Notes en mathématiques d'élèves du secondaire

Les facteurs de succès scolaire, notamment en mathématiques figurent parmi les questions les plus controversées dans le domaine de l'éducation. Ce rapport se penche sur cette question.

Le présent rapport explore un certain nombre de questions autour de ce thème :

- Comment évoluent les notes des élèves au cours d'une année scolaire ?
- Parmi les possibles facteurs explicatifs sélectionnés, certains sont-ils corrélés de manière significative avec les résultats d'un élève en mathématiques ?
- Est-il possible de prédire la note finale (variable « **G3 »** ci-dessous) d'un élève en mathématiques à partir de ces facteurs ?

Les données ont été recueillies grâce à une enquête qui a rassemblé les réponses de 395 élèves du secondaire au Portugal.

La source du jeu de données est :

https://www.kaggle.com/uciml/student-alcohol-consumption

Le jeu de données initial regroupe 33 variables. Toutefois, dans le cadre de ce projet, 12 variables seront retenues. La description des variables présélectionnées est la suivante :

G1: Note en mathématiques de la période 1 (de 0 à 20)

G2: Note en mathématiques de la période 2 (de 0 à 20)

G3: Note finale en mathématiques (de 0 à 20)

Medu: Niveau d'éducation de la mère (0 : aucun, 1 : éducation primaire, 2 : collège, 3 : éducation secondaire, 4 : éducation supérieure)

Fedu : Niveau d'éducation de la mère (0 : aucun, 1 : éducation primaire, 2 : collège, 3 : éducation secondaire, 4 : éducation supérieure)

Studytime: temps passé à étudier par semaine (1 : inférieur à 2h, 2 : de 2h à 5h, 3 : de 5h à 10h, 4 : plus de 10h)

failures: nombre de redoublements par le passé (borne à 4, même si le nombre de redoublement est supérieur à 4)

Famrel: Qualité des relations familiales (de 1 : très mauvaises à 5 : très bonnes)

Absences: nombre d'absences à l'école (de 0 à 93)

goout: sorties entre amis (de 1: très rares à 5: très régulières)

age: âge de l'élève (de 15 à 22 ans)

Walc: consommation d'alcool durant le week-end (de 1: très rare à 5: très régulière)

Prétraitement des données: Les traitements suivants ont été effectués sur les données.

- **Vérification de l'absence de NaN dans les données** : aucune variable ne contenait de valeurs NaN
- **Vérification du type des variables censées être numériques** : les variables de type numériques sont déjà toutes sous format numérique exploitable
- **Vérification de la dispersion des variables (transformation logarithmique) :** comme présenté dans leur description, les variables sont toutes contenues dans un intervalle assez restreint, mis à part le nombre d'absences. La représentation sous forme de nuage de point de cette variable suggère la création d'une nouvelle variable log_absences qui est la transformation logarithmique de la variable (absences +1). Le 1 a été rajouté car certains élèves ont eu 0 absences sur la période.
- **Création de nouvelles variables** :
    - **Création d'une variable « evol »**, qui vaut G3 – G1, et qui matérialise l'évolution de la note d'un élève au cours de l'année. Cette variable sera complémentaire de la variable G3 qui est la note finale de l'élève. En effet, intuitivement, la note de l'élève au cours d'une année donnée dépend d'un ensemble de facteurs historiques qui ne sont pas nécessairement capturés dans les variables de ce jeu de données. L'évolution de la note permettrait de capturer les progrès ou régressions d'un élève au cours de l'année scolaire étudiée. Cependant cette variable devra être considérée uniquement comme complémentaire de la variable **G3.** En effet, on ne peut pas attendre d'un élève qui commence l'année à plus de 18/20 de faire d'énormes

progrès et symétriquement, un élève qui débute l'année à 0/20 ne peut faire que des progrès.

  - **Pour les besoins de l'ACP, une variable « G3_class» sera également créée**. Elle catégorisera les notes finales en 3 groupes : note basse si elle est inférieure au 1er quartile, note élevée si elle est supérieure au 3ème quartile et note moyenne sinon. Cela permettra de différencier les points sur les nuages de points par des couleurs différentes en fonction du niveau de la note

- **Sélection des observations** : Les nuages de points et une sélection de données montre que 35 élèves ont fini l'année avec la note 0 alors qu'ils ont eu une note supérieure à 5 à au moins une des 2 périodes. Il a été choisi de ne pas conserver ces observations dans les données à traiter.
- Comme indiqué plus haut, sur les 33 variables initiales, seules 12 ont été présélectionnées. La liste complète des variables est présentée en annexe.

