

SUP DE VINCI - INGENIERIE DES SYSTEMES
D'INFORMATION

LES FACTEURS DE REUSSITE SCOLAIRE DES ETUDIANTS



Mamakan COULIBALY, Estelle NTSAMA

30 octobre 2022

Table des matières

1	Introduction	2
2	Librairies	2
3	Importation des données	2
4	Description du dataset	3
4.1	Aperçu des données	3
4.2	Données manquantes	3
4.3	Description statistique des données quantitatives	5
5	Analyse des données	6
5.1	Première modélisation : arbre de décision	6
5.2	Deuxième modélisation : matrice de corrélation	8
5.3	Statistiques	9
5.3.1	Alcool	9
5.3.2	Résultat	10
5.3.3	Etablissement	11
5.3.4	Education des parents	13
5.3.5	Temps de transport	15
5.3.6	Géographie	16
5.3.7	Couple	16
5.3.8	Age et assiduité	18
6	Conclusion	19
7	Source	19

1 Introduction

Notre étude porte sur les performances scolaires des étudiants. Pour ce faire, nous avons utilisé une base de données basées sur les caractéristiques d'élèves dans deux écoles portugaises. Nous disposons d'attributs tels que le sexe, l'âge, la consommation d'alcool, l'environnement familial, le statut relationnel, la zone d'habitation, les notes etc. Compte tenu de l'ensemble de données, nous avons pour objectif de relever les facteurs mélioratifs ou péjoratifs sur les résultats des élèves.

2 Librairies

Nous avons choisi l'éditeur du site Kaggle comme environnement de développement et comme langage de programmation Python. Ci-dessous les librairies que nous avons utilisées :

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
from matplotlib import pyplot as plt #graph production
import plotly.express as px #makes interactive, publication-quality graphs.
import seaborn as sns #graph production
from sklearn.neighbors import KNeighborsClassifier #for K-NN method
from sklearn.model_selection import cross_val_score #Evaluate a score by cross-validation
from sklearn.model_selection import train_test_split #returns a list of train-test splits of the inputs
from sklearn.ensemble import RandomForestClassifier #for decision trees method
from sklearn.metrics import confusion_matrix #to evaluate the accuracy of a classification.
```

FIGURE 1 – Les librairies utilisées.

3 Importation des données

Les données que nous avons utilisées proviennent de Kaggle, plus précisément du dossier alcohol-effects-on-study dans lequel nous avons téléchargé le fichier Maths.csv.

```
#importation du document
pth = '/kaggle/input/alcohol-effects-on-study/'
df = pd.read_csv (pth + 'Maths.csv')
```

FIGURE 2 – Importation du document.

4 Description du dataset

4.1 Aperçu des données

La première étape de notre étude a été d'afficher un premier aperçu de la représentation des données dans notre base, grâce à la fonction `head()`. Nous avons ainsi pris connaissance des 5 premières lignes de notre base et avons eu comme information qu'elle se compose de 33 colonnes.



