```

[138]:
```
data_por.head()
```

[138]:
```
  school sex  age address famsize Pstatus  Medu  Fedu     Mjob      Fjob  \
0     GP   F   18       U     GT3       A     4     4  at_home   teacher
1     GP   F   17       U     GT3       T     1     1  at_home     other
2     GP   F   15       U     LE3       T     1     1  at_home     other
3     GP   F   15       U     GT3       T     4     2   health  services
4     GP   F   16       U     GT3       T     3     3    other     other

   reason guardian  traveltime  studytime  failures schoolsup famsup paid  \
0  course   mother           2          2         0       yes     no   no
1  course   father           1          2         0        no    yes   no
2   other   mother           1          2         0       yes     no   no
3    home   mother           1          3         0        no    yes   no
4    home   father           1          2         0        no    yes   no

  activities nursery higher internet romantic  famrel  freetime  goout  Dalc  \
0         no     yes    yes       no       no       4         3      4     1
1         no      no    yes      yes       no       5         3      3     1
2         no     yes    yes      yes       no       4         3      2     2
3        yes     yes    yes      yes      yes       3         2      2     1
4         no     yes    yes       no       no       4         3      2     1

   Walc  health  absences  G1  G2  G3
0     1       3         4   0  11  11
1     1       3         2   9  11  11
2     3       3         6  12  13  12
3     1       5         0  14  14  14
4     2       5         0  11  13  13
```

[139]:
```
data_mat.tail()
```

[139]:
```
    school sex  age address famsize Pstatus  Medu  Fedu      Mjob      Fjob  \
390     MS   M   20       U     LE3       A     2     2  services  services
391     MS   M   17       U     LE3       T     3     1  services  services
392     MS   M   21       R     GT3       T     1     1     other     other
393     MS   M   18       R     LE3       T     3     2  services     other
394     MS   M   19       U     LE3       T     1     1     other   at_home

     reason guardian  traveltime  studytime  failures schoolsup famsup paid  \
390  course    other           1          2         2        no    yes  yes
391  course   mother           2          1         0        no     no   no
392  course    other           1          1         3        no     no   no
```

```
393  course    mother           3          1         0        no     no   no
394  course    father           1          1         0        no     no   no

    activities nursery higher internet romantic  famrel  freetime  goout  \
390         no     yes    yes       no       no       5         5      4
391         no      no    yes      yes       no       2         4      5
392         no      no    yes       no       no       5         5      3
393         no      no    yes      yes       no       4         4      1
394         no     yes    yes      yes       no       3         2      3

     Dalc  Walc  health  absences  G1  G2  G3
390     4     5       4        11   9   9   9
391     3     4       2         3  14  16  16
392     3     3       3         3  10   8   7
393     3     4       5         0  11  12  10
394     3     3       5         5   8   9   9
```

[140]:
```
data_por.tail()
```

[140]:
```
    school sex  age address famsize Pstatus  Medu  Fedu      Mjob      Fjob  \
644     MS   F   19       R     GT3       T     2     3  services     other
645     MS   F   18       U     LE3       T     3     1   teacher  services
646     MS   F   18       U     GT3       T     1     1     other     other
647     MS   M   17       U     LE3       T     3     1  services  services
648     MS   M   18       R     LE3       T     3     2  services     other

     reason guardian  traveltime  studytime  failures schoolsup famsup paid  \
644  course   mother           1          3         1        no     no   no
645  course   mother           1          2         0        no    yes   no
646  course   mother           2          2         0        no     no   no
647  course   mother           2          1         0        no     no   no
648  course   mother           3          1         0        no     no   no

    activities nursery higher internet romantic  famrel  freetime  goout  \
644        yes      no    yes      yes       no       5         4      2
645         no     yes    yes      yes       no       4         3      4
646        yes     yes    yes       no       no       1         1      1
647         no      no    yes      yes       no       2         4      5
648         no      no    yes      yes       no       4         4      1

     Dalc  Walc  health  absences  G1  G2  G3
644     1     2       5         4  10  11  10
645     1     1       1         4  15  15  16
646     1     1       5         6  11  12   9
647     3     4       2         6  10  10  10
648     3     4       5         4  10  11  11
```

Vérifier si les colonnes des deux fichiers correspondent

[141]:
```
sum(list(data_mat.columns != data_por.columns))
```

[141]: 0

On prend un échantillon aléatoire de chaque dataset:

[142]:
```
data_mat.sample()
```

[142]:
```
  school sex  age address famsize Pstatus  Medu  Fedu      Mjob    Fjob  \
38     GP   F   15       R     GT3       T     3     4  services  health

    reason guardian  traveltime  studytime  failures schoolsup famsup paid  \
38  course   mother           1          3         0       yes    yes  yes

  activities nursery higher internet romantic  famrel  freetime  goout  Dalc  \
38        yes     yes    yes      yes       no       4         3      2     1

    Walc  health  absences  G1  G2  G3
38     1       5         2  12  12  11
```

[143]:
```
data_por.sample()
```

[143]:
```
    school sex  age address famsize Pstatus  Medu  Fedu   Mjob   Fjob  reason  \
355     GP   F   17       U     GT3       T     2     3  other  other  course

    guardian  traveltime  studytime  failures schoolsup famsup paid  \
355   father           2          2         0        no     no   no

    activities nursery higher internet romantic  famrel  freetime  goout  \
355        yes     yes    yes      yes      yes       4         2      1

     Dalc  Walc  health  absences  G1  G2  G3
355     1     1       3         2  11  12  14
```

Avec le .info() on peut voir un sommaire de type data de chaque colonnes, le nombre de valeurs non nulles, et l'utilisation de la mémoire. Il y a des objets dans le dataset, ce qui signifie que nous avons des catégories que nous devons transformer en int.

[144]:
```
data_mat.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 395 entries, 0 to 394
Data columns (total 33 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   school      395 non-null    object
 1   sex         395 non-null    object
```

```
 2   age         395 non-null    int64
 3   address     395 non-null    object
 4   famsize     395 non-null    object
 5   Pstatus     395 non-null    object
 6   Medu        395 non-null    int64
 7   Fedu        395 non-null    int64
 8   Mjob        395 non-null    object
 9   Fjob        395 non-null    object
 10  reason      395 non-null    object
 11  guardian    395 non-null    object
 12  traveltime  395 non-null    int64
 13  studytime   395 non-null    int64
 14  failures    395 non-null    int64
 15  schoolsup   395 non-null    object
 16  famsup      395 non-null    object
 17  paid        395 non-null    object
 18  activities  395 non-null    object
 19  nursery     395 non-null    object
 20  higher      395 non-null    object
 21  internet    395 non-null    object
 22  romantic    395 non-null    object
 23  famrel      395 non-null    int64
 24  freetime    395 non-null    int64
 25  goout       395 non-null    int64
 26  Dalc        395 non-null    int64
 27  Walc        395 non-null    int64
 28  health      395 non-null    int64
 29  absences    395 non-null    int64
 30  G1          395 non-null    int64
 31  G2          395 non-null    int64
 32  G3          395 non-null    int64
dtypes: int64(16), object(17)
memory usage: 102.0+ KB
```

[145]:
```
data_por.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 649 entries, 0 to 648
Data columns (total 33 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   school      649 non-null    object
 1   sex         649 non-null    object
 2   age         649 non-null    int64
 3   address     649 non-null    object
 4   famsize     649 non-null    object
 5   Pstatus     649 non-null    object
 6   Medu        649 non-null    int64
```

```
 7   Fedu        649 non-null    int64
 8   Mjob        649 non-null    object
 9   Fjob        649 non-null    object
 10  reason      649 non-null    object
 11  guardian    649 non-null    object
 12  traveltime  649 non-null    int64
 13  studytime   649 non-null    int64
 14  failures    649 non-null    int64
 15  schoolsup   649 non-null    object
 16  famsup      649 non-null    object
 17  paid        649 non-null    object
 18  activities  649 non-null    object
 19  nursery     649 non-null    object
 20  higher      649 non-null    object
 21  internet    649 non-null    object
 22  romantic    649 non-null    object
 23  famrel      649 non-null    int64
 24  freetime    649 non-null    int64
 25  goout       649 non-null    int64
 26  Dalc        649 non-null    int64
 27  Walc        649 non-null    int64
 28  health      649 non-null    int64
 29  absences    649 non-null    int64
 30  G1          649 non-null    int64
 31  G2          649 non-null    int64
 32  G3          649 non-null    int64
dtypes: int64(16), object(17)
memory usage: 167.4+ KB
```

Puisqu'on a vérifié que les 2 datasets ont la même structure, on peut concaténer les 2 datasets pour faciliter leurs manipulations.

[146]:
```python
data = pd.concat([data_mat, data_por], ignore_index=True)
data
```

[146]:
```
     school sex  age address famsize Pstatus  Medu  Fedu      Mjob      Fjob  \
0        GP   F   18       U     GT3       A     4     4   at_home   teacher
1        GP   F   17       U     GT3       T     1     1   at_home     other
2        GP   F   15       U     LE3       T     1     1   at_home     other
3        GP   F   15       U     GT3       T     4     2    health  services
4        GP   F   16       U     GT3       T     3     3     other     other
…       …  ..   …       …      …      …     …    …       …         …
1039     MS   F   19       R     GT3       T     2     3  services     other
1040     MS   F   18       U     LE3       T     3     1   teacher  services
1041     MS   F   18       U     GT3       T     1     1     other     other
1042     MS   M   17       U     LE3       T     3     1  services  services
1043     MS   M   18       R     LE3       T     3     2  services     other
```

```
     reason guardian  traveltime  studytime  failures schoolsup famsup paid  \
0    course   mother           2          2         0       yes     no   no
1    course   father           1          2         0        no    yes   no
2     other   mother           1          2         3       yes     no  yes
3      home   mother           1          3         0        no    yes  yes
4      home   father           1          2         0        no    yes  yes
…       …      …         …          …         …        …      …    …
1039 course   mother           1          3         1        no     no   no
1040 course   mother           1          2         0        no    yes   no
1041 course   mother           2          2         0        no     no   no
1042 course   mother           2          1         0        no     no   no
1043 course   mother           3          1         0        no     no   no

     activities nursery higher internet romantic  famrel  freetime  goout  \
0            no     yes    yes       no       no       4         3      4
1            no      no    yes      yes       no       5         3      3
2            no     yes    yes      yes       no       4         3      2
3           yes     yes    yes      yes      yes       3         2      2
4            no     yes    yes       no       no       4         3      2
…          …      …      …       …        …       …        …      …
1039        yes      no    yes      yes       no       5         4      2
1040         no     yes    yes      yes       no       4         3      4
1041        yes     yes    yes       no       no       1         1      1
1042         no      no    yes      yes       no       2         4      5
1043         no      no    yes      yes       no       4         4      1

      Dalc  Walc  health  absences  G1  G2  G3
0        1     1       3         6   5   6   6
1        1     1       3         4   5   5   6
2        2     3       3        10   7   8  10
3        1     1       5         2  15  14  15
4        1     2       5         4   6  10  10
…    …    …     …       …         ..  ..  ..
1039     1     2       5         4  10  11  10
1040     1     1       1         4  15  15  16
1041     1     1       5         6  11  12   9
1042     3     4       2         6  10  10  10
1043     3     4       5         4  10  11  11

[1044 rows x 33 columns]
```

On affiche les domaines des valeurs du dataset:

[147]:
```python
for col in data.columns:
    print(col,"  ", data[col].unique())
```

```
school    ['GP' 'MS']
sex    ['F' 'M']
```

```
age    [18 17 15 16 19 22 20 21]
address    ['U' 'R']
famsize    ['GT3' 'LE3']
Pstatus    ['A' 'T']
Medu    [4 1 3 2 0]
Fedu    [4 1 2 3 0]
Mjob    ['at_home' 'health' 'other' 'services' 'teacher']
Fjob    ['teacher' 'other' 'services' 'health' 'at_home']
reason    ['course' 'other' 'home' 'reputation']
guardian    ['mother' 'father' 'other']
traveltime    [2 1 3 4]
studytime    [2 3 1 4]
failures    [0 3 2 1]
schoolsup    ['yes' 'no']
famsup    ['no' 'yes']
paid    ['no' 'yes']
activities    ['no' 'yes']
nursery    ['yes' 'no']
higher    ['yes' 'no']
internet    ['no' 'yes']
romantic    ['no' 'yes']
famrel    [4 5 3 1 2]
freetime    [3 2 4 1 5]
goout    [4 3 2 1 5]
Dalc    [1 2 5 3 4]
Walc    [1 3 2 4 5]
health    [3 5 1 2 4]
absences    [ 6  4 10  2  0 16 14  7  8 25 12 54 18 26 20 56 24 28  5 13 15 22
3 21
  1 75 30 19  9 11 38 40 23 17 32]
G1    [ 5  7 15  6 12 16 14 10 13  8 11  9 17 19 18  4  3  0]
G2    [ 6  5  8 14 10 15 12 18 16 13  9 11  7 19 17  4  0]
G3    [ 6 10 15 11 19  9 12 14 16  5  8 17 18 13 20  7  0  4  1]
```

On remarque que par rapport au fichier de renseignement fourni avec les 2 databases, la colonne de failures prend les valeurs de 0 à 3 au lieu de 1 à 4. On considère que l'erreur est fait au niveau du fichier de renseignement, et pas au niveau des datasets.

[148]:
```python
data.duplicated().sum()
```

[148]:
```
0
```

On remarque qu'il n'y a pas de duplication au niveau de nos données.

[150]:
```python
data.describe()
```

[150]:
```
               age         Medu         Fedu   traveltime    studytime  \
count  1044.000000  1044.000000  1044.000000  1044.000000  1044.000000
```