Dans un premier temps, ce rapport décrira individuellement les variables ***G1, G2 et G3***, qui représentent les notes obtenues par les élèves au cours de l'année et leur note finale. Il décrira également la variable ***evol*** qui est la différence entre la note finale G3 et la note de la 1ère période G1.

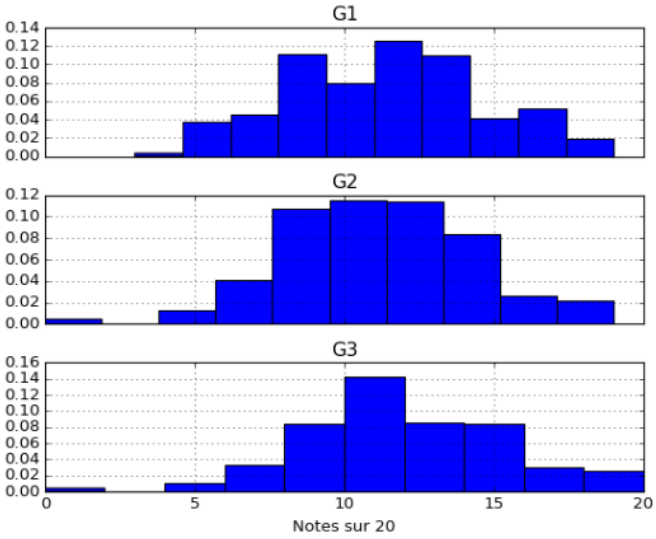
Le lien entre ces variables sera ensuite étudié à l'aide d'outils de statistique bivariée. La corrélation entre ces variables et les autres variables du modèle sera également abordée, en guise d'introduction à la dernière partie.

Ce rapport effectuera enfin une analyse multivariée afin d'analyser si les notes finales ***G3*** peuvent être expliquées par les variables du modèle autres que G1 et G2.

I. Analyse univariée : L'évolution des notes montre une tendance à l'aplatissement

L'objectif de cette partie est de décrire les notes des élèves ainsi que leur évolution. Les variables sélectionnées sont donc G1, G2, G3 et evol.

Comme nous le voyons sur l'histogramme ci-dessous, les notes ont grossièrement l'allure d'une loi normale au début de l'année scolaire (note G1). L'amplitude des notes est déjà grande avec des notes allant de 3 à 19. Mais la majorité des élèves est située entre 9 et 14.



	G1	G2	G3
count	360.000000	360.000000	360.000000
mean	11.213889	11.263889	11.427778
std	3.282659	3.300177	3.381138
min	3.000000	0.000000	0.000000
25%	9.000000	9.000000	9.000000
50%	11.000000	11.000000	11.000000
75%	14.000000	13.250000	14.000000
max	19.000000	19.000000	20.000000

Au fur et à mesure de l'année, la moyenne reste constante à 11/20, de même que la médiane, également à 11/20. La position des différents quartiles change très peu.

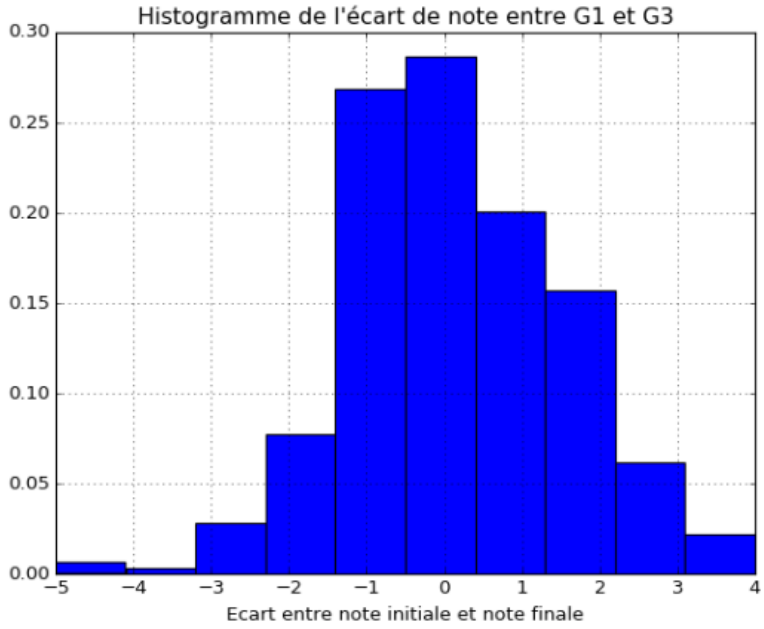
Au fur et à mesure de l'année, les notes s'étalent de plus en plus vers les extrémités : il y a plus de notes très faibles et très élevées.

