Rapport dataset Projet ANDO

C. GAUTHERET, G. HUNOUT, V. TOURNIER, B. GALIBERT, Y. BOICRA

October 23, 2024

1 Contexte

Le dataset que nous avons choisi est une collection complète de données visant à analyser les facteurs influençant la performance des étudiants selon différents paramètres. Ce jeu de données regroupe une variété d'attributs incluant des informations démographiques, des habitudes d'étude, la participation à des activités extrascolaires, et les résultats académiques d'un échantillon d'étudiants. L'objectif de ce dataset est de dégager des corrélations entre les origines, les comportements et la réussite scolaire des étudiants.

2 Description du dataset

Ce dataset comporte 15 attributs et 2392 observations. Cette quantité de données semble raisonnable ($10^3 >> 10^1$) ce qui permettra d'obtenir des résultats pertinents.

- *StudentID* : Identifiant unique pour chaque étudiant (ex. : 2737, 1403).
- *Age* : Âge de l'étudiant (ex. : 16, 17, 18).
- *Gender* : Genre de l'étudiant (0 pour femme, 1 pour homme).
- *Ethnicity* : Représentation catégorielle de l'origine ethnique de l'étudiant (ex. : 0, 1, 3 ; chaque numéro représentant un groupe différent).
- *ParentalEducation* : Niveau d'éducation des parents (échelle de 1 à 5, du moins au plus éduqué).
- *StudyTimeWeekly* : Temps d'étude hebdomadaire en heures (ex. : 10.82, 11.51).
- *Absences* : Nombre d'absences (ex. : 3, 18).
- *Tutoring* : Indique si l'étudiant bénéficie de tutorat (0 pour non, 1 pour oui).
- *ParentalSupport* : Niveau de soutien parental (échelle de 0 à 3, de aucun à élevé).
- *Extracurricular* : Participation à des activités extrascolaires (0 pour non, 1 pour oui).
- *Sports* : Participation à des activités sportives (0 pour non, 1 pour oui).
- *Music* : Participation à des activités musicales (0 pour non, 1 pour oui).
- *Volunteering* : Participation à des activités bénévoles (0 pour non, 1 pour oui).
- *GPA* : Moyenne générale de l'étudiant (ex. : 1.25, 2.85).
- *GradeClass* : Classification annuelle des notes (ex. : 2.0, 4.0 ; sur une échelle de 1 à 5).

3 Summray du dataset

	StudentID	Age	Gender	Ethnicity	ParentalEducation	\
count	2392.000000	2392.000000	2392.000000	2392.000000	2392.000000	
mean	2196.500000	16.468645	0.510870	0.877508	1.746237	
std	690.655244	1.123798	0.499986	1.028476	1.000411	

```
min    1001.000000    15.000000     0.000000     0.000000           0.000000
25%    1598.750000    15.000000     0.000000     0.000000           1.000000
50%    2196.500000    16.000000     1.000000     0.000000           2.000000
75%    2794.250000    17.000000     1.000000     2.000000           2.000000
max    3392.000000    18.000000     1.000000     3.000000           4.000000

       StudyTimeWeekly     Absences     Tutoring  ParentalSupport  \
count      2392.000000  2392.000000  2392.000000      2392.000000
mean          9.771992    14.541388     0.301421         2.122074
std           5.652774     8.467417     0.458971         1.122813
min           0.001057     0.000000     0.000000         0.000000
25%           5.043079     7.000000     0.000000         1.000000
50%           9.705363    15.000000     0.000000         2.000000
75%          14.408410    22.000000     1.000000         3.000000
max          19.978094    29.000000     1.000000         4.000000

       Extracurricular       Sports        Music  Volunteering          GPA  \
count      2392.000000  2392.000000  2392.000000   2392.000000  2392.000000
mean          0.383361     0.303512     0.196906      0.157191     1.906186
std           0.486307     0.459870     0.397744      0.364057     0.915156
min           0.000000     0.000000     0.000000      0.000000     0.000000
25%           0.000000     0.000000     0.000000      0.000000     1.174803
50%           0.000000     0.000000     0.000000      0.000000     1.893393
75%           1.000000     1.000000     0.000000      0.000000     2.622216
max           1.000000     1.000000     1.000000      1.000000     4.000000

        GradeClass
count  2392.000000
mean      2.983696
std       1.233908
min       0.000000
25%       2.000000
50%       4.000000
75%       4.000000
max       4.000000
```

On peut ici dégager quelques informations utiles de ce premier summary. En regardant les moyennes, on constate : les plupart des étudiants ont 16.5 ans, il y a légèrement plus d'hommes que de femmes, ils étudients en moyenne 9.7h par semaines, moins de la moitié pratique une activité extrascolaire

Source : https://openml.org/search?type=data&status=active&id=46255&sort=runs