```
mean     16.726054     2.603448     2.387931     1.522989     1.970307
std       1.239975     1.124907     1.099938     0.731727     0.834353
min      15.000000     0.000000     0.000000     1.000000     1.000000
25%      16.000000     2.000000     1.000000     1.000000     1.000000
50%      17.000000     3.000000     2.000000     1.000000     2.000000
75%      18.000000     4.000000     3.000000     2.000000     2.000000
max      22.000000     4.000000     4.000000     4.000000     4.000000

          failures       famrel     freetime        goout         Dalc  \
count  1044.000000  1044.000000  1044.000000  1044.000000  1044.000000
mean      0.264368     3.935824     3.201149     3.156130     1.494253
std       0.656142     0.933401     1.031507     1.152575     0.911714
min       0.000000     1.000000     1.000000     1.000000     1.000000
25%       0.000000     4.000000     3.000000     2.000000     1.000000
50%       0.000000     4.000000     3.000000     3.000000     1.000000
75%       0.000000     5.000000     4.000000     4.000000     2.000000
max       3.000000     5.000000     5.000000     5.000000     5.000000

              Walc       health     absences           G1           G2  \
count  1044.000000  1044.000000  1044.000000  1044.000000  1044.000000
mean      2.284483     3.543103     4.434866    11.213602    11.246169
std       1.285105     1.424703     6.210017     2.983394     3.285071
min       1.000000     1.000000     0.000000     0.000000     0.000000
25%       1.000000     3.000000     0.000000     9.000000     9.000000
50%       2.000000     4.000000     2.000000    11.000000    11.000000
75%       3.000000     5.000000     6.000000    13.000000    13.000000
max       5.000000     5.000000    75.000000    19.000000    19.000000

                G3
count  1044.000000
mean     11.341954
std       3.864796
min       0.000000
25%      10.000000
50%      11.000000
75%      14.000000
max      20.000000
```

On remarque que le std de la colonne absences est élevé par rapport aux autres colonnes(= environ 6).

[151]:
```python
data.describe(include="object")
```

[151]:
```
       school   sex address famsize Pstatus   Mjob   Fjob  reason guardian  \
count    1044  1044    1044    1044    1044   1044   1044    1044     1044
unique      2     2       2       2       2      5      5       4        3
top        GP     F       U     GT3       T  other  other  course   mother
```

```
freq       772    591      759      738      923     399     584      430       728

       schoolsup famsup  paid activities nursery higher internet romantic
count        1044   1044  1044       1044    1044   1044     1044     1044
unique          2      2     2          2       2      2        2        2
top            no    yes    no         no     yes    yes      yes       no
freq          925    640   824        528     835    955      827      673
```

dtypes nous permet d'explorer les types de données qu'on a. Object correspond à une variable catégorique et int est une variable numérique.

[152]:
```python
data.dtypes
```

[152]:
```
school         object
sex            object
age             int64
address        object
famsize        object
Pstatus        object
Medu            int64
Fedu            int64
Mjob           object
Fjob           object
reason         object
guardian       object
traveltime      int64
studytime       int64
failures        int64
schoolsup      object
famsup         object
paid           object
activities     object
nursery        object
higher         object
internet       object
romantic       object
famrel          int64
freetime        int64
goout           int64
Dalc            int64
Walc            int64
health          int64
absences        int64
G1              int64
G2              int64
G3              int64
dtype: object
```

```
[153]: data.isna().sum()
```

```
[153]: school        0
       sex           0
       age           0
       address       0
       famsize       0
       Pstatus       0
       Medu          0
       Fedu          0
       Mjob          0
       Fjob          0
       reason        0
       guardian      0
       traveltime    0
       studytime     0
       failures      0
       schoolsup     0
       famsup        0
       paid          0
       activities    0
       nursery       0
       higher        0
       internet      0
       romantic      0
       famrel        0
       freetime      0
       goout         0
       Dalc          0
       Walc          0
       health        0
       absences      0
       G1            0
       G2            0
       G3            0
       dtype: int64
```

Chercher les valeurs nulles

```
[154]: data.isnull().sum()
```

```
[154]: school        0
       sex           0
       age           0
       address       0
       famsize       0
       Pstatus       0
```

```
Medu          0
Fedu          0
Mjob          0
Fjob          0
reason        0
guardian      0
traveltime    0
studytime     0
failures      0
schoolsup     0
famsup        0
paid          0
activities    0
nursery       0
higher        0
internet      0
romantic      0
famrel        0
freetime      0
goout         0
Dalc          0
Walc          0
health        0
absences      0
G1            0
G2            0
G3            0
dtype: int64
```

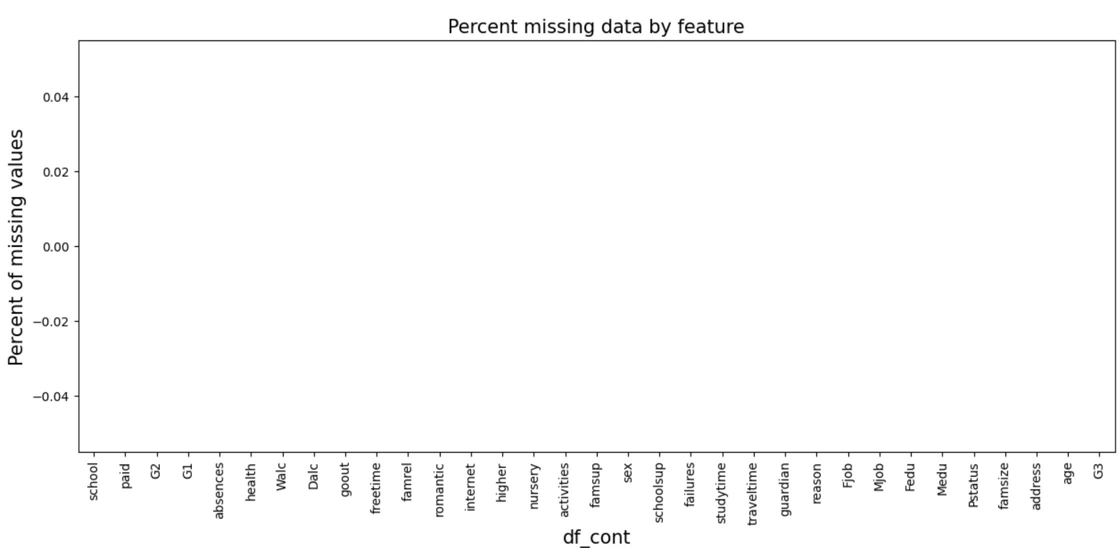
[155]:

```
total = data.isnull().sum().sort_values(ascending=False)
percent = (data.isnull().sum()/data.isnull().count()).
  ↪sort_values(ascending=False)
missing_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
f, ax = plt.subplots(figsize=(15, 6))

plt.xticks(rotation=90)

sns.barplot(x=missing_data.index, y=missing_data['Percent'])
plt.xlabel('df_cont', fontsize=15)
plt.ylabel('Percent of missing values', fontsize=15)
plt.title('Percent missing data by feature', fontsize=15)
missing_data
```

[155]:

```
            Total  Percent
school          0      0.0
paid            0      0.0
G2              0      0.0
```

```
G1              0       0.0
absences        0       0.0
health          0       0.0
Walc            0       0.0
Dalc            0       0.0
goout           0       0.0
freetime        0       0.0
famrel          0       0.0
romantic        0       0.0
internet        0       0.0
higher          0       0.0
nursery         0       0.0
activities      0       0.0
famsup          0       0.0
sex             0       0.0
schoolsup       0       0.0
failures        0       0.0
studytime       0       0.0
traveltime      0       0.0
guardian        0       0.0
reason          0       0.0
Fjob            0       0.0
Mjob            0       0.0
Fedu            0       0.0
Medu            0       0.0
Pstatus         0       0.0
famsize         0       0.0
address         0       0.0
age             0       0.0
G3              0       0.0
```

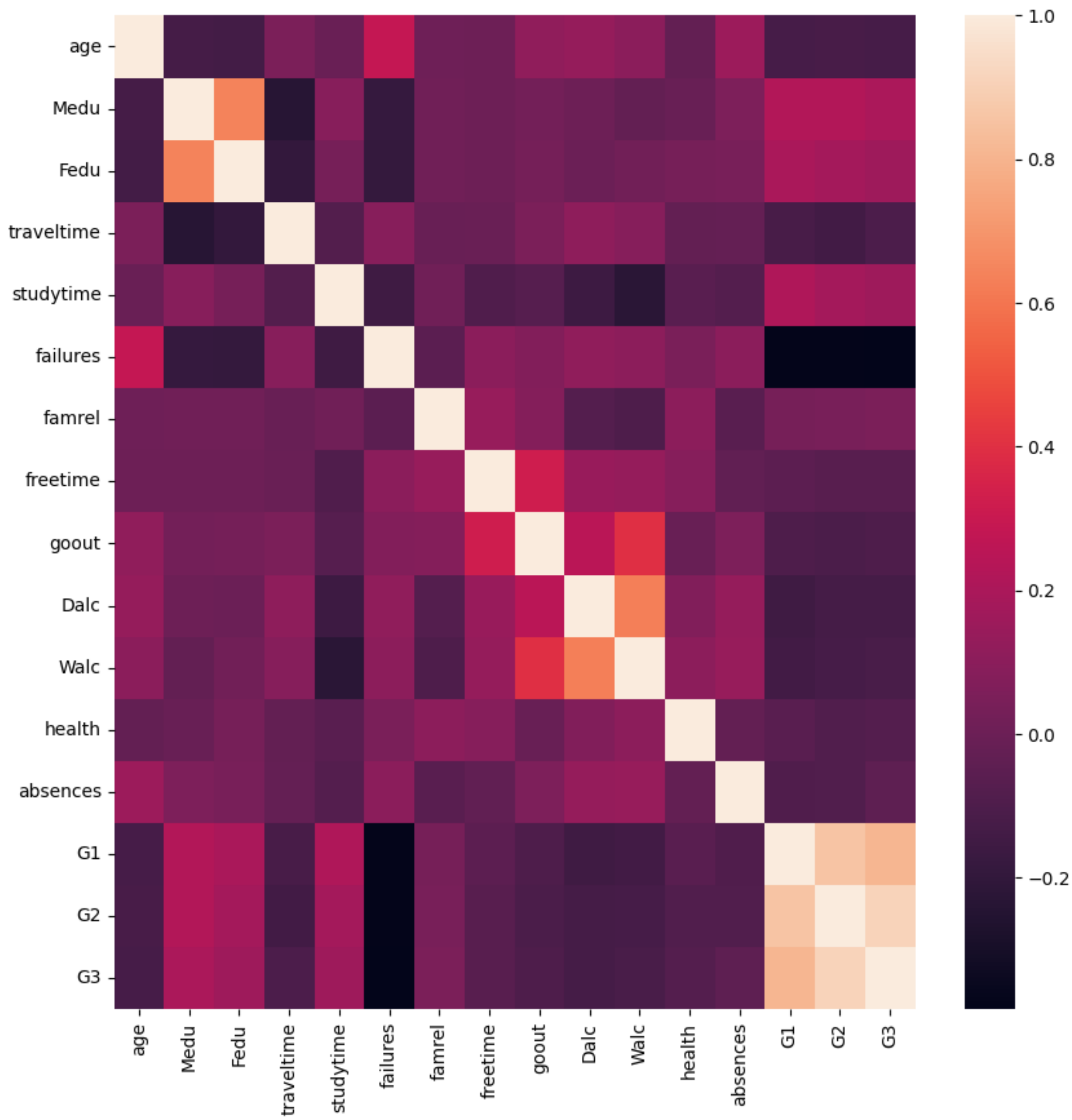


Aucune valeur manquante ou nulle dans l'ensemble de données, mais les notes = 0 pourraient être des absences, Nous traiterons les absences potentielles (note = 0 plus tard dans le processus, pour l'instant nous allons juste explorer les données et détecter les anomalies).

[156]:
```python
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure( figsize = (10,10))
sns.heatmap(data.corr())
```

```
The default value of numeric_only in DataFrame.corr is deprecated. In a future
version, it will default to False. Select only valid columns or specify the
value of numeric_only to silence this warning.
```

[156]:
```
<Axes: >
```



On note les corrélations entre: Medu et Fedu (éducation de la mère et du père) Dalc et Walc qui correspondent à la consommation d'alcool et les niveaux scolaires G1, G2, et G3

Il existe aussi une légère corrélation entre goout et Walc, et une très faible corrélation entre les failures et G1, G2, G3.

[157]:
```
data = pd.DataFrame(data)

print(data.corr())
```

```
                 age      Medu      Fedu  traveltime  studytime  failures  \
age         1.000000 -0.130196 -0.138521    0.049216  -0.007870  0.282364
Medu       -0.130196  1.000000  0.642063   -0.238181   0.090616 -0.187769
```