Le coefficient d'asymétrie est positif à la première période : donc la distribution est plus étalée à droite : il y a quelques très bons qui se détachent du reste. Mais au cours de la seconde et 3ème période, la tendance est inversée comme on le voit sur l'histogramme ci-dessus : de très mauvaises notes se détachent des autres (A noter que cet effet était encore plus prononcé avant que nous n'écartions les observations correspondant aux élèves ayant eu une supérieure à 5 à l'une des 2 périodes et 0 en note finale.).

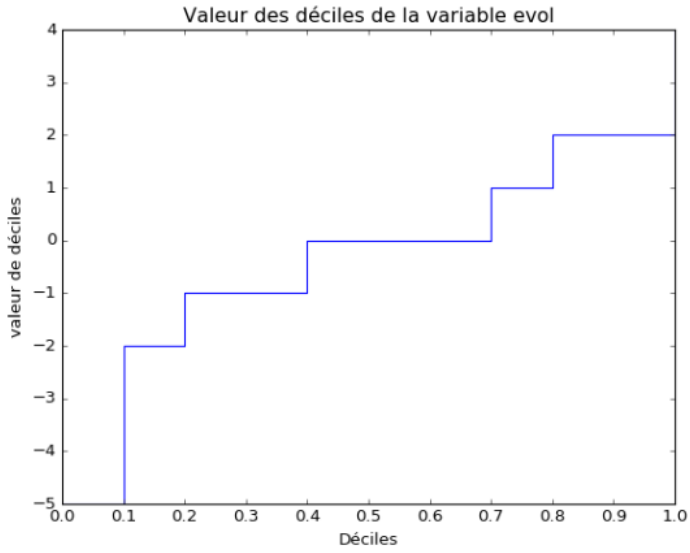
Le coefficient d'aplatissement montre qu'au cours la période 1 (notes G1), la distribution a moins de valeurs très petites et très grandes comparée à une loi normale standard. Les notes sont bien tassées autour de la moyenne. Cependant, à partir de la deuxième période (notes G2 et G3), comme il est possible de le voir sur les histogrammes, la distribution des notes tend à s'aplatir davantage comparée à une distribution normale.

	Skewness	**Kurtosis**
G1	0.137932	-0.660242
G2	-0.090875	0.210885
G3	-0.066184	0.248949

Concernant l'évolution des notes au cours de l'année, il est clair que la majorité des étudiants ont des notes qui évoluent très peu.



Le graphique ci-dessous montre que les déciles 0.1 et 0.9 valent respectivement -2 et 2. La variable « evol » est approximativement centrée avec une moyenne de 0,2. Un grand nombre d'élèves voit leur moyenne légèrement en baisse (de 1 point), mais cela est contrebalancé quelques élèves qui ont leur moyenne qui augmente un peu plus significativement (de 2 à 3 points sur leur note).