```
df.head()
```

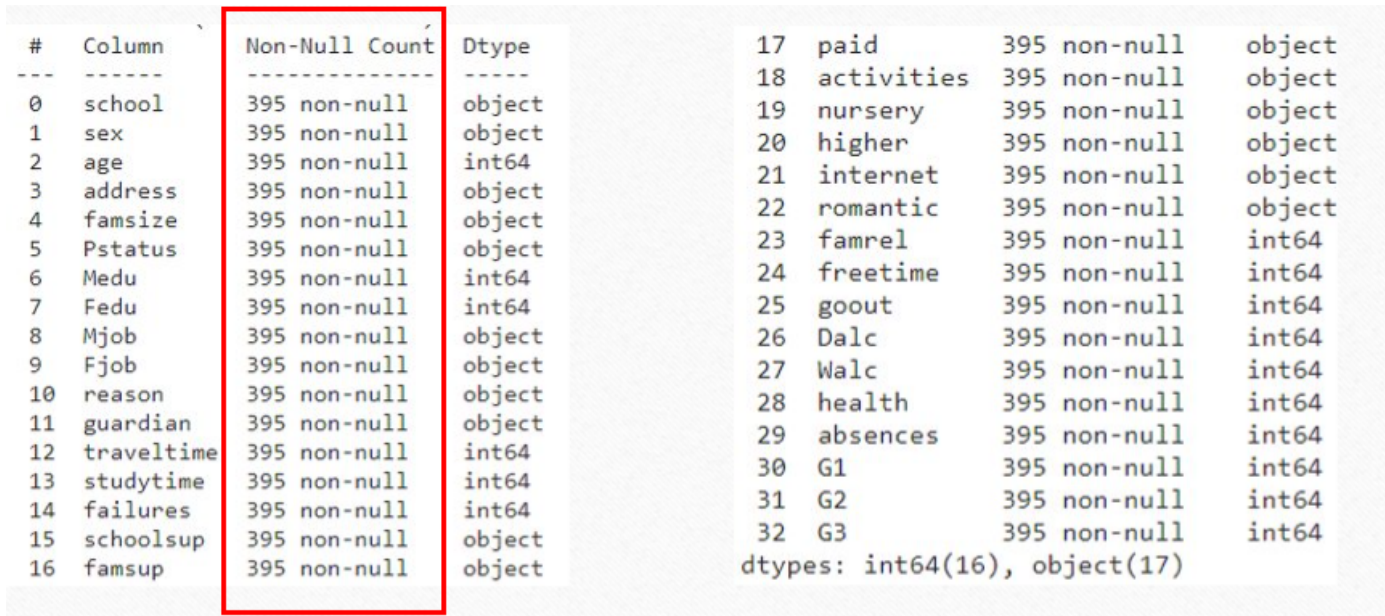
	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	6	5	6	6
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	4	5	5	6
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	10	7	8	10
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	2	15	14	15
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	4	6	10	10

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')

FIGURE 3 – Entête de la base de données.

4.2 Données manquantes

La fonction `info()` nous a permis de connaître le nombre et le type de donnée auxquelles on avait à faire. On constate que l'on dispose de 395 enregistrements et aucune valeur non nulle.



#	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)

FIGURE 4 – Nombre et type des données.

Après cette première analyse, nous n'avions donc pas à faire à des données manquantes, ce qui est une bonne chose dans le cadre d'une étude de données car celles-ci ont tendance à fausser les résultats obtenus.

Columns	Description
school	student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
sex	student's sex (binary: 'F' - female or 'M' - male)
age	student's age (numeric: from 15 to 22)
address	student's home address type (binary: 'U' - urban or 'R' - rural)
famsize	family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Pstatus	parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
Fedu	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
Mjob	mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Fjob	father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
reason	reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
guardian	student's guardian (nominal: 'mother', 'father' or 'other')
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)
studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures	number of past class failures (numeric: n if 1<=n<3, else 4)
schoolsup	extra educational support (binary: yes or no)
famsup	family educational support (binary: yes or no)
paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
internet	Internet access at home (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	free time after school (numeric: from 1 - very low to 5 - very high)
goout	going out with friends (numeric: from 1 - very low to 5 - very high)
Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	current health status (numeric: from 1 - very bad to 5 - very good)
absences	number of school absences (numeric: from 0 to 93)

FIGURE 5 – Description des données.

La dataset maths contient des données quantitatives et qualitatives. D'une part, nous avons des renseignements sur l'identité de l'élève : son école, son sexe, son âge, son adresse, sa situation amoureuse... et plus précisément sur son mode de vie ce qui inclut ses sorties, son temps libre, ses activités et sa consommation d'alcool que l'on peut assimiler à sa santé en plus ses colonnes infirmerie et santé. Nous avons également des informations sur son environnement familial dont le niveau d'éducation des parents, le métier des parents, la situation des parents (vivant ensemble ou