```
Fedu       -0.138521  0.642063  1.000000   -0.196328   0.033458 -0.191390
traveltime  0.049216 -0.238181 -0.196328    1.000000  -0.081328  0.087177
studytime  -0.007870  0.090616  0.033458   -0.081328   1.000000 -0.152024
failures    0.282364 -0.187769 -0.191390    0.087177  -0.152024  1.000000
famrel      0.007162  0.015004  0.013066   -0.012578   0.012324 -0.053676
freetime    0.002645  0.001054  0.002142   -0.007403  -0.094429  0.102679
goout       0.118510  0.025614  0.030075    0.049740  -0.072941  0.074683
Dalc        0.133453  0.001515 -0.000165    0.109423  -0.159665  0.116336
Walc        0.098291 -0.029331  0.019524    0.084292  -0.229073  0.107432
health     -0.029129 -0.013254  0.034288   -0.029002  -0.063044  0.048311
absences    0.153196  0.059708  0.040829   -0.022669  -0.075594  0.099998
G1         -0.124121  0.226101  0.195898   -0.121053   0.211314 -0.374175
G2         -0.119475  0.224662  0.182634   -0.140163   0.183167 -0.377172
G3         -0.125282  0.201472  0.159796   -0.102627   0.161629 -0.383145

              famrel  freetime     goout      Dalc      Walc    health  \
age         0.007162  0.002645  0.118510  0.133453  0.098291 -0.029129
Medu        0.015004  0.001054  0.025614  0.001515 -0.029331 -0.013254
Fedu        0.013066  0.002142  0.030075 -0.000165  0.019524  0.034288
traveltime -0.012578 -0.007403  0.049740  0.109423  0.084292 -0.029002
studytime   0.012324 -0.094429 -0.072941 -0.159665 -0.229073 -0.063044
failures   -0.053676  0.102679  0.074683  0.116336  0.107432  0.048311
famrel      1.000000  0.136901  0.080619 -0.076483 -0.100663  0.104101
freetime    0.136901  1.000000  0.323556  0.144979  0.130377  0.081517
goout       0.080619  0.323556  1.000000  0.253135  0.399794 -0.013736
Dalc       -0.076483  0.144979  0.253135  1.000000  0.627814  0.065515
Walc       -0.100663  0.130377  0.399794  0.627814  1.000000  0.106669
health      0.104101  0.081517 -0.013736  0.065515  0.106669  1.000000
absences   -0.062171 -0.032079  0.056142  0.132867  0.139703 -0.027479
G1          0.036947 -0.051985 -0.101163 -0.150943 -0.142401 -0.060478
G2          0.042054 -0.068952 -0.108411 -0.131576 -0.128114 -0.088001
G3          0.054461 -0.064890 -0.097877 -0.129642 -0.115740 -0.080079

            absences        G1        G2        G3
age         0.153196 -0.124121 -0.119475 -0.125282
Medu        0.059708  0.226101  0.224662  0.201472
Fedu        0.040829  0.195898  0.182634  0.159796
traveltime -0.022669 -0.121053 -0.140163 -0.102627
studytime  -0.075594  0.211314  0.183167  0.161629
failures    0.099998 -0.374175 -0.377172 -0.383145
famrel     -0.062171  0.036947  0.042054  0.054461
freetime   -0.032079 -0.051985 -0.068952 -0.064890
goout       0.056142 -0.101163 -0.108411 -0.097877
Dalc        0.132867 -0.150943 -0.131576 -0.129642
Walc        0.139703 -0.142401 -0.128114 -0.115740
health     -0.027479 -0.060478 -0.088001 -0.080079
absences    1.000000 -0.092425 -0.089332 -0.045671
G1         -0.092425  1.000000  0.858739  0.809142
```

```
G2         -0.089332  0.858739  1.000000  0.910743
G3         -0.045671  0.809142  0.910743  1.000000
```

```
The default value of numeric_only in DataFrame.corr is deprecated. In a future
version, it will default to False. Select only valid columns or specify the
value of numeric_only to silence this warning.
```

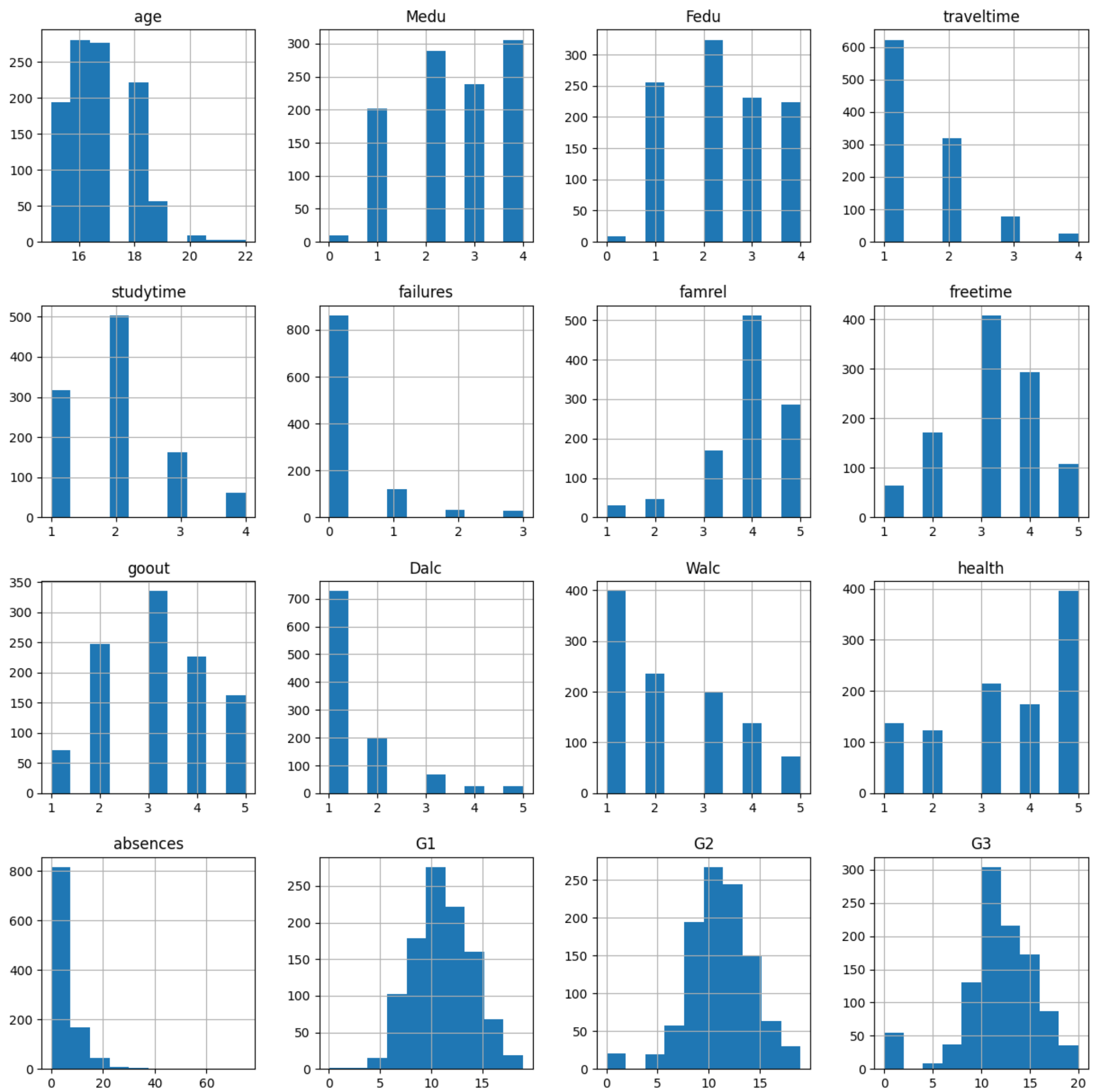
On peut voir qu'il existe une très forte corrélation entre les notes des semestre 1, 2 et la note finale. Or nous ne voulons pas entrainer un modèle qui se base principalement sur les notes de l'année pour prédire la note finale mais plutôt avoir un modèle qui s'appuie sur l'ensemble des autres données. C'est pourquoi dans la suite, nous n'inclurons pas les notes des semestres 1 et 2 comme features dans nos modèles.

1 Statistiques

1.1 Histogramme

[158]:
```
data.hist(figsize = (15,15))
```

[158]:
```
array([[<Axes: title={'center': 'age'}>,
        <Axes: title={'center': 'Medu'}>,
        <Axes: title={'center': 'Fedu'}>,
        <Axes: title={'center': 'traveltime'}>],
       [<Axes: title={'center': 'studytime'}>,
        <Axes: title={'center': 'failures'}>,
        <Axes: title={'center': 'famrel'}>,
        <Axes: title={'center': 'freetime'}>],
       [<Axes: title={'center': 'goout'}>,
        <Axes: title={'center': 'Dalc'}>,
        <Axes: title={'center': 'Walc'}>,
        <Axes: title={'center': 'health'}>],
       [<Axes: title={'center': 'absences'}>,
        <Axes: title={'center': 'G1'}>, <Axes: title={'center': 'G2'}>,
        <Axes: title={'center': 'G3'}>]], dtype=object)
```



Afin de pouvoir mieux analyser ces données, nous avons d'abord utilisé l'histogramme pour montrer la distribution de chaque type de manière générale. Nous remarquons par exemple que les absences possèdent des valeurs rares > 20 et que la plupart des valeurs sont < 20, que les notes sont centrées sur 11 environ… Toutes ces colonnes vont ensuite être vues plus en détail grâce aux boxplots.

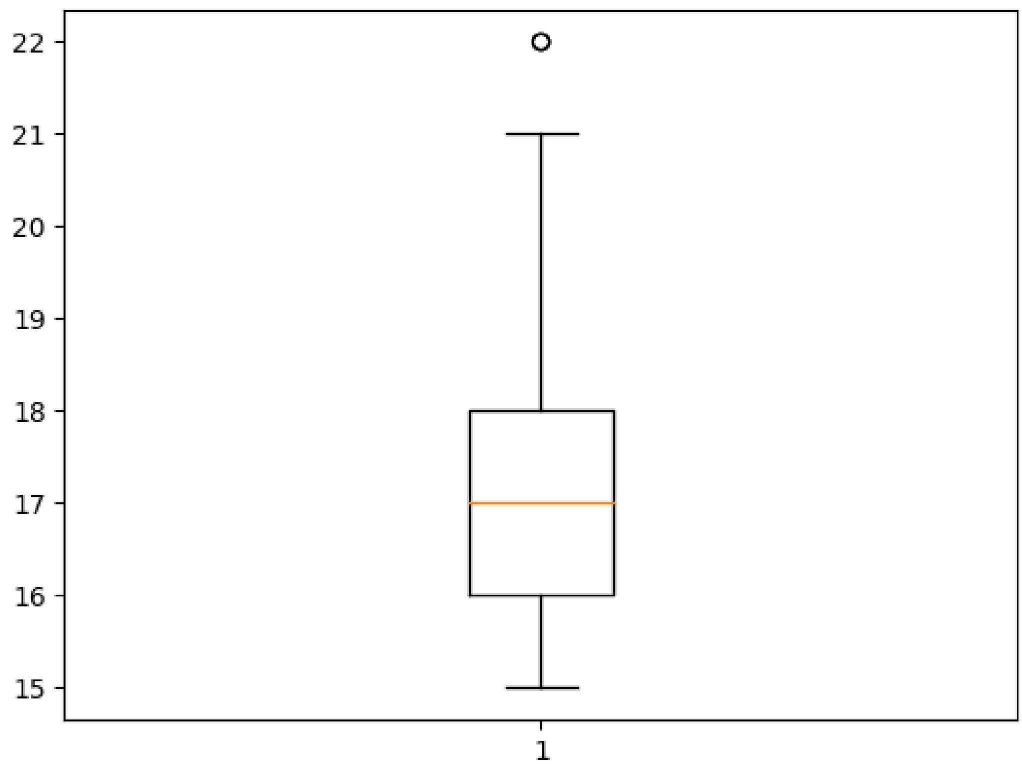
1.2 Box plot

1.2.1 Age des étudiants

[159]:
```
plt.boxplot(data['age'])
```

[159]:
```
{'whiskers': [<matplotlib.lines.Line2D at 0x7ff1b30bebf0>,
  <matplotlib.lines.Line2D at 0x7ff1b30beec0>],
```

```
 'caps': [<matplotlib.lines.Line2D at 0x7ff1b30bf160>,
  <matplotlib.lines.Line2D at 0x7ff1b30bf400>],
 'boxes': [<matplotlib.lines.Line2D at 0x7ff1b30bea70>],
 'medians': [<matplotlib.lines.Line2D at 0x7ff1b30bf6a0>],
 'fliers': [<matplotlib.lines.Line2D at 0x7ff1b30bf940>],
 'means': []}
```



On observe une médiane de 17 ans, une valeur minimale de 15 ans et une valeur maximale de 21 ans, avec une valeur hors de boxplot de 22 ans. On va par la suite vérifier cette valeur pour déterminer s'il s'agit d'un outlier (bruit) .

[160]:
```
data[data['age']>21]
```

[160]:
```
    school sex  age address famsize Pstatus  Medu  Fedu      Mjob      Fjob  \
247     GP   M   22       U     GT3       T     3     1  services  services
674     GP   M   22       U     GT3       T     3     1  services  services

    reason guardian  traveltime  studytime  failures schoolsup famsup paid  \
247  other   mother           1          1         3        no     no   no
674  other   mother           1          1         3        no     no   no

    activities nursery higher internet romantic  famrel  freetime  goout  \
```

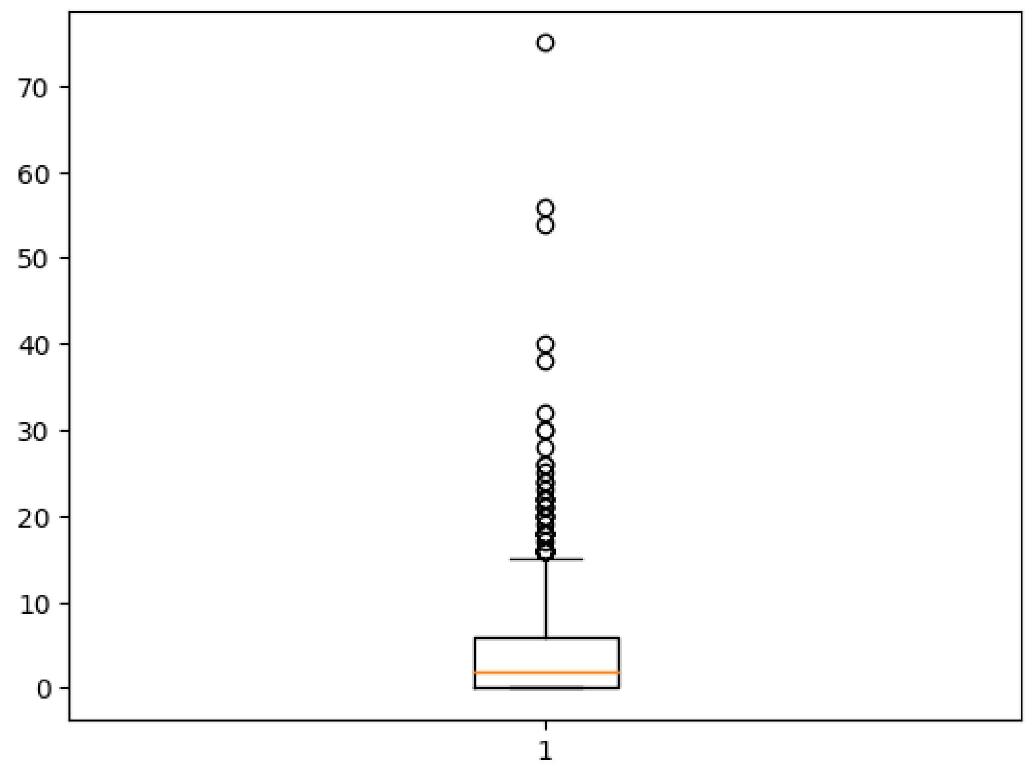
```
247          no       no      no       yes       yes        5          4       5
674          no       no      no       yes       yes        5          4       5