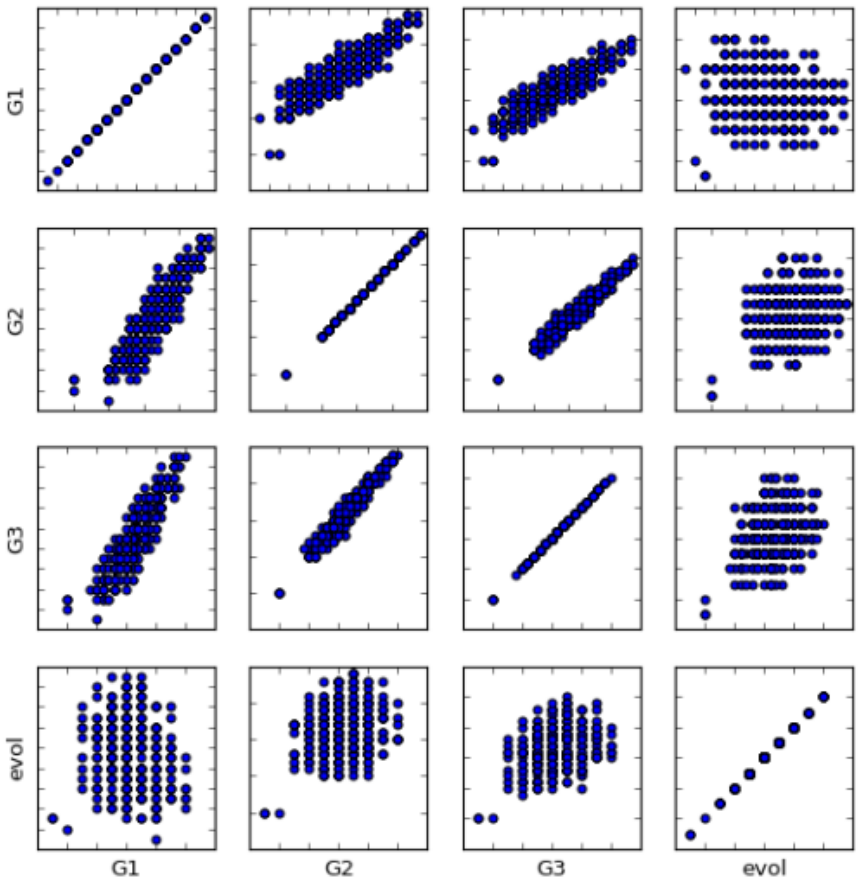


En conclusion de cette partie, au fur et à mesure de l'année, les notes évoluent peu. Cependant, cette légère évolution se fait vers les valeurs extrêmes : l'écart se creuse. Il y a de plus en plus de très bonnes notes et de très mauvaises notes, même après l'ajustement des données effectué en introduction.

II. Analyse bivariée : Les notes obtenues en période 2 sont déterminantes pour la note finale

Les nuages de points effectués entre les 3 variables G1, G2 et G3 montrent que les notes G2 et G3 sont davantage corrélées entre elles que G1 et G2 ou G1 et G3.

L'évolution l'écart de la note entre G1 et G3 (variable 'evol') semble est peu corrélée avec les notes obtenues.



Cela tout cela est confirmé par la matrice de corrélation ci-dessous :

	G1	**G2**	**G3**	**evol**
G1	1.000000	0.899337	0.890196	-0.174377
G2	0.899337	1.000000	0.968923	0.207017
G3	0.890196	0.968923	1.000000	0.293369
evol	-0.174377	0.207017	0.293369	1.000000

Une matrice de corrélation complète entre les 12 variables du modèle a été réalisée (en annexes). Mis à part les corrélations entre les notes des différents semestres, les variables du modèle sont relativement peu corrélées. La corrélation la plus significative rencontrée en-dehors de celle entre les notes et celle entre le niveau d'éducation de la mère et celui du père.

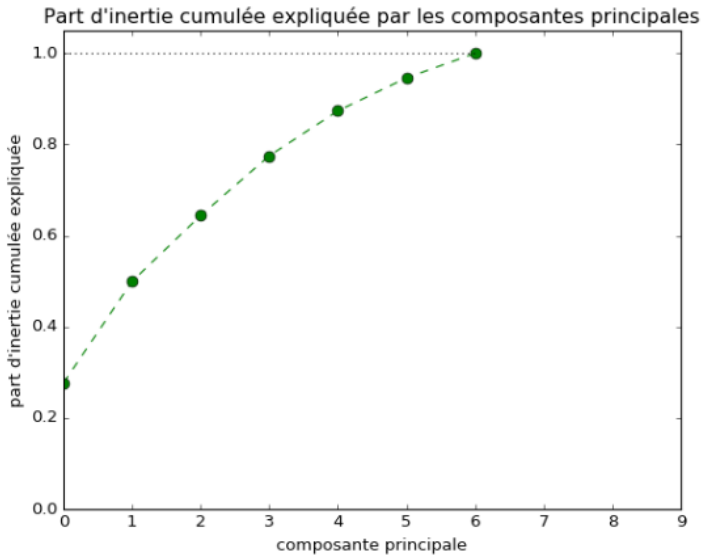
III. Analyse multivariée : la note finale G3 est difficilement prévisible dans son intégralité. Cependant il plus aisé d'expliquer séparément les caractéristiques des différents niveaux de notes