séparément), la garde l'enfant, le soutien de la part de sa famille concernant les études, la qualité des liens familiaux, la taille de la famille. Il existe également trois colonnes G1, G2 et G3 représentant des notes en mathématiques allant de 0 à 20. Au niveau de la scolarité, nous avons d'autres données comme les absences, le temps de révision etc. La base de données contient énormément d'informations. Nous nous focaliserons donc sur les données que nous jugeons plus pertinentes sans pour autant supprimer les autres colonnes que nous allons garder sous la main.

4.3 Description statistique des données quantitatives

A l'aide de la fonction `describe()`, nous avons calculé pour chaque attribut des éléments tels que :

- Le nombre total d'enregistrements
- La moyenne
- L'écart-type
- Le minimum
- Le premier quartile
- La médiane
- Le troisième quartile
- Le maximum

	count	mean	std	min	25%	50%	75%	max
age	395.000000	16.696203	1.276043	15.000000	16.000000	17.000000	18.000000	22.000000
Medu	395.000000	2.749367	1.094735	0.000000	2.000000	3.000000	4.000000	4.000000
Fedu	395.000000	2.521519	1.088201	0.000000	2.000000	2.000000	3.000000	4.000000
traveltime	395.000000	1.448101	0.697505	1.000000	1.000000	1.000000	2.000000	4.000000
studytime	395.000000	2.035443	0.839240	1.000000	1.000000	2.000000	2.000000	4.000000
failures	395.000000	0.334177	0.743651	0.000000	0.000000	0.000000	0.000000	3.000000
famrel	395.000000	3.944304	0.896659	1.000000	4.000000	4.000000	5.000000	5.000000
freetime	395.000000	3.235443	0.998862	1.000000	3.000000	3.000000	4.000000	5.000000
goout	395.000000	3.108861	1.113278	1.000000	2.000000	3.000000	4.000000	5.000000
Dalc	395.000000	1.481013	0.890741	1.000000	1.000000	1.000000	2.000000	5.000000
Walc	395.000000	2.291139	1.287897	1.000000	1.000000	2.000000	3.000000	5.000000
goout	395.000000	3.108861	1.113278	1.000000	2.000000	3.000000	4.000000	5.000000
Dalc	395.000000	1.481013	0.890741	1.000000	1.000000	1.000000	2.000000	5.000000
Walc	395.000000	2.291139	1.287897	1.000000	1.000000	2.000000	3.000000	5.000000
health	395.000000	3.554430	1.390303	1.000000	3.000000	4.000000	5.000000	5.000000
absences	395.000000	5.708861	8.003096	0.000000	0.000000	4.000000	8.000000	75.000000
G1	395.000000	10.908861	3.319195	3.000000	8.000000	11.000000	13.000000	19.000000
G2	395.000000	10.713924	3.761505	0.000000	9.000000	11.000000	13.000000	19.000000
G3	395.000000	10.415190	4.581443	0.000000	8.000000	11.000000	14.000000	20.000000

FIGURE 6 – Description des données.

5 Analyse des données

5.1 Première modélisation : arbre de décision

Nous nous sommes intéressées à la question de savoir si le nombre de sorties avait-il une influence sur le taux de consommation d'alcool et quelle tranche de la population était touchée par ce phénomène. L'idée était donc de modéliser la solution du problème de machine learning que l'on a traité comme une suite de décisions à prendre. Une décision étant représentée par une feuille dans l'arbre, avons donc choisi comme attribut de sortie `goout` : le taux de sortie avec des amis (numérique : de 1 - très faible à 5 - très élevé)