     Dalc  Walc  health  absences  G1  G2  G3
247     5     5       1        16   6   8   8
674     5     5       1        12   7   8   5
```

Les données correspondantes à l'age 22 semblent pas hors la norme, donc on décide de garder ce point.

1.2.2 Nombre d'absences des étudiants

[161]:
```python
plt.boxplot(data['absences'])
```

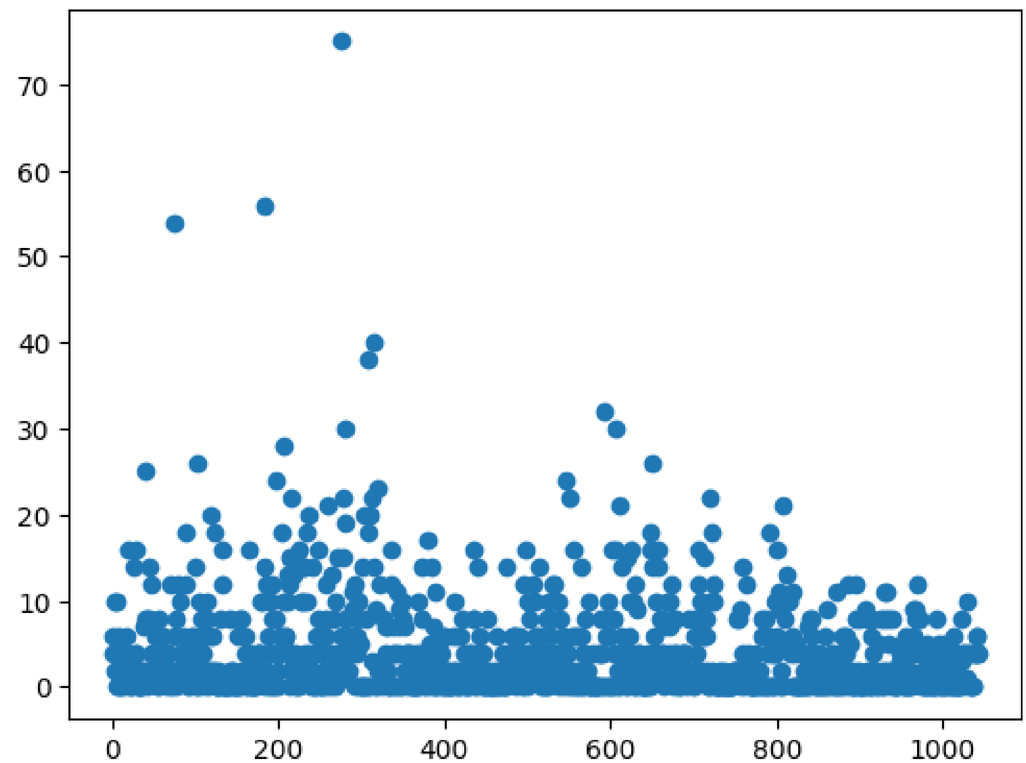
[161]:
```
{'whiskers': [<matplotlib.lines.Line2D at 0x7ff1b38603d0>,
  <matplotlib.lines.Line2D at 0x7ff1b3af8880>],
 'caps': [<matplotlib.lines.Line2D at 0x7ff1b3afb5e0>,
  <matplotlib.lines.Line2D at 0x7ff1b3afad40>],
 'boxes': [<matplotlib.lines.Line2D at 0x7ff1b3862140>],
 'medians': [<matplotlib.lines.Line2D at 0x7ff1b3afab90>],
 'fliers': [<matplotlib.lines.Line2D at 0x7ff1b33ec310>],
 'means': []}
```



On observe qu'ici il y a beaucoup de valeurs hors boxplot. On essaye de trouver s'il s'agit d'anomalies.

[163]:
```
plt.plot(data['absences'], 'o')
```

[163]:
```
[<matplotlib.lines.Line2D at 0x7ff1b7144c70>]
```



Les valeurs des absences sont plutôt concentrées entre 0 et 20.

[164]:
```
data[data['absences'] > 20]
```

[164]:
```
    school sex  age address famsize Pstatus  Medu  Fedu      Mjob      Fjob  \
40      GP   F   16       U     LE3       T     2     2     other     other
74      GP   F   16       U     GT3       T     3     3     other  services
103     GP   F   15       U     GT3       T     3     2  services     other
183     GP   F   17       U     LE3       T     3     3     other     other
198     GP   F   17       U     GT3       T     4     4  services   teacher
205     GP   F   17       U     GT3       T     3     4   at_home  services
216     GP   F   17       U     GT3       T     4     3     other     other
260     GP   F   18       U     GT3       T     4     3  services     other
276     GP   F   18       R     GT3       A     3     2     other  services
277     GP   M   18       U     GT3       T     4     4   teacher  services
280     GP   M   17       U     LE3       A     4     1  services     other
```

307	GP	M	19	U	GT3	T	4	4	teacher	services
313	GP	F	19	U	LE3	T	3	2	services	other
315	GP	F	19	R	GT3	T	2	3	other	other
320	GP	F	17	U	GT3	A	4	3	services	services
545	GP	F	15	U	GT3	A	3	3	services	services
550	GP	M	17	U	GT3	T	2	1	other	other
592	GP	F	17	U	LE3	T	3	3	other	other
607	GP	F	17	U	GT3	T	4	4	services	teacher
612	GP	F	17	R	GT3	T	2	2	other	other
651	GP	M	18	U	GT3	T	2	2	other	at_home
720	GP	M	17	U	LE3	A	4	1	services	other
808	GP	M	21	R	LE3	T	1	1	at_home	other

	reason	guardian	traveltime	studytime	failures	schoolsup	famsup \
40	home	mother	2	2	1	no	yes
74	home	mother	1	2	0	yes	yes
103	home	mother	2	2	0	yes	yes
183	reputation	mother	1	2	0	no	yes
198	home	mother	2	1	1	no	yes
205	home	mother	1	3	1	no	yes
216	reputation	mother	1	2	2	no	no
260	home	father	1	2	0	no	yes
276	home	mother	2	2	0	no	no
277	home	mother	2	1	0	no	no
280	home	mother	2	1	0	no	no
307	reputation	other	2	1	1	no	yes
313	reputation	other	2	2	1	no	yes
315	reputation	other	1	3	1	no	no
320	course	mother	1	2	0	no	yes
545	home	mother	1	2	0	no	no
550	home	mother	1	1	0	no	yes
592	reputation	mother	1	2	0	no	yes
607	home	mother	2	1	1	no	yes
612	reputation	mother	1	1	0	no	yes
651	course	other	1	1	1	no	yes
720	home	mother	2	1	0	no	no
808	course	other	2	2	2	no	yes

	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout \
40	no	yes	no	yes	yes	yes	3	3	3
74	yes	yes	yes	yes	yes	no	4	3	3
103	yes	no	yes	yes	yes	no	4	3	5
183	no	yes	yes	yes	yes	yes	5	3	3
198	no	no	yes	yes	yes	no	4	2	4
205	yes	no	yes	yes	yes	yes	4	4	3
216	yes	no	yes	yes	yes	yes	3	4	5
260	yes	no	yes	yes	yes	yes	3	1	2

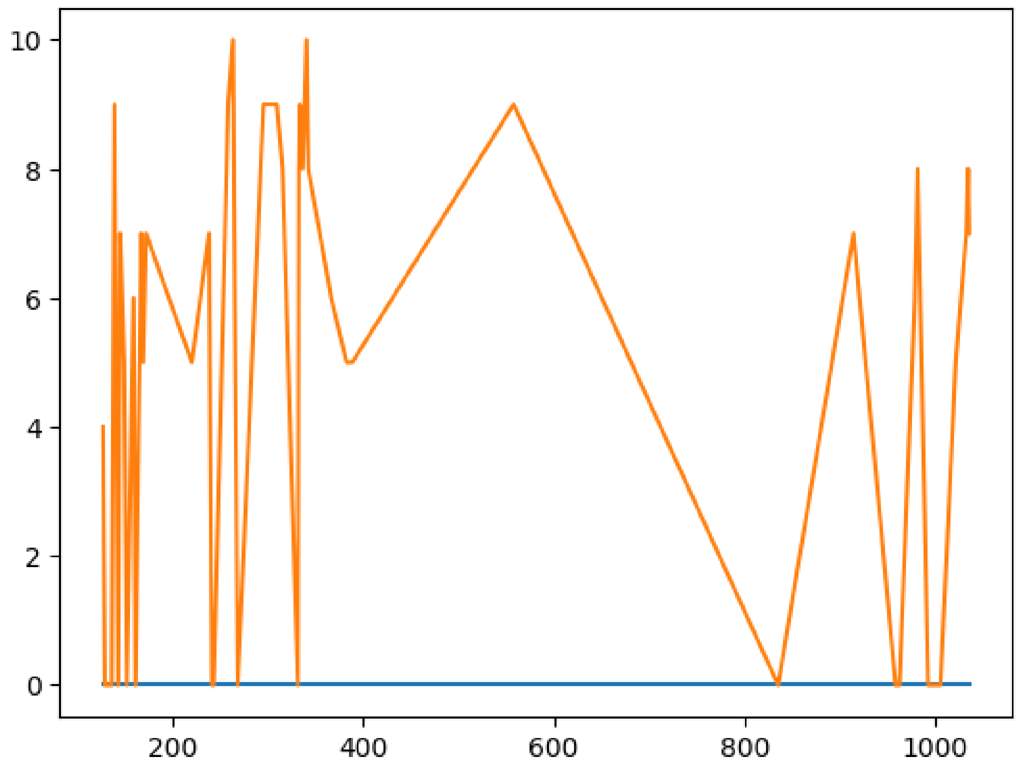
276	no	no	no	no	yes	yes	4	1	1
277	yes	yes	yes	yes	yes	no	3	2	4
280	yes	yes	yes	yes	yes	yes	4	5	4
307	yes	no	yes	yes	yes	yes	4	3	4
313	yes	no	no	yes	yes	yes	4	2	2
315	no	no	yes	yes	yes	yes	4	1	2
320	yes	no	yes	yes	yes	yes	5	2	2
545	no	no	no	yes	no	yes	1	3	2
550	no	no	yes	yes	yes	no	5	4	5
592	no	yes	yes	yes	yes	yes	5	3	3
607	no	no	yes	yes	yes	no	4	2	4
612	no	no	yes	yes	yes	no	5	3	2
651	no	yes	no	no	yes	yes	4	4	3
720	no	yes	yes	yes	yes	yes	4	5	4
808	no	yes	yes	no	yes	yes	5	3	3

	Dalc	Walc	health	absences	G1	G2	G3
40	1	2	3	25	7	10	11
74	2	4	5	54	11	12	11
103	1	1	2	26	7	6	6
183	2	3	1	56	9	9	8
198	2	3	2	24	18	18	18
205	3	4	5	28	10	9	9
216	2	4	1	22	6	6	4
260	1	3	2	21	17	18	18
276	1	1	5	75	10	9	9
277	1	4	3	22	9	9	9
280	2	4	5	30	8	8	8
307	1	1	4	38	8	9	8
313	1	2	1	22	13	10	11
315	1	1	3	40	13	11	11
320	1	2	5	23	13	13	13
545	2	3	1	24	9	8	9
550	1	2	5	22	9	7	6
592	2	3	1	32	14	13	14
607	2	3	2	30	14	15	16
612	1	2	3	21	13	13	13
651	2	2	1	26	7	8	8
720	2	4	5	22	11	11	10
808	5	2	4	21	9	10	10

On essaye de voir si G3 = 0 a une relation avec les absences.

[162]:
```
data[data['G3'] == 0 ]['absences'].plot()