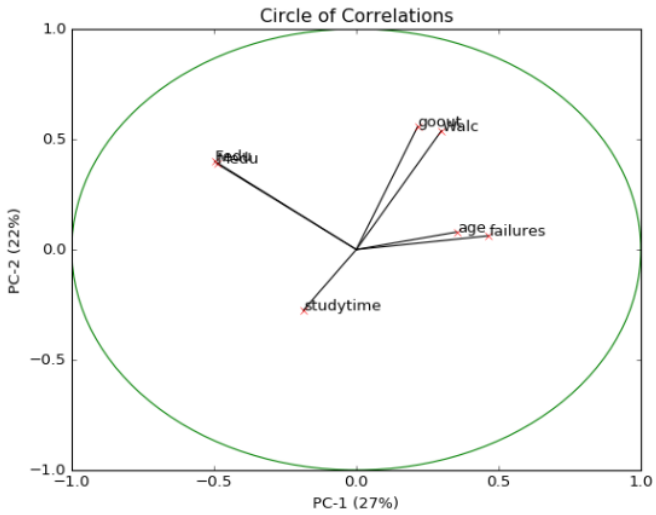
Pour l'analyse multivariée, la variable « famrel » sera écartée car elle est particulièrement faiblement corrélée avec les notes. La variable « evol », que nous avions créée uniquement afin de décrire l'évolution des notes, sera également écartée de l'étude multivariée.

La variable 'G3_class' a été introduite comme présenté dans l'introduction de ce rapport.

Cette ACP est réalisée en écartant les variables G1, G2 et G3 car nous savons que G1 et G2 sont fortement corrélées à G3. L'objectif de cette ACP sera d'étudier si la variable G3 peut bien être expliquée par la variabilité d'autres variables du modèle.

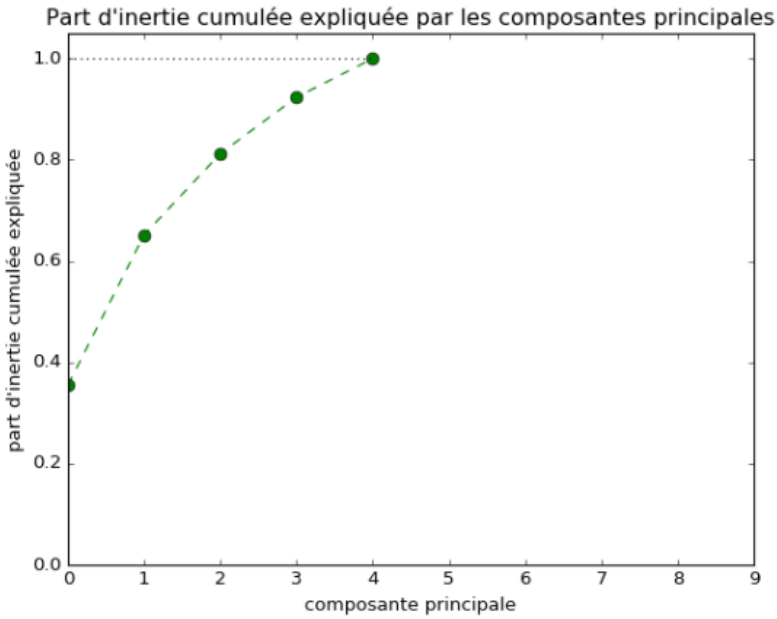
Comme le montre le graphique ci-dessous qui représente la part d'inertie expliquée de manière cumulée par les composantes principales, il faut retenir au moins les 4 premières composantes principales pour s'approcher des 80% de variance expliquée. Cela n'est pas satisfaisant pour les représentations. En ne retenant que 2 variables, 50% de l'inertie totale est expliquée (et environ 65% en en retenant 3).





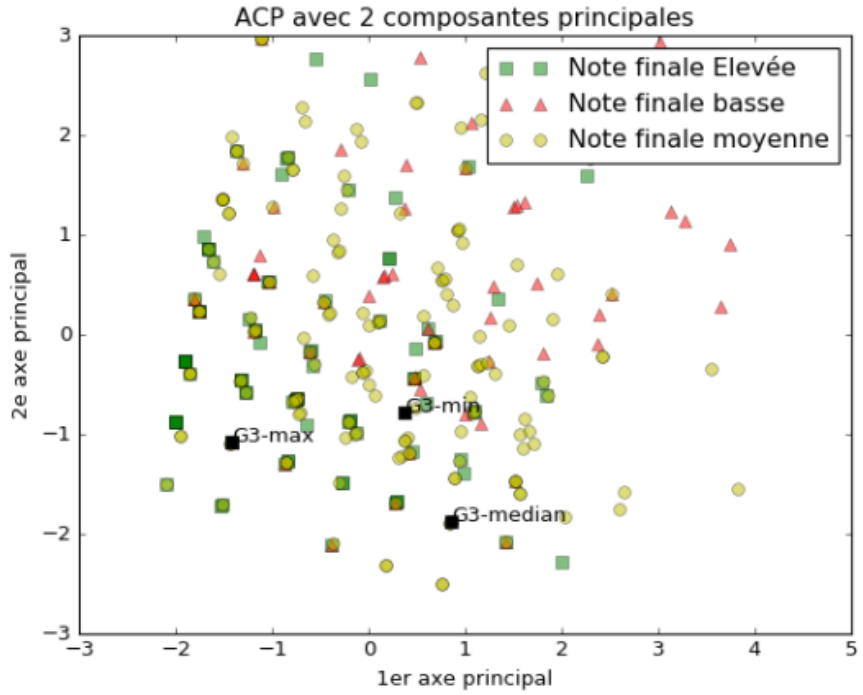
Le cercle de corrélation indique que les variables les moins bien représentées par les deux premières composantes principales sont ‘studytime’ et ‘age’. Afin d’obtenir un modèle mieux représentable, ces 2 variables sont donc écartées de l’ACP.