```
label = df['goout'] #critère de recherche
label
```

FIGURE 7 – Label.

Ensuite nous avons choisis 5 attributs d'entrée pour notre étude :

```
data.head()
```

	school	sex	age	Dalc	Walc
338	GP	F	18	1	1
210	GP	F	19	1	2
121	GP	M	15	1	2
282	GP	F	18	1	1
67	GP	F	16	1	2

FIGURE 8 – Echantillon d'étude

Puis nous avons remplacé tous les attributs qui étaient de type String en type int pour pouvoir les manier correctement :

data['school']		data['sex']	
338	1	338	0
210	1	210	0
121	1	121	0
282	1	282	0
67	1	67	0
..
194	1	194	0
196	1	196	0
198	1	198	0
203	1	203	0
394	0	394	0
..

FIGURE 9 – Conversion str to int.

Et enfin, grâce à la librairie `DecisionTreeClassifier`, nous avons obtenu l'arbre de décision contenant les

résultats de prédiction relatives à la question posée. ce graphe il ressort que les jeunes de moins de 17 ans sont ceux qui sortent le moins et donc qui consomment le moins. Plus le nombre de sortie augmente, plus la consommation d'alcool augmente également, ce qui influence négativement les résultats des étudiants.

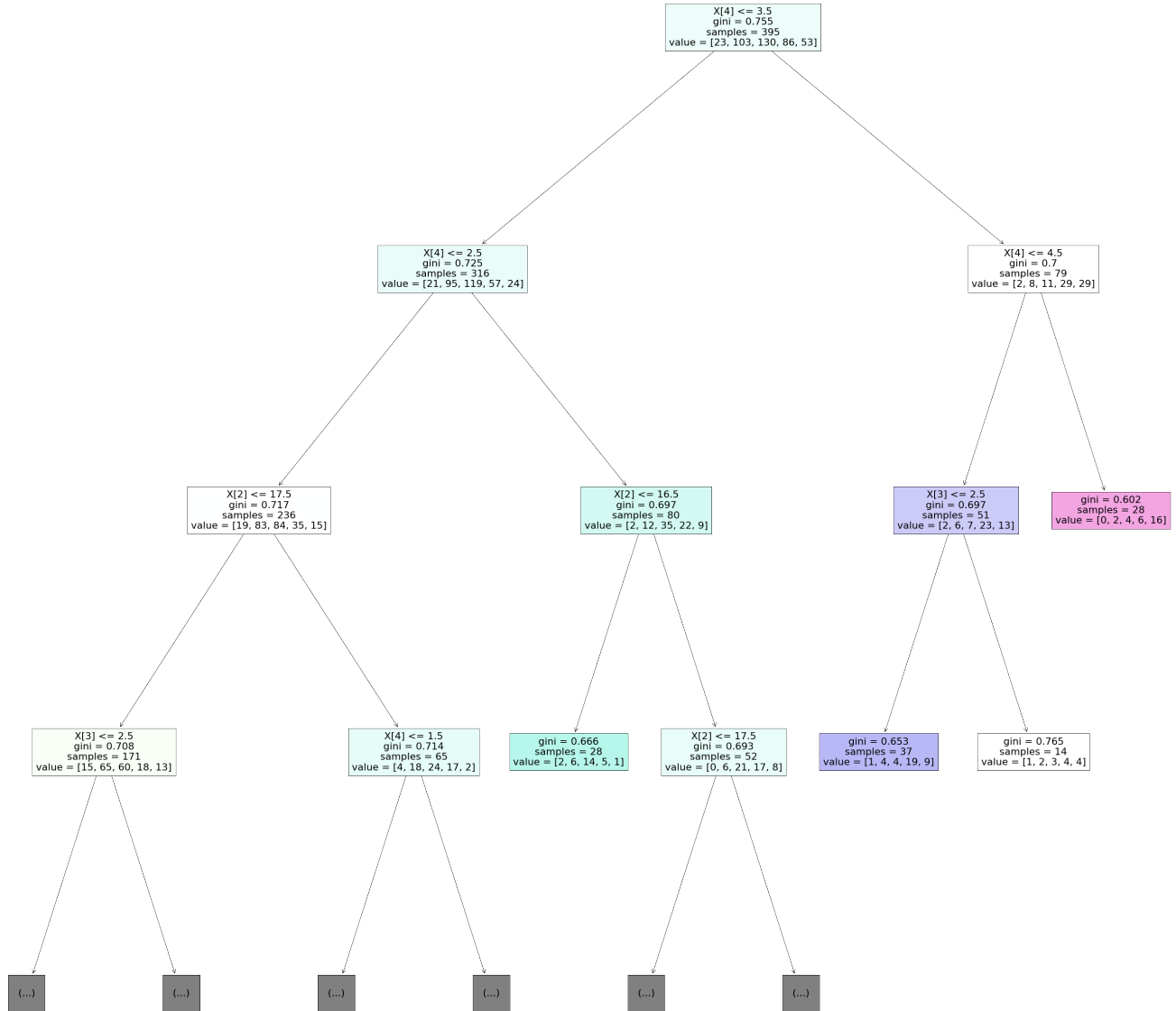


FIGURE 10 – Arbre de décision.

5.2 Deuxième modélisation : matrice de corrélation