data[data['G3'] == 0 ]['G2'].plot()
```

[162]: <Axes: >



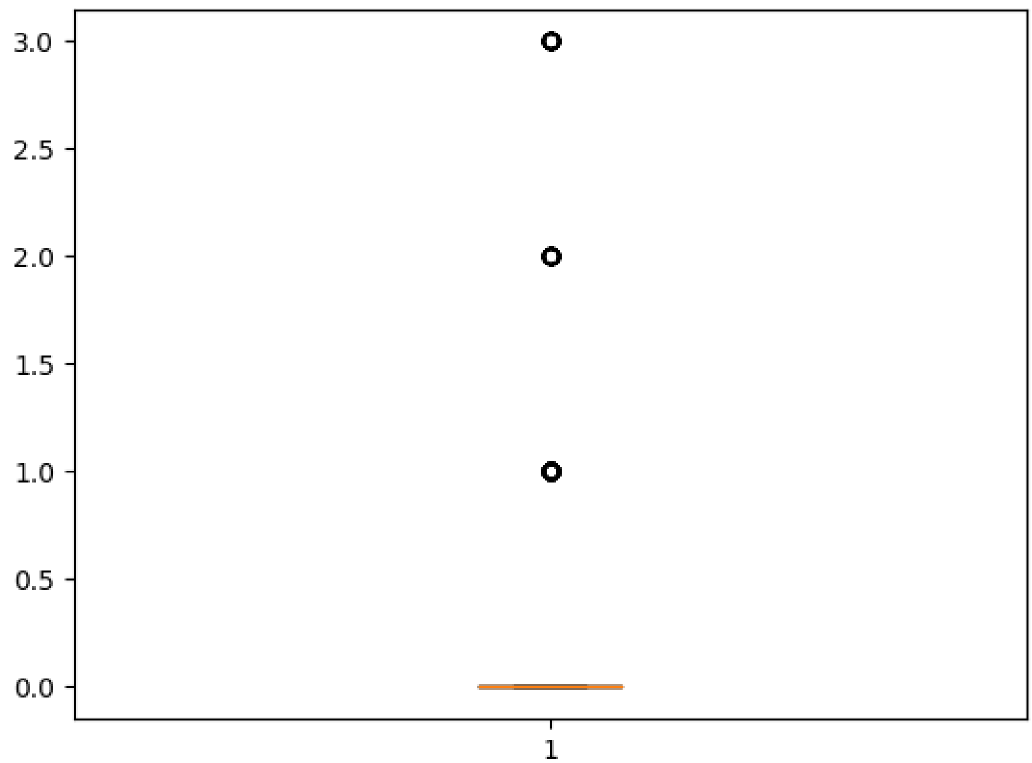
Nous ne voyons pas de relation entre les absences et le fait que G3 soit égal à 0. Nous remarquons que si G3 = 0, G2 = 0 également. Cependant, comme nous n'avons pas l'intention d'utiliser G1 et G2 comme caractéristiques, les valeurs 0 pourraient influencer négativement les prédictions. Nous décidons donc de supprimer les valeurs zéros par la suite.

Avec un nombre élevé d'absences, nous pouvons voir que G3 a des résultats variables, ce qui pourrait entraîner des erreurs dans les prédictions de nos modèles. Nous décidons d'éliminer les absences > 20 dans la partie élimination des valeurs aberrantes.

1.2.3 Nombre d'échecs

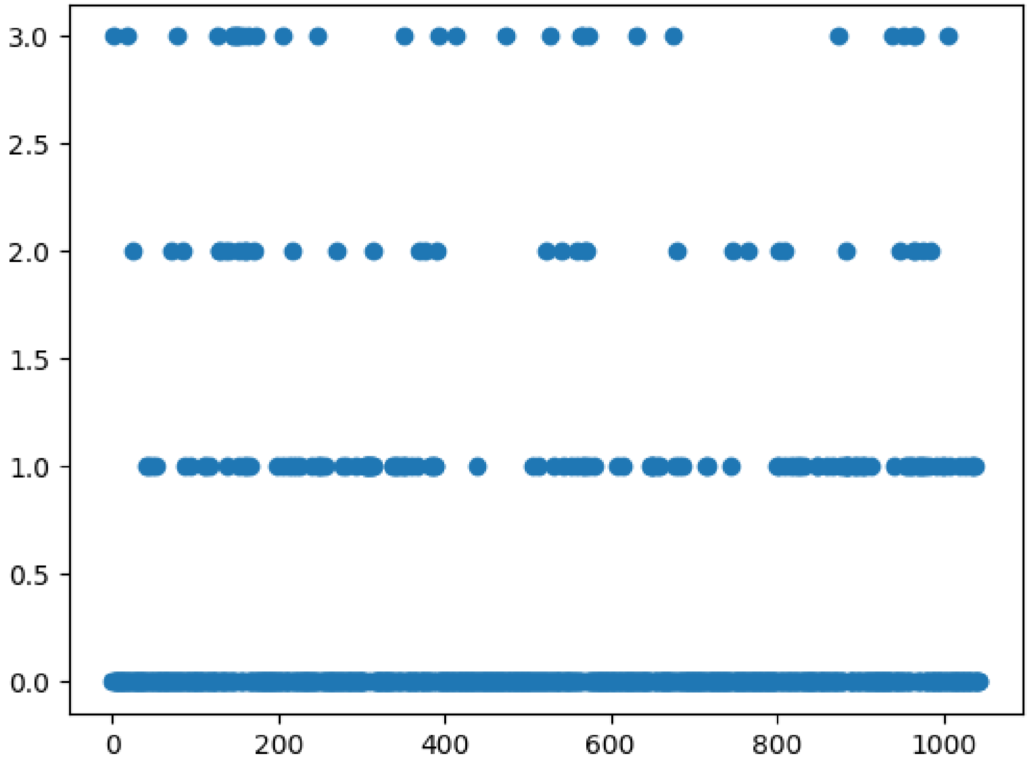
[165]:
```
plt.boxplot(data['failures'])
```

[165]:
```
{'whiskers': [<matplotlib.lines.Line2D at 0x7ff1b45c1600>,
  <matplotlib.lines.Line2D at 0x7ff1b45c2410>],
 'caps': [<matplotlib.lines.Line2D at 0x7ff1b45c1b70>,
  <matplotlib.lines.Line2D at 0x7ff1b45c2860>],
 'boxes': [<matplotlib.lines.Line2D at 0x7ff1b45c16f0>],
 'medians': [<matplotlib.lines.Line2D at 0x7ff1b45c2500>],
 'fliers': [<matplotlib.lines.Line2D at 0x7ff1b45c2740>],
 'means': []}
```




```
[166]: plt.plot(data['failures'], 'o')
```

```
[166]: [<matplotlib.lines.Line2D at 0x7ff1b4605870>]
```



[167]:
```
data[data['failures'] > 0]
```

[167]:
```
      school sex  age address famsize Pstatus  Medu  Fedu      Mjob      Fjob  \
2         GP   F   15       U     LE3       T     1     1   at_home     other
18        GP   M   17       U     GT3       T     3     2  services  services
25        GP   F   16       U     GT3       T     2     2  services  services
40        GP   F   16       U     LE3       T     2     2     other     other
44        GP   F   16       U     LE3       T     2     2     other   at_home
…        …  . .  …    …      …     …      …    …     …     …
1019      MS   F   17       R     GT3       T     1     1     other  services
1027      MS   F   19       R     GT3       T     1     1   at_home     other
1034      MS   M   19       R     GT3       T     1     1     other  services
1035      MS   M   18       R     GT3       T     4     2     other     other
1039      MS   F   19       R     GT3       T     2     3  services     other

          reason guardian  traveltime  studytime  failures schoolsup famsup  \
2          other   mother           1          2         3       yes     no
18        course   mother           1          1         3        no    yes
25          home   mother           1          1         2        no    yes
40          home   mother           2          2         1        no    yes
44        course   father           2          2         1       yes     no
…            …      …         …          …        …       …     …
```

```
1019  reputation   mother          3          1          1         no    yes
1027      course    other          2          2          1         no    yes
1034       other   mother          2          1          1         no     no
1035        home   father          2          1          1         no     no
1039      course   mother          1          3          1         no     no

     paid activities nursery higher internet romantic  famrel  freetime  \
2     yes         no     yes    yes      yes       no       4         3
18     no        yes     yes    yes      yes       no       5         5
25    yes         no      no    yes      yes       no       1         2
40     no        yes      no    yes      yes      yes       3         3
44     no        yes     yes    yes      yes       no       4         3
…    …        …      …     …      …      …        …       …
1019   no         no     yes    yes      yes      yes       5         2
1027   no         no     yes    yes      yes      yes       4         3
1034   no         no     yes    yes       no       no       4         3
1035  yes         no     yes    yes       no       no       5         4
1039   no        yes      no    yes      yes       no       5         4

      goout  Dalc  Walc  health  absences  G1  G2  G3
2         2     2     3       3        10   7   8  10
18        5     2     4       5        16   6   5   5
25        2     1     3       5        14   6   9   8
40        3     1     2       3        25   7  10  11
44        3     2     2       5        14  10  10   9
…     …    …    …      …       …  ..  ..  ..
1019      1     1     2       1         0   8   8   9
1027      3     1     1       3         4   7   8   9
1034      2     1     3       5         0   5   8   0
1035      3     4     3       3         0   7   7   0
1039      2     1     2       5         4  10  11  10

[183 rows x 33 columns]
```

[168]:
```
data[data['failures']  == 0]
```

[168]:
```
     school sex  age address famsize Pstatus  Medu  Fedu      Mjob      Fjob  \
0        GP   F   18       U     GT3       A     4     4   at_home   teacher
1        GP   F   17       U     GT3       T     1     1   at_home     other
3        GP   F   15       U     GT3       T     4     2    health  services
4        GP   F   16       U     GT3       T     3     3     other     other
5        GP   M   16       U     LE3       T     4     3  services     other
…     …  ..  …    …       …      …     …    …      …       …
1038     MS   F   18       R     GT3       T     4     4   teacher   at_home
1040     MS   F   18       U     LE3       T     3     1   teacher  services
1041     MS   F   18       U     GT3       T     1     1     other     other
1042     MS   M   17       U     LE3       T     3     1  services  services
```

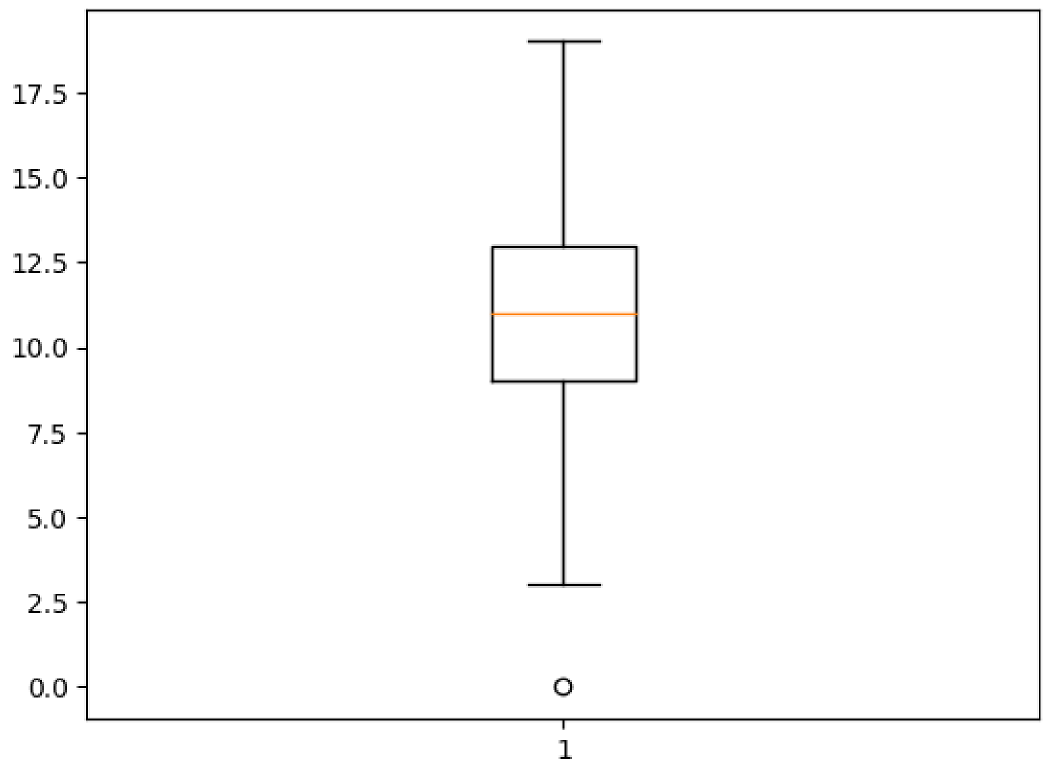
```
1043      MS   M   18       R     LE3       T     3     2  services     other

          reason guardian  traveltime  studytime  failures schoolsup famsup  \
0         course   mother           2          2         0       yes     no
1         course   father           1          2         0        no    yes
3           home   mother           1          3         0        no    yes
4           home   father           1          2         0        no    yes
5     reputation   mother           1          2         0        no    yes
…          …       …         …       …       …       …    …
1038  reputation   mother           3          1         0        no    yes
1040      course   mother           1          2         0        no    yes
1041      course   mother           2          2         0        no     no
1042      course   mother           2          1         0        no     no
1043      course   mother           3          1         0        no     no

     paid activities nursery higher internet romantic  famrel  freetime  \
0      no         no     yes    yes       no       no       4         3
1      no         no      no    yes      yes       no       5         3
3     yes        yes     yes    yes      yes      yes       3         2
4     yes         no     yes    yes       no       no       4         3
5     yes        yes     yes    yes      yes       no       5         4
…   …         …     …     …      …      …      …      …
1038   no        yes     yes    yes      yes      yes       4         4
1040   no         no     yes    yes      yes       no       4         3
1041   no        yes     yes    yes       no       no       1         1
1042   no         no      no    yes      yes       no       2         4
1043   no         no      no    yes      yes       no       4         4

      goout  Dalc  Walc  health  absences  G1  G2  G3
0         4     1     1       3         6   5   6   6
1         3     1     1       3         4   5   5   6
3         2     1     1       5         2  15  14  15
4         2     1     2       5         4   6  10  10
5         2     1     2       5        10  15  15  15
…     …    …    …      …       …  ..  ..  ..
1038      3     2     2       5         4   7   9  10
1040      4     1     1       1         4  15  15  16
1041      1     1     1       5         6  11  12   9
1042      5     3     4       2         6  10  10  10
1043      1     3     4       5         4  10  11  11