Sans ces variables, le schéma ci-dessous montre qu’avec les 2 premières CP, 65% de l’inertie est expliquée et qu’avec les 3 premières, 80% de l’inertie est expliquée.

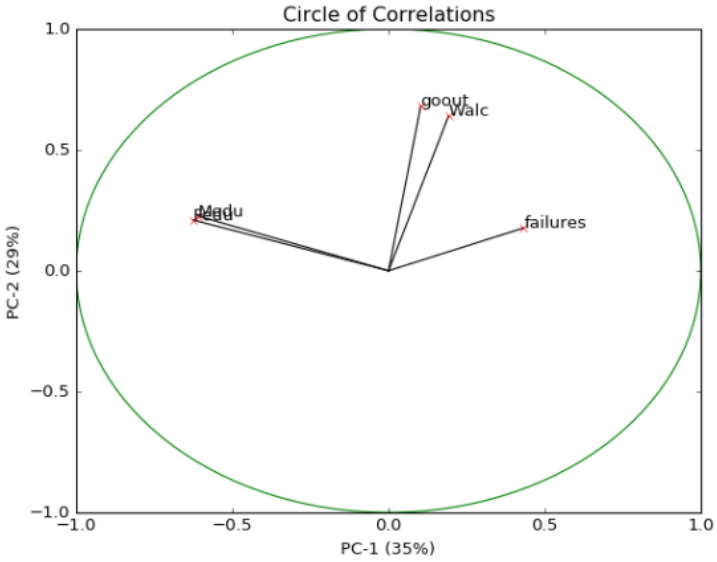


La représentation du nuage de points projeté sur les 2 premières composantes principales montre qu’il est assez difficile d’expliquer les notes G3 à partir des variations des 2 premières composantes principales. Cependant les notes G3

élevées sont plutôt regroupées du côté gauche du graphique tandis que les notes faibles ont tendance à être dans la partie supérieure du graphique.

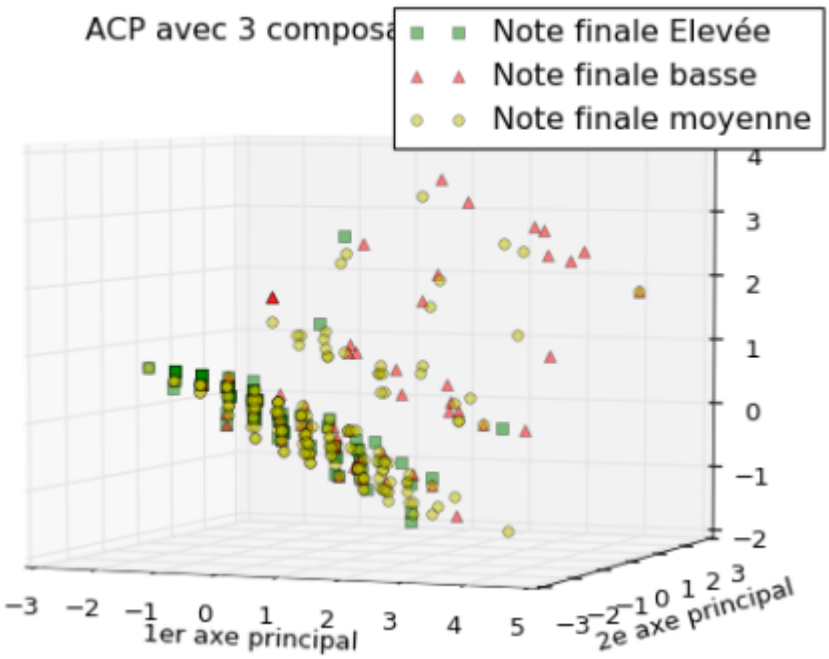


En analysant le cercle des corrélations, ce sont les variables ‘Medu’, ‘Fedu’ et ‘failures’, qui sont le mieux représentées par la 1ère composante principale (CP). ‘failure’ évolue dans le même sens que cette composante principale. La remarque du paragraphe précédent indique donc que les notes élevées ont tendance à être inversement corrélées avec le nombre de redoublements. Elles sont plus fréquentes du côté des élèves ayant des parents avec des niveaux d’éducation élevés.