```
import pandas as pd
import numpy as np
from matplotlib import pyplot
df.corr(method='spearman').style.format("{:.2}").background_gradient(cmap=pyplot.get_cmap('coolwarm'))
```

	age	Medu	Fedu	travelttime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	Grade
age	1.0	-0.16	-0.15	0.11	0.032	0.24	0.031	0.0003	0.14	0.097	0.13	-0.075	0.15	-0.14
Medu	-0.16	1.0	0.63	-0.15	0.063	-0.24	0.012	0.028	0.065	0.023	-0.044	-0.036	0.098	0.23
Fedu	-0.15	0.63	1.0	-0.15	0.018	-0.24	0.011	-0.017	0.048	0.004	-0.014	0.018	0.0036	0.18
travelttime	0.11	-0.15	-0.15	1.0	-0.11	0.08	-0.039	-0.022	-0.0014	0.066	0.064	-0.015	-0.025	-0.12
studytime	0.032	0.063	0.018	-0.11	1.0	-0.16	0.058	-0.13	-0.066	-0.22	-0.26	-0.091	-0.046	0.12
failures	0.24	-0.24	-0.24	0.08	-0.16	1.0	-0.051	0.088	0.11	0.19	0.13	0.08	0.096	-0.37
famrel	0.031	0.012	0.011	-0.039	0.058	-0.051	1.0	0.14	0.064	-0.11	-0.12	0.085	-0.087	0.032
freetime	0.0003	0.028	-0.017	-0.022	-0.13	0.088	0.14	1.0	0.29	0.19	0.13	0.089	0.013	0.00068
goout	0.14	0.065	0.048	-0.0014	-0.066	0.11	0.064	0.29	1.0	0.26	0.39	-0.019	0.13	-0.16
Dalc	0.097	0.023	0.004	0.066	-0.22	0.19	-0.11	0.19	0.26	1.0	0.64	0.095	0.13	-0.11
Walc	0.13	-0.044	-0.014	0.064	-0.26	0.13	-0.12	0.13	0.39	0.64	1.0	0.094	0.21	-0.097
health	-0.075	-0.036	0.018	-0.015	-0.091	0.08	0.085	0.089	-0.019	0.095	0.094	1.0	-0.07	-0.048
absences	0.15	0.098	0.0036	-0.025	-0.046	0.096	-0.087	0.013	0.13	0.13	0.21	-0.07	1.0	0.026
Grade	-0.14	0.23	0.18	-0.12	0.12	-0.37	0.032	0.00068	-0.16	-0.11	-0.097	-0.048	0.026	1.0

FIGURE 11 – Matrice de corrélation

La matrice de corrélation de Spearman nous permet d'évaluer les relations entre deux variables quantitatives. Le coefficient de corrélation représente cette relation. Nous nous intéressons à l'importance et au signe de celui-ci.

Tout d'abord, nous avons regroupé les trois colonnes G1, G2 et G3 de notre dataset en une unique colonne représentant la moyenne des 3 que l'on a nommé « Grade ». Les coefficients de corrélation reliés avec la variable « Grade » ne sont pas très élevés. Nous n'avons donc pas de forte corrélation mais nous allons tout de même tenter d'identifier les relations. La variable « Grade » est négativement corrélée avec la variable « age » ce qui signifie que plus un étudiant est âgé, plus ses notes auront tendance à diminuer. Le coefficient de corrélation est de 0.23 entre « Medu » et « Grade » et de 0.18 entre « Fedu » et « Grade ». Cela signifie que plus le niveau d'éducation des parents augmente, plus les résultats de l'étudiant augmentent. Ensuite, plus un étudiant fait des sorties, plus ses notes baissent et pareillement pour la consommation d'alcool. Au-delà de la variable « Grade », on peut relever d'autres relations dans cette matrice dont celles coloré en rouge clair. Les variables « Medu » et « Fedu » sont fortement corrélées avec un coefficient de corrélation de 0.63. Plus le niveau d'études de la mère est élevé, plus le niveau d'études du père est généralement élevé. D'autre part, les variables « Walc » et « Grade » sont également fortement corrélées. Leur coefficient de corrélation est de 0.64. Plus la consommation d'alcool en semaine est élevée, plus la consommation d'alcool le week-end est généralement élevée. Pour finir, nous pouvons faire la même observation que sur l'arbre de décision : plus un étudiant a tendance à sortir, plus sa consommation d'alcool est élevée et plus ses notes diminuent.

5.3 Statistiques

5.3.1 Alcool

Nous nous sommes également intéressées au taux de consommation d'alcool par jour en fonction du sexe. Ici on constate que les filles consomment moins d'alcool que les garçons.