[861 rows x 33 columns]
```

On observe que la majorité des valeurs des failures sont = 0, par contre les autres valeurs ne représentent pas d'anomalies.

1.2.4 Note au premier trimestre

[169]:
```
plt.boxplot(data['G1'])
```

[169]:
```
{'whiskers': [<matplotlib.lines.Line2D at 0x7ff1b4714640>,
  <matplotlib.lines.Line2D at 0x7ff1b4714d90>],
 'caps': [<matplotlib.lines.Line2D at 0x7ff1b4714f40>,
  <matplotlib.lines.Line2D at 0x7ff1b4714280>],
 'boxes': [<matplotlib.lines.Line2D at 0x7ff1b4717370>],
 'medians': [<matplotlib.lines.Line2D at 0x7ff1b4714eb0>],
 'fliers': [<matplotlib.lines.Line2D at 0x7ff1b4b4a800>],
 'means': []}
```



[170]:
```
data[data['G1']  == 0]
```

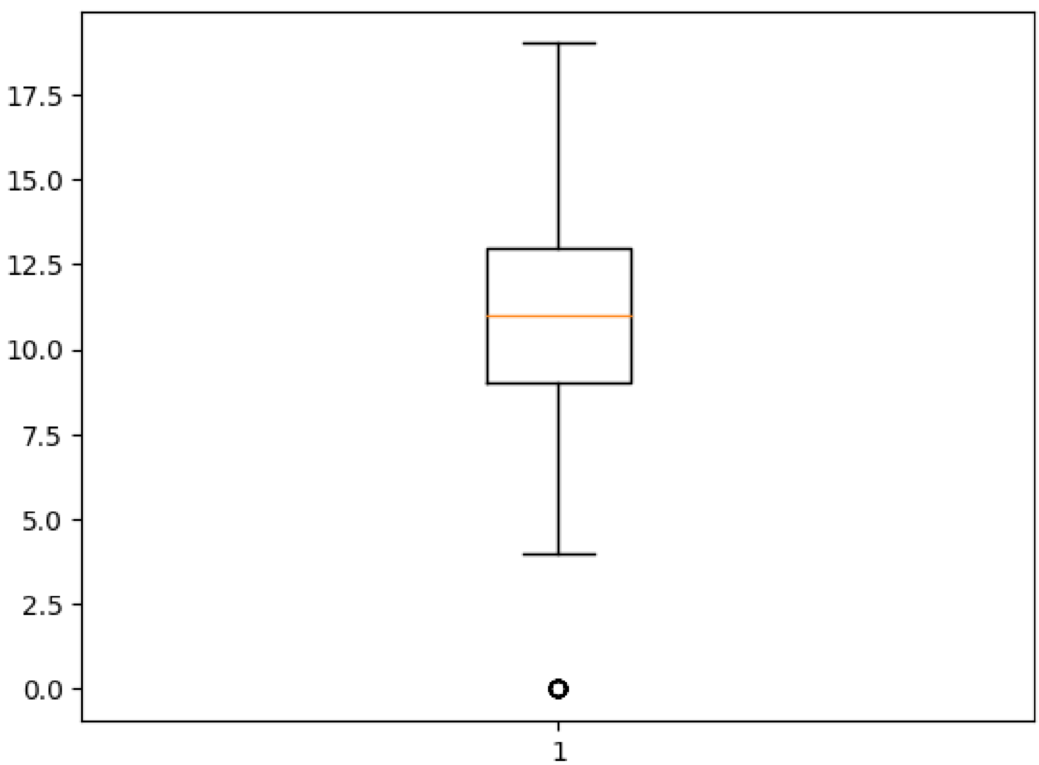
[170]:
```
    school sex  age address famsize Pstatus  Medu  Fedu     Mjob     Fjob  \
395     GP   F   18       U     GT3       A     4     4  at_home  teacher

     reason guardian  traveltime  studytime  failures schoolsup famsup paid  \
395  course   mother           2          2         0       yes     no   no

    activities nursery higher internet romantic  famrel  freetime  goout  \
```

```
395         no      yes     yes        no        no        4          3       4

     Dalc  Walc  health  absences  G1  G2  G3
395     1     1       3         4   0  11  11
```

1.2.5 Note au deuxième trimestre

[171]:
```python
plt.boxplot(data['G2'])
```

[171]:
```
{'whiskers': [<matplotlib.lines.Line2D at 0x7ff1b4665f60>,
  <matplotlib.lines.Line2D at 0x7ff1b4666200>],
 'caps': [<matplotlib.lines.Line2D at 0x7ff1b46664a0>,
  <matplotlib.lines.Line2D at 0x7ff1b4666740>],
 'boxes': [<matplotlib.lines.Line2D at 0x7ff1b4665de0>],
 'medians': [<matplotlib.lines.Line2D at 0x7ff1b46669e0>],
 'fliers': [<matplotlib.lines.Line2D at 0x7ff1b4666c80>],
 'means': []}
```



[207]:
```python
data[(data['G2'] == 0 )]
```

[207]:
```
    school sex  age address famsize Pstatus  Medu  Fedu      Mjob      Fjob  \
130      GP   F   15       R     GT3       T     3     4  services   teacher
```

131	GP	F	15	U	GT3	T	1	1	at_home	other
134	GP	M	15	R	GT3	T	3	4	at_home	teacher
135	GP	F	15	U	GT3	T	4	4	services	at_home
136	GP	M	17	R	GT3	T	3	4	at_home	other
137	GP	F	16	U	GT3	A	3	3	other	other
144	GP	M	17	U	GT3	T	2	1	other	other
153	GP	M	19	U	GT3	T	3	2	services	at_home
162	GP	M	16	U	LE3	T	1	2	other	other
242	GP	M	16	U	LE3	T	4	3	teacher	other
244	GP	F	18	U	GT3	T	2	1	other	other
269	GP	F	18	R	GT3	T	2	1	other	other
332	GP	F	18	U	GT3	T	3	3	services	services
835	MS	M	16	U	GT3	T	1	1	at_home	services
958	MS	M	17	U	GT3	T	2	2	other	other
962	MS	M	18	R	GT3	T	3	2	services	other
992	MS	F	18	R	GT3	T	2	2	at_home	other
998	MS	F	18	R	LE3	A	4	2	teacher	other
1000	MS	F	19	U	GT3	T	1	1	at_home	services
1005	MS	F	19	R	GT3	A	1	1	at_home	at_home

	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	\
130	course	father	2	3	2	no	yes	
131	course	mother	3	1	0	no	yes	
134	course	mother	4	2	0	no	yes	
135	course	mother	1	3	0	no	yes	
136	course	mother	3	2	0	no	no	
137	course	other	2	1	2	no	yes	
144	home	mother	1	1	3	no	yes	
153	home	mother	1	1	3	no	yes	
162	course	mother	2	1	1	no	no	
242	course	mother	1	1	0	no	no	
244	course	other	2	3	0	no	yes	
269	reputation	mother	2	2	0	no	yes	
332	home	mother	1	2	0	no	no	
835	home	mother	2	2	0	no	yes	
958	course	mother	1	1	1	no	no	
962	course	mother	1	1	1	no	no	
992	course	mother	3	2	1	no	no	
998	reputation	mother	1	2	0	no	no	
1000	other	father	2	1	1	no	no	
1005	course	other	2	2	3	no	yes	

	paid	activities	nursery	higher	internet	romantic	famrel	freetime	\
130	no	no	yes	yes	yes	yes	4	2	
131	no	yes	no	yes	yes	yes	4	3	
134	no	no	yes	yes	no	yes	5	3	
135	no	yes	yes	yes	yes	yes	4	3	

```
136     no    no   yes   yes    no    no   5   4
137     no   yes    no   yes   yes   yes   4   3
144     no    no   yes   yes   yes    no   5   4
153     no    no   yes    no   yes   yes   4   5
162     no   yes   yes   yes    no    no   4   4
242     no   yes    no   yes   yes    no   5   4
244    yes    no    no   yes   yes   yes   4   4
269     no    no   yes    no   yes   yes   4   3
332     no   yes   yes   yes   yes    no   5   3
835     no   yes   yes   yes    no   yes   5   4
958     no   yes   yes   yes    no   yes   1   2
962     no    no   yes    no   yes    no   2   3
992     no   yes   yes   yes    no   yes   4   3
998     no   yes   yes   yes   yes   yes   5   3
1000    no    no   yes    no    no    no   5   5
1005    no   yes   yes    no    no   yes   3   5

      goout  Dalc  Walc  health  absences  G1  G2  G3
130       2     2     2       5         0  12   0   0
131       3     1     2       4         0   8   0   0
134       3     1     1       5         0   9   0   0
135       3     1     1       5         0  11   0   0
136       5     2     4       5         0  10   0   0
137       2     1     1       5         0   4   0   0
144       5     1     2       5         0   5   0   0
153       4     1     1       4         0   5   0   0
162       4     2     4       5         0   7   0   0
242       5     1     1       3         0   6   0   0
244       4     1     1       3         0   7   0   0
269       5     1     2       3         0   6   0   0
332       4     1     1       4         0   7   0   0
835       5     4     5       3         0   7   0   0
958       1     2     3       5         0   7   0   0
962       1     2     2       5         0   4   0   0
992       3     1     1       4         0   9   0   0
998       1     1     1       5         0   5   0   0
1000      5     2     3       2         0   5   0   0
1005      4     1     4       1         0   8   0   0
```

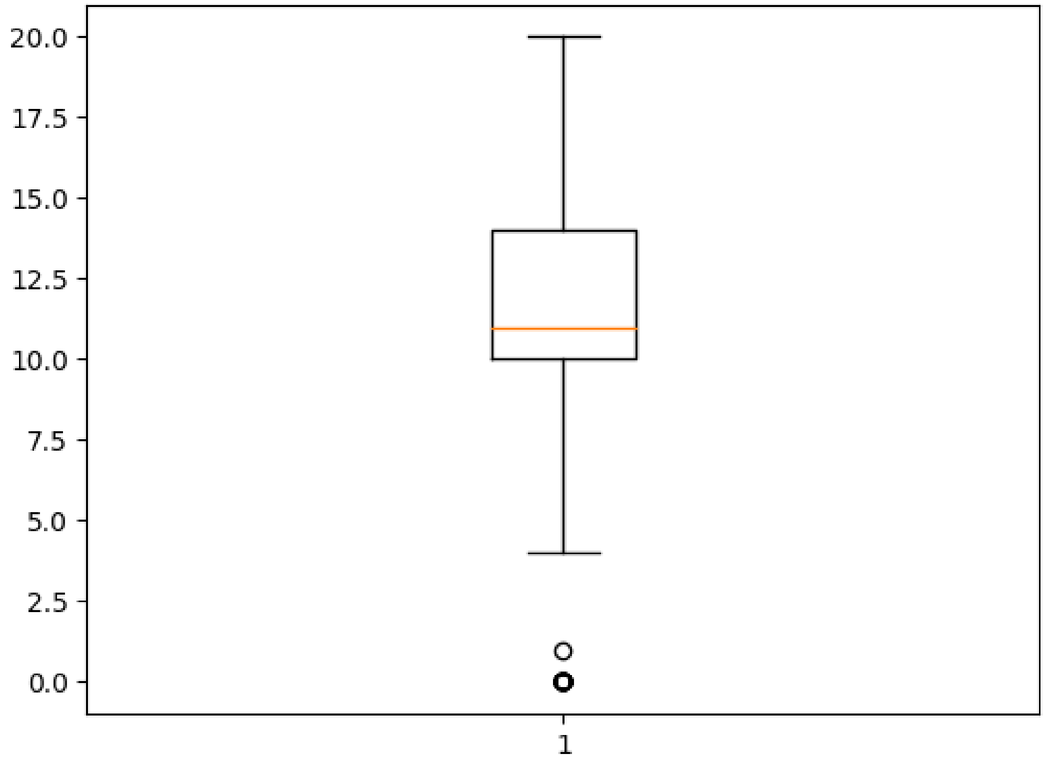
On décide de ne pas éliminer ces données car ils ne sont pas des anomalies.

1.2.6 Note finale

[173]:
```
plt.boxplot(data['G3'])
```

[173]:
```
{'whiskers': [<matplotlib.lines.Line2D at 0x7ff1b2f89db0>,
  <matplotlib.lines.Line2D at 0x7ff1b2f8a050>],
```