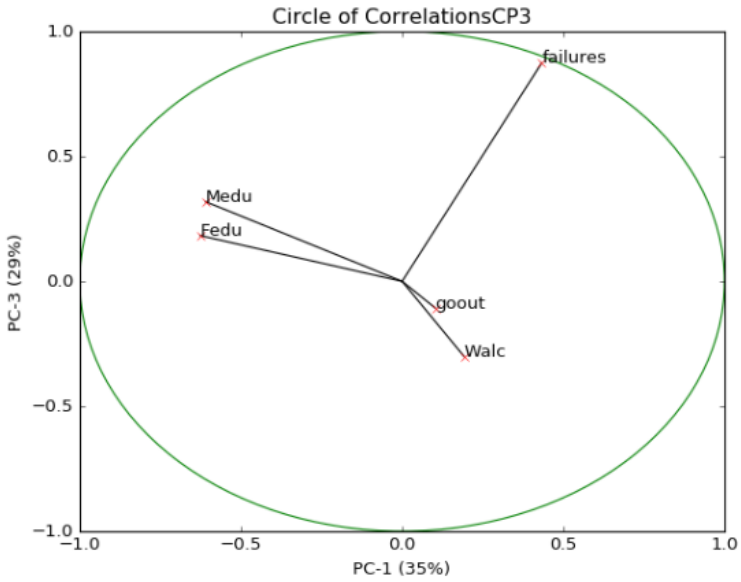


De même, le cercle des corrélations montre que les variables les mieux représentées par la deuxième CP sont ‘goout’ and ‘Walc’. Ainsi, les notes négatives ont tendance à être regroupées du côté des étudiants qui sortent fréquemment et consomment de l’alcool durant le week-end.

Le nuage de points projeté sur les 3 premières CP permet de mieux séparer encore mieux les observations de la note G3 :



- les mauvaises notes sont plutôt du côté des valeurs élevées de la 1ère et 2ème CP (ce qui correspond à des valeurs plutôt élevés pour les variables ‘goout’, ‘Walc’ et ‘failures’ comme l’indique le cercle des corrélations).
- Les bonnes/moyennes notes sont plutôt du côté des valeurs basses pour la 3ème et 1ère CP. Dans le cercle des corrélations avec la 1ère et 3ème CP, nous voyons qu’elles moins représentée quand le nombre de redoublements ‘failures’ augmente. D’ailleurs la variable ‘failure’ est très bien représentée par ces composantes principales. Puis nous pouvons encore relever, comme indiqué plus haut, l’effet de la 1ère CP, qui représente bien les variables Medu et Fedu.



En conclusion, ce rapport a fourni une description des notes des élèves et de leur évolution. Il est montré que les notes sont bien regroupées autour de la moyenne de 11/20 dès le début de l'année, mais qu'au fur et à mesure, les bonnes notes et mauvaises notes se font plus fréquentes : la distribution des notes a tendance à s'aplatir.

La note finale des élèves est fortement corrélée à chacune des notes obtenues durant la 1ère période et durant la 2nde période. Cependant la corrélation est plus forte avec la note de la seconde période. Les corrélations sont très faibles avec les autres variables du modèle.

Cependant, l'analyse multivariée a permis de repérer certaines tendances :

- Les élèves ayant de bonnes notes sont davantage représentés parmi ceux qui ont le moins redoublé auparavant et qui ont des parents avec des niveaux d'éducation élevés.
- Les élèves ayant des notes faibles sont les plus fréquents parmi les élèves qui sortent fréquemment, consomment de l'alcool le week-end et ont redoublé plusieurs fois par le passé.
- Pour les notes moyennes, leur profil est moins tranché, cependant il est davantage similaire à celui des élèves ayant de bonnes notes.