```
 'caps': [<matplotlib.lines.Line2D at 0x7ff1b2f8a2f0>,
  <matplotlib.lines.Line2D at 0x7ff1b2f8a590>],
 'boxes': [<matplotlib.lines.Line2D at 0x7ff1b2f89b10>],
 'medians': [<matplotlib.lines.Line2D at 0x7ff1b2f8a830>],
 'fliers': [<matplotlib.lines.Line2D at 0x7ff1b2f8aad0>],
 'means': []}
```



[208]:
```
data[data['G3']  < 2]
```

[208]:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	\
128	GP	M	18	R	GT3	T	2	2	services	other	
130	GP	F	15	R	GT3	T	3	4	services	teacher	
131	GP	F	15	U	GT3	T	1	1	at_home	other	
134	GP	M	15	R	GT3	T	3	4	at_home	teacher	
135	GP	F	15	U	GT3	T	4	4	services	at_home	
136	GP	M	17	R	GT3	T	3	4	at_home	other	
137	GP	F	16	U	GT3	A	3	3	other	other	
140	GP	M	15	U	GT3	T	4	3	teacher	services	
144	GP	M	17	U	GT3	T	2	1	other	other	
146	GP	F	15	U	GT3	T	3	2	health	services	
148	GP	M	16	U	GT3	T	4	4	teacher	teacher	
150	GP	M	18	U	LE3	T	1	1	other	other	

153	GP	M	19	U	GT3	T	3	2	services	at_home
160	GP	M	17	R	LE3	T	2	1	at_home	other
162	GP	M	16	U	LE3	T	1	2	other	other
168	GP	F	16	U	GT3	T	2	2	other	other
170	GP	M	16	U	GT3	T	3	4	other	other
173	GP	F	16	U	GT3	T	1	3	at_home	services
221	GP	F	17	U	GT3	T	1	1	at_home	other
239	GP	M	18	U	GT3	T	2	2	other	services
242	GP	M	16	U	LE3	T	4	3	teacher	other
244	GP	F	18	U	GT3	T	2	1	other	other
259	GP	F	17	U	LE3	T	2	2	services	services
264	GP	F	18	U	GT3	T	2	2	at_home	services
269	GP	F	18	R	GT3	T	2	1	other	other
296	GP	F	19	U	GT3	T	4	4	health	other
310	GP	F	19	U	LE3	T	1	2	services	services
316	GP	F	18	U	GT3	T	2	1	services	other
332	GP	F	18	U	GT3	T	3	3	services	services
333	GP	F	18	U	LE3	T	2	2	other	other
334	GP	F	18	R	GT3	T	2	2	at_home	other
337	GP	F	17	U	GT3	T	3	2	other	other
341	GP	M	18	U	GT3	T	4	4	teacher	services
343	GP	F	17	U	GT3	A	2	2	at_home	at_home
367	MS	F	17	R	GT3	T	1	1	other	services
383	MS	M	19	R	GT3	T	1	1	other	services
387	MS	F	19	R	GT3	T	2	3	services	other
389	MS	F	18	U	GT3	T	1	1	other	other
558	GP	M	18	U	LE3	T	1	1	other	other
567	GP	M	16	U	GT3	T	3	3	other	services
835	MS	M	16	U	GT3	T	1	1	at_home	services
914	MS	M	16	R	GT3	T	2	1	other	services
958	MS	M	17	U	GT3	T	2	2	other	other
962	MS	M	18	R	GT3	T	3	2	services	other
978	MS	F	18	R	GT3	T	2	2	other	other
981	MS	F	17	U	GT3	T	4	2	teacher	services
992	MS	F	18	R	GT3	T	2	2	at_home	other
998	MS	F	18	R	LE3	A	4	2	teacher	other
1000	MS	F	19	U	GT3	T	1	1	at_home	services
1005	MS	F	19	R	GT3	A	1	1	at_home	at_home
1021	MS	F	18	R	GT3	T	4	4	other	teacher
1032	MS	M	18	R	GT3	T	2	1	other	other
1034	MS	M	19	R	GT3	T	1	1	other	services
1035	MS	M	18	R	GT3	T	4	2	other	other

	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	\
128	reputation	mother	1	1	2	no	yes	
130	course	father	2	3	2	no	yes	
131	course	mother	3	1	0	no	yes	

134	course	mother	4	2	0	no	yes
135	course	mother	1	3	0	no	yes
136	course	mother	3	2	0	no	no
137	course	other	2	1	2	no	yes
140	course	father	2	4	0	yes	yes
144	home	mother	1	1	3	no	yes
146	home	father	1	2	3	no	yes
148	course	mother	1	1	0	no	yes
150	course	mother	1	1	3	no	no
153	home	mother	1	1	3	no	yes
160	course	mother	2	1	2	no	no
162	course	mother	2	1	1	no	no
168	home	mother	1	2	0	no	yes
170	course	father	3	1	2	no	yes
173	home	mother	1	2	3	no	no
221	reputation	mother	1	3	1	no	yes
239	reputation	father	1	2	1	no	no
242	course	mother	1	1	0	no	no
244	course	other	2	3	0	no	yes
259	course	father	1	4	0	no	no
264	home	mother	1	3	0	no	yes
269	reputation	mother	2	2	0	no	yes
296	reputation	other	2	2	0	no	yes
310	home	other	1	2	1	no	no
316	course	mother	2	2	0	no	yes
332	home	mother	1	2	0	no	no
333	home	other	1	2	0	no	no
334	course	mother	2	4	0	no	no
337	home	mother	1	2	0	no	yes
341	home	father	1	2	1	no	yes
343	home	father	1	2	1	no	yes
367	reputation	mother	3	1	1	no	yes
383	other	mother	2	1	1	no	no
387	course	mother	1	3	1	no	no
389	course	mother	2	2	1	no	no
558	course	mother	1	1	2	no	no
567	course	father	1	2	1	no	yes
835	home	mother	2	2	0	no	yes
914	reputation	mother	2	2	0	no	no
958	course	mother	1	1	1	no	no
962	course	mother	1	1	1	no	no
978	other	mother	2	1	1	no	no
981	home	mother	1	2	0	yes	yes
992	course	mother	3	2	1	no	no
998	reputation	mother	1	2	0	no	no
1000	other	father	2	1	1	no	no
1005	course	other	2	2	3	no	yes

1021	other	father	3	2	0	no	yes
1032	other	mother	2	1	0	no	no
1034	other	mother	2	1	1	no	no
1035	home	father	2	1	1	no	no

	paid	activities	nursery	higher	internet	romantic	famrel	freetime \
128	no	yes	yes	yes	yes	no	3	3
130	no	no	yes	yes	yes	yes	4	2
131	no	yes	no	yes	yes	yes	4	3
134	no	no	yes	yes	no	yes	5	3
135	no	yes	yes	yes	yes	yes	4	3
136	no	no	yes	yes	no	no	5	4
137	no	yes	no	yes	yes	yes	4	3
140	no	no	yes	yes	yes	no	2	2
144	no	no	yes	yes	yes	no	5	4
146	no	no	yes	yes	yes	no	3	3
148	no	no	yes	no	yes	yes	3	3
150	no	no	yes	no	yes	yes	2	3
153	no	no	yes	no	yes	yes	4	5
160	no	yes	yes	no	yes	yes	3	3
162	no	yes	yes	yes	no	no	4	4
168	yes	no	no	yes	yes	no	5	1
170	no	yes	no	yes	yes	no	3	4
173	no	yes	no	yes	yes	yes	4	3
221	no	yes	yes	yes	no	yes	4	3
239	no	no	yes	no	yes	no	5	5
242	no	yes	no	yes	yes	no	5	4
244	yes	no	no	yes	yes	yes	4	4
259	yes	yes	yes	yes	yes	yes	3	4
264	yes	yes	yes	yes	yes	yes	4	3
269	no	no	yes	no	yes	yes	4	3
296	yes	yes	yes	yes	yes	no	2	3
310	no	yes	no	yes	no	yes	4	2
316	yes	yes	yes	yes	yes	no	5	3
332	no	yes	yes	yes	yes	no	5	3
333	no	yes	no	yes	yes	yes	4	3
334	no	yes	yes	yes	no	no	4	4
337	yes	no	yes	yes	yes	yes	4	3
341	no	yes	yes	yes	yes	no	4	3
343	no	no	yes	yes	yes	yes	3	3
367	yes	no	yes	yes	yes	yes	5	2
383	no	no	yes	yes	no	no	4	3
387	no	yes	no	yes	yes	no	5	4
389	no	yes	yes	yes	no	no	1	1
558	no	no	yes	no	yes	yes	2	3
567	no	no	yes	yes	yes	yes	4	5
835	no	yes	yes	yes	no	yes	5	4

914	no	yes	yes	yes	yes	no	5	2
958	no	yes	yes	yes	no	yes	1	2
962	no	no	yes	no	yes	no	2	3
978	no	no	yes	no	yes	yes	5	5
981	no	yes	yes	yes	yes	no	5	5
992	no	yes	yes	yes	no	yes	4	3
998	no	yes	yes	yes	yes	yes	5	3
1000	no	no	yes	no	no	no	5	5
1005	no	yes	yes	no	no	yes	3	5
1021	no	no	no	yes	yes	yes	3	2
1032	no	yes	no	yes	yes	yes	4	4
1034	no	no	yes	yes	no	no	4	3
1035	yes	no	yes	yes	no	no	5	4

	goout	Dalc	Walc	health	absences	G1	G2	G3
128	3	1	2	4	0	7	4	0
130	2	2	2	5	0	12	0	0
131	3	1	2	4	0	8	0	0
134	3	1	1	5	0	9	0	0
135	3	1	1	5	0	11	0	0
136	5	2	4	5	0	10	0	0
137	2	1	1	5	0	4	0	0
140	2	1	1	3	0	7	9	0
144	5	1	2	5	0	5	0	0
146	2	1	1	3	0	6	7	0
148	2	2	1	5	0	7	6	0
150	5	2	5	4	0	6	5	0
153	4	1	1	4	0	5	0	0
160	2	2	2	5	0	7	6	0
162	4	2	4	5	0	7	0	0
168	5	1	1	4	0	6	7	0
170	5	2	4	2	0	6	5	0
173	5	1	1	3	0	8	7	0
221	4	1	1	5	0	6	5	0
239	4	3	5	2	0	7	7	0
242	5	1	1	3	0	6	0	0
244	4	1	1	3	0	7	0	0
259	1	1	1	2	0	10	9	0
264	3	1	1	3	0	9	10	0
269	5	1	2	3	0	6	0	0
296	4	2	3	2	0	10	9	0
310	4	2	2	3	0	9	9	0
316	3	1	2	1	0	8	8	0
332	4	1	1	4	0	7	0	0
333	3	1	1	2	0	8	8	0
334	4	1	1	4	0	10	9	0
337	2	2	3	2	0	7	8	0

341	3	2	2	2	0	10	10	0
343	1	1	2	4	0	9	8	0
367	1	1	2	1	0	7	6	0
383	2	1	3	5	0	6	5	0
387	2	1	2	5	0	7	5	0
389	1	1	1	5	0	6	5	0
558	5	2	5	4	0	11	9	0
567	5	4	4	5	0	10	10	1
835	5	4	5	3	0	7	0	0
914	1	1	1	2	0	8	7	0
958	1	2	3	5	0	7	0	0
962	1	2	2	5	0	4	0	0
978	5	1	1	3	0	8	6	0
981	5	1	3	5	0	8	8	0
992	3	1	1	4	0	9	0	0
998	1	1	1	5	0	5	0	0
1000	5	2	3	2	0	5	0	0
1005	4	1	4	1	0	8	0	0
1021	2	4	2	5	0	7	5	0
1032	3	1	3	5	0	7	7	0
1034	2	1	3	5	0	5	8	0
1035	3	4	3	3	0	7	7	0

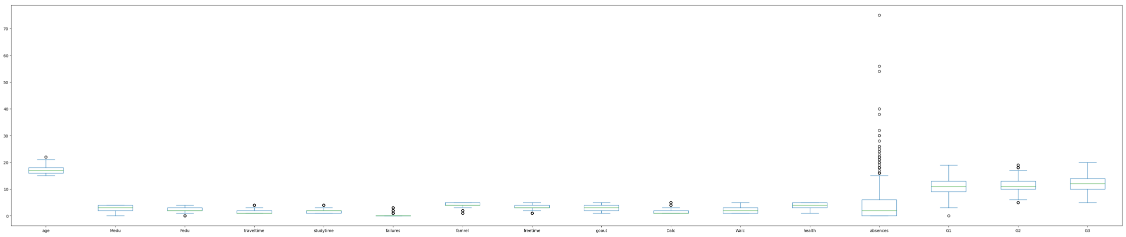
On décide de ne pas éliminer ces données car ils ne sont pas des anomalies.

1.2.7 Tous les boxplots et filtrage initial

Sélectionner les dtypes de données

[175]:
```
result = data.select_dtypes(include='number')

for i in result.columns:
    percentile25 = data[i].quantile(0.25)
    percentile75 = data[i].quantile(0.75)

    iqr = percentile75-percentile25

    upper_limit = percentile75 + 1.5 * iqr
    lower_limit = percentile25 - 1.5 * iqr

    data[data[i] > upper_limit]
    data[data[i] < lower_limit]

    dataset_new = data[data[i] < upper_limit ]
    dataset_new = data[data[i] > lower_limit ]
dataset_new.plot(kind='box',figsize=(50,10))
```

[175]:
```
<Axes: >
```



On a représenté tous les boxplots ensemble et on a décidé de supprimer les outliers par la méthode d'IQR.

L'écart interquartile (IQR) est une mesure de l'étendue des données. Elle est calculée en soustrayant le 25ème percentile du 75ème percentile des données. Les points de données qui se trouvent en dehors d'une certaine plage (par exemple, 1,5 fois l'IQR) peuvent être considérés comme des valeurs aberrantes.

2 Création d'un dataset filtré

2.0.1 Filtre des zeros à G3

Nous avons remarqué dans les analyses ci dessus que de nombreux étudiants ont eu la note de 0 pour leur note finale. Cette note est dans la majorité des cas due à une absence lors de l'examen ou même encore à une valeur manquante. Nous avons donc choisi de retirer des données les étudiants qui ont eu un 0 comme note finale.

[176]:
```
filtered_data = data[(data.G3 != 0) & (data.absences <=20)]
```

2.0.2 Transformation des catégories non-numériques en catégories binaire en utilisant des dummies

Plusieurs des features sont des litéraux. Il faut donc les encoder dans des feature "dummies" numériques pour pouvoir les exploiter.

[177]:
```
filtered_data_with_dummies = pd.get_dummies(filtered_data,drop_first=True)
filtered_data_with_dummies.head()
```

[177]:
```
   age  Medu  Fedu  traveltime  studytime  failures  famrel  freetime  goout  \
0   18     4     4           2          2         0       4         3      4
1   17     1     1           1          2         0       5         3      3
2   15     1     1           1          2         3       4         3      2
3   15     4     2           1          3         0       3         2      2
4   16     3     3           1          2         0       4         3      2

   Dalc  Walc  health  absences  G1  G2  G3  school_MS  sex_M  address_U  \
0     1     1       3         6   5   6   6          0      0          1
1     1     1       3         4   5   5   6          0      0          1
2     2     3       3        10   7   8  10          0      0          1
```

```
3     1     1        5         2  15  14  15           0      0          1
4     1     2        5         4   6  10  10           0      0          1

   famsize_LE3  Pstatus_T  Mjob_health  Mjob_other  Mjob_services  \
0            0          0            0           0              0
1            0          1            0           0              0
2            1          1            0           0              0
3            0          1            1           0              0
4            0          1            0           1              0

   Mjob_teacher  Fjob_health  Fjob_other  Fjob_services  Fjob_teacher  \
0             0            0           0              0             1
1             0            0           1              0             0
2             0            0           1              0             0
3             0            0           0              1             0
4             0            0           1              0             0

   reason_home  reason_other  reason_reputation  guardian_mother  \
0            0             0                  0                1
1            0             0                  0                0
2            0             1                  0                1
3            1             0                  0                1
4            1             0                  0                0

   guardian_other  schoolsup_yes  famsup_yes  paid_yes  activities_yes  \
0               0              1           0         0               0
1               0              0           1         0               0
2               0              1           0         1               0
3               0              0           1         1               1
4               0              0           1         1               0

   nursery_yes  higher_yes  internet_yes  romantic_yes
0            1           1             0             0
1            0           1             1             0
2            1           1             1             0
3            1           1             1             1
4            1           1             0             0
```

2.0.3 Extraction de la cible

[178]:
```python
y_filtered_data = filtered_data_with_dummies['G3']
y_filtered_data.head()
```

[178]:
```
0     6
1     6
2    10
3    15
```