Les limitations de ces résultats sont qu'il n'a pas été possible d'expliquer les notes dans leur globalité. Nous n'avons pu que caractériser les groupes de notes de manière individuelle. En effet, pour ce qui est par exemple de la 1ère composante

principale de l'ACP, celle qui explique le plus de variabilité, nous avons vu certes que les bonnes notes étaient surtout représentées parmi les valeurs faibles de cette CP, cependant les mauvaises notes sont décorrélées de cette CP : on en rencontre aussi bien dans les valeurs faibles que dans les valeurs élevées.

Par ailleurs, les conclusions sur l'ACP effectuée n'ont pas été concrètement chiffrées. Elles reposent surtout sur une analyse graphique. Par exemple les niveaux de corrélation entre la 1ère CP et le fait qu'un élève a une bonne note n'a pas été quantifiée. Pour aller plus loin, il faudrait par exemple créer une variable qui vaut 1 si la note fait partie des bonnes notes et 0 sinon, et analyser les corrélations entre cette nouvelle variable et la 1ère composante principale. Une analyse complémentaire consisterait également à effectuer une régression multilinéaire sur les 3 premières composantes principales afin de tester si le modèle obtenu a un niveau d'erreur élevé ou pas.

ANNEXES

Liste complète des variables initiales

1. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. sex - student's sex (binary: 'F' - female or 'M' - male)
3. age - student's age (numeric: from 15 to 22)
4. address - student's home address type (binary: 'U' - urban or 'R' - rural)
5. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
9. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11. reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12. guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13. traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

14. studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16. schoolsup - extra educational support (binary: yes or no)
17. famsup - family educational support (binary: yes or no)
18. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. activities - extra-curricular activities (binary: yes or no)
20. nursery - attended nursery school (binary: yes or no)
21. higher - wants to take higher education (binary: yes or no)
22. internet - Internet access at home (binary: yes or no)
23. romantic - with a romantic relationship (binary: yes or no)
24. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26. goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. health - current health status (numeric: from 1 - very bad to 5 - very good)
30. absences - number of school absences (numeric: from 0 to 93)

Matrice de corrélations complète

	evol	G1	G2	G3	Medu	Fedu	studytime	famrel	goout	Walc	log_absences	age	failures
evol	1.000000	-0.174377	0.207017	0.293369	0.038913	0.005410	0.003061	0.044747	-0.069886	-0.006623	-0.141098	-0.239700	-0.071889
G1	-0.174377	1.000000	0.899337	0.890196	0.171081	0.167058	0.158179	0.003081	-0.154370	-0.159948	-0.127211	-0.038034	-0.337620
G2	0.207017	0.899337	1.000000	0.968923	0.196261	0.183505	0.148634	-0.017231	-0.161154	-0.145265	-0.191150	-0.164747	-0.368458
G3	0.293369	0.890196	0.968923	1.000000	0.184102	0.164695	0.154988	0.023694	-0.182208	-0.158354	-0.188787	-0.147827	-0.361047
Medu	0.038913	0.171081	0.196261	0.184102	1.000000	0.608582	0.056580	-0.008913	0.073846	-0.049685	0.049493	-0.139002	-0.208096
Fedu	0.005410	0.167058	0.183505	0.164695	0.608582	1.000000	-0.023126	-0.014183	0.032822	-0.017559	-0.029563	-0.141591	-0.265814
studytime	0.003061	0.158179	0.148634	0.154988	0.056580	-0.023126	1.000000	0.047252	-0.052652	-0.237266	-0.047882	-0.005087	-0.158765
famrel	0.044747	0.003081	-0.017231	0.023694	-0.008913	-0.014183	0.047252	1.000000	0.036154	-0.127123	-0.135509	0.066825	0.006148
goout	-0.069886	-0.154370	-0.161154	-0.182208	0.073846	0.032822	-0.052652	0.036154	1.000000	0.439367	0.153063	0.133517	0.141704
Walc	-0.006623	-0.159948	-0.145265	-0.158354	-0.049685	-0.017559	-0.237266	-0.127123	0.439367	1.000000	0.195087	0.115873	0.144101
log_absences	-0.141098	-0.127211	-0.191150	-0.188787	0.049493	-0.029563	-0.047882	-0.135509	0.153063	0.195087	1.000000	0.211821	0.132562
age	-0.239700	-0.038034	-0.164747	-0.147827	-0.139002	-0.141591	-0.005087	0.066825	0.133517	0.115873	0.211821	1.000000	0.275724
failures	-0.071889	-0.337620	-0.368458	-0.361047	-0.208096	-0.265814	-0.158765	0.006148	0.141704	0.144101	0.132562	0.275724	1.000000