```
4    10
Name: G3, dtype: int64
```

2.0.4 Extraction des features

[179]:
```
X_filtered_data = filtered_data_with_dummies.drop(['G3','G2','G1'],axis=1)
X_filtered_data.head()
```

[179]:
```
   age  Medu  Fedu  traveltime  studytime  failures  famrel  freetime  goout  \
0   18     4     4           2          2         0       4         3      4
1   17     1     1           1          2         0       5         3      3
2   15     1     1           1          2         3       4         3      2
3   15     4     2           1          3         0       3         2      2
4   16     3     3           1          2         0       4         3      2

   Dalc  Walc  health  absences  school_MS  sex_M  address_U  famsize_LE3  \
0     1     1       3         6          0      0          1            0
1     1     1       3         4          0      0          1            0
2     2     3       3        10          0      0          1            1
3     1     1       5         2          0      0          1            0
4     1     2       5         4          0      0          1            0

   Pstatus_T  Mjob_health  Mjob_other  Mjob_services  Mjob_teacher  \
0          0            0           0              0             0
1          1            0           0              0             0
2          1            0           0              0             0
3          1            1           0              0             0
4          1            0           1              0             0

   Fjob_health  Fjob_other  Fjob_services  Fjob_teacher  reason_home  \
0            0           0              0             1            0
1            0           1              0             0            0
2            0           1              0             0            0
3            0           0              1             0            1
4            0           1              0             0            1

   reason_other  reason_reputation  guardian_mother  guardian_other  \
0             0                  0                1               0
1             0                  0                0               0
2             1                  0                1               0
3             0                  0                1               0
4             0                  0                0               0

   schoolsup_yes  famsup_yes  paid_yes  activities_yes  nursery_yes  \
0              1           0         0               0            1
1              0           1         0               0            0
2              1           0         1               0            1
```

```
3              0            1          1                1             1
4              0            1          1                0             1

   higher_yes  internet_yes  romantic_yes
0           1             0             0
1           1             1             0
2           1             1             0
3           1             1             1
4           1             0             0
```

3 Création des jeu d'entrainement et de validation

[180]:
```
X_filtered_train, X_filtered_test, y_filtered_train, y_filtered_test = 
 ↪train_test_split(X_filtered_data, y_filtered_data, test_size=0.
 ↪2,random_state=2023)
```

3.1 Régression linéaire simple

Nous construisons d'abord un modèle de régression linéaire car c'est le modèle le plus simple et le plus facilement interprétable.

[181]:
```
regFiltered = LinearRegression().fit(X_filtered_train, y_filtered_train)
regFiltered.score(X_filtered_train, y_filtered_train)
```

[181]: 0.32839393127148553

[182]:
```
regFiltered.score(X_filtered_test,y_filtered_test)
```

[182]: 0.29092295912393107

[183]:
```
y_pred = regFiltered.predict(X_filtered_test)
```

[184]:
```
# Calculate mean squared error and R-squared score
mse = mean_squared_error(y_filtered_test, y_pred)
r2 = r2_score(y_filtered_test, y_pred)

print("Mean squared error: ", mse)
print("Mean error", math.sqrt(mse))
print("R-squared score: ", r2)
```

```
Mean squared error:  6.211621137892372
Mean error 2.492312407763596
R-squared score:  0.29092295912393107
```

3.1.1 Validation croisée

[187]:
```
kfold = KFold(n_splits=10, shuffle = True)
cv_results_filtered = cross_val_score(LinearRegression(), X_filtered_train,␣
 ↪y_filtered_train, cv=kfold, scoring='neg_mean_absolute_error')
print(f"{cv_results_filtered.mean():.2f} {cv_results_filtered.std():.2f}")
```

```
-1.96 0.23
```

3.2 Regression polynomiale à régularisation Ridge

Création d'un fonction custom pour générer un modèle polynomial à régularisation Ridge. Une normalization est appliquée avant.

[188]:
```
def polynomial_ridge_regression(degree, alpha):
    return make_pipeline(StandardScaler(), PolynomialFeatures(degree),␣
 ↪Ridge(alpha=alpha))
```

Validation croisée

[189]:
```
# Create a range of degrees and alphas for cross-validation
degrees = np.arange(1, 6)
alphas = np.logspace(-4, 4, 9)
```

[190]:
```
# Initialize GridSearchCV with the custom model, hyperparameters, and the␣
 ↪number of folds for cross-validation
grid_search = GridSearchCV(polynomial_ridge_regression(None, None),
                           param_grid={'polynomialfeatures__degree': degrees,
                                       'ridge__alpha': alphas},
                           scoring='neg_mean_squared_error',
                           cv=5)

# Fit the grid search using the training data
grid_search.fit(X_filtered_train, y_filtered_train)

best_degree = grid_search.best_params_['polynomialfeatures__degree']
best_alpha = grid_search.best_params_['ridge__alpha']

print("Best degree: ", best_degree)
print("Best alpha: ", best_alpha)
```

```
Best degree:  2
Best alpha:  1000.0
```

[194]:
```
model_polynomial_ridge = polynomial_ridge_regression(best_degree, best_alpha)
model_polynomial_ridge.fit(X_filtered_train, y_filtered_train)
```

[194]:
```
Pipeline(steps=[('standardscaler', StandardScaler()),
                ('polynomialfeatures', PolynomialFeatures()),
                ('ridge', Ridge(alpha=1000.0))])
```

[195]:
```
y_pred_polynomial_ridge = model_polynomial_ridge.predict(X_filtered_test)
```

[196]:
```
# Calculate mean squared error and R-squared score
mse = mean_squared_error(y_filtered_test, y_pred_polynomial_ridge)
r2 = r2_score(y_filtered_test, y_pred_polynomial_ridge)

print("Mean squared error: ", mse)
print("Mean error", math.sqrt(mse))
print("R-squared score: ", r2)
```

```
Mean squared error:  5.343158845522805
Mean error 2.311527383684391
R-squared score:  0.39006079427445106
```

Le meilleur paramètre de degré du polynome est de 2 avec un alpha de 1000. Le modèle ainsi obtenu à un score R2 de 0.39 ce qui est mieux de 0.1 par rapport au modèle linéaire.

3.3 Regression random forest

[200]:
```
n_estimators_range = [10, 25, 50, 75, 100, 200, 400]
cv_scores = []

for n_estimators in n_estimators_range:
    model = RandomForestRegressor(n_estimators=n_estimators, random_state=42)
    scores = cross_val_score(model, X_filtered_train, y_filtered_train, cv=5,␣
 ↪scoring='neg_mean_squared_error')
    cv_scores.append(np.mean(scores))

# Find the best n_estimators based on the highest cross-validation score
best_n_estimators = n_estimators_range[np.argmax(cv_scores)]
print("Best n_estimators: ", best_n_estimators)
```

```
Best n_estimators:  400
```

[201]:
```
model_random_forest = RandomForestRegressor(n_estimators=best_n_estimators,␣
 ↪random_state=0)
model_random_forest.fit(X_filtered_train, y_filtered_train)
```

[201]:
```
RandomForestRegressor(n_estimators=400, random_state=0)
```

[202]:
```
y_pred_random_forest = model_random_forest.predict(X_filtered_test)
```

[203]:
```
# Calculate mean squared error and R-squared score
mse = mean_squared_error(y_filtered_test, y_pred_random_forest)
r2 = r2_score(y_filtered_test, y_pred_random_forest)

print("Mean squared error: ", mse)
print("Mean error", math.sqrt(mse))
print("R-squared score: ", r2)
```

```
Mean squared error:  4.957514442076605
Mean error 2.2265476509782145
R-squared score:  0.43408337491091786
```

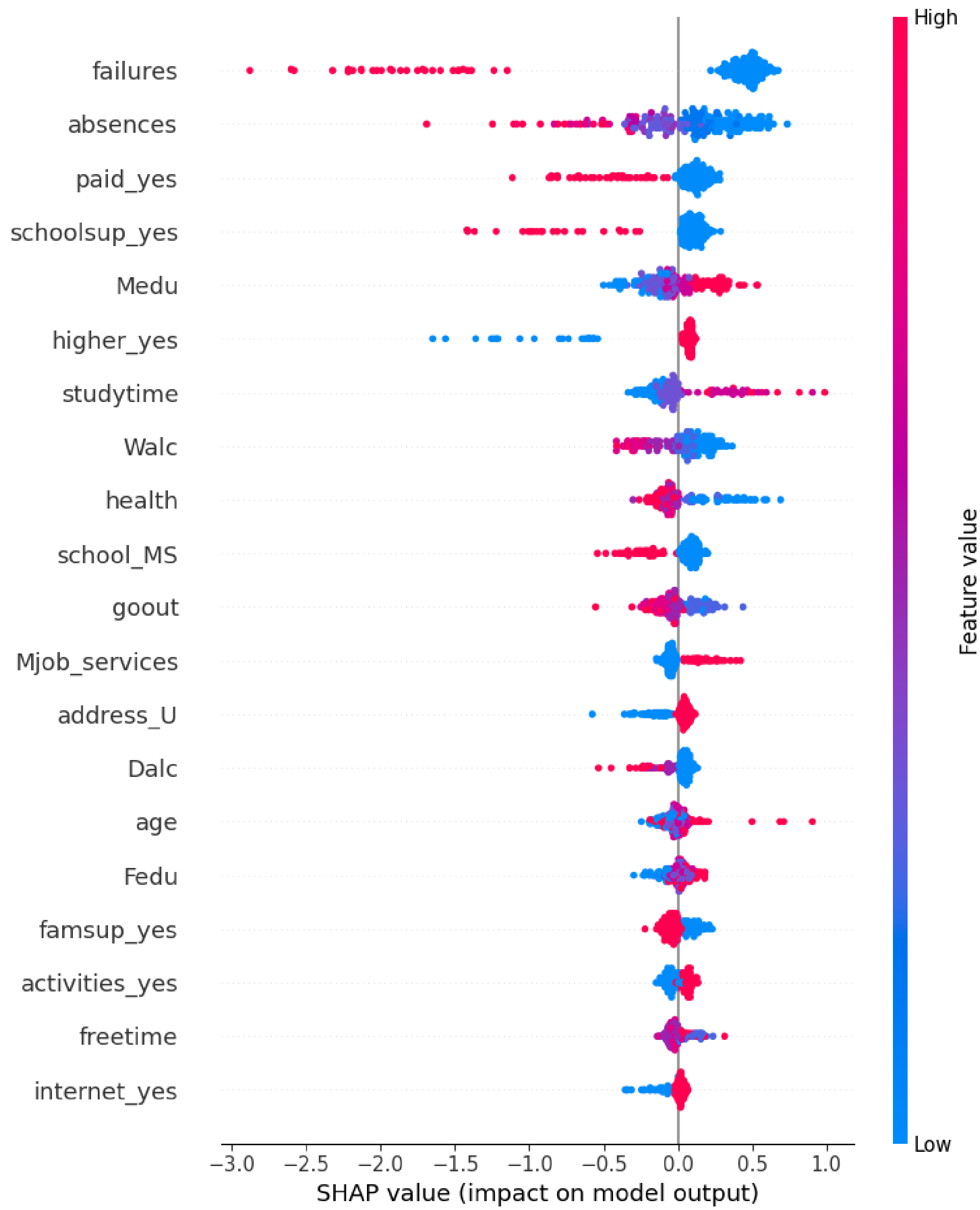
On obtien alors un modèle avec un R2 de 0.43 soit 0.04 points de plus que le modèle polynomial précédent.

4 Explication du modèle random forest en utilisant shap

[197]:
```
explainer = shap.Explainer(model_random_forest)
shap_values = explainer(X_filtered_test)
```

[198]:
```
shap.summary_plot(shap_values, X_filtered_test)
```

```
No data for colormapping provided via 'c'. Parameters 'vmin', 'vmax' will be
ignored
```



On peut alors voir que les features les plus importantes pour décider de la note d'un étudiant sont : son nombre passé d'echecs scolaire, son nombre d'absences, les cours supplémentaires en dehors de l'école, le niveau d'éducation de la mère, le temps passé à étudier chaque semaine ou bien encore la consommation d'alcool.

[199]:
```python
instance_index = 0
shap.initjs()
shap.force_plot(explainer.expected_value, shap_values.values[instance_index],
    X_filtered_test.iloc[instance_index])
```

```
<IPython.core.display.HTML object>
```

[199]:
```
<shap.plots._force.AdditiveForceVisualizer at 0x7ff1b2de77c0>
```

On peut voir ici pour une prédiction individuelle les contributions individuelle de chaque paramètres sur la prédiction de sa note finale. on retrouve les features vu dans l'analyse précedente.

5 Conclusion

Pour conclure, parmis les modèles que nous avons entrainés, le modèle random forest est celui qui obtient les meilleurs résultats. Cependant le R2 du modèle est de 0.43 ce qui signifie qu'il explique 43% de la variance des données. Ce n'est pas suffisant pour que l'on puisse utiliser le modèle pour des tâches prédictives qui ont besoin d'être fiables (Si on avait utilisé G1 et G2, on aurait bien évidemment obtenu de meilleurs résultats de prédiction) . Cependant on notera que le modèle s'explique plutôt bien comme vu grâce à l'outil shap. Cette capacité peut permettre de tout de même utiliser le modèle afin de mieux comprendre ce qui pourrai nuire à la note finale d'un étudiant (par exemple -> beaucoup d'absences -> le modèle applique une forte pénalité à la note finale -> il y a tout intéret à réduire le nombre d'abcences de l'élève). Pour continuer l'étude, on aurrai pu essayer d'entrainer un modèle de réseaux de neurone simple ou analyser plus finement les données pour vérifier s'il n'y a pas des outliers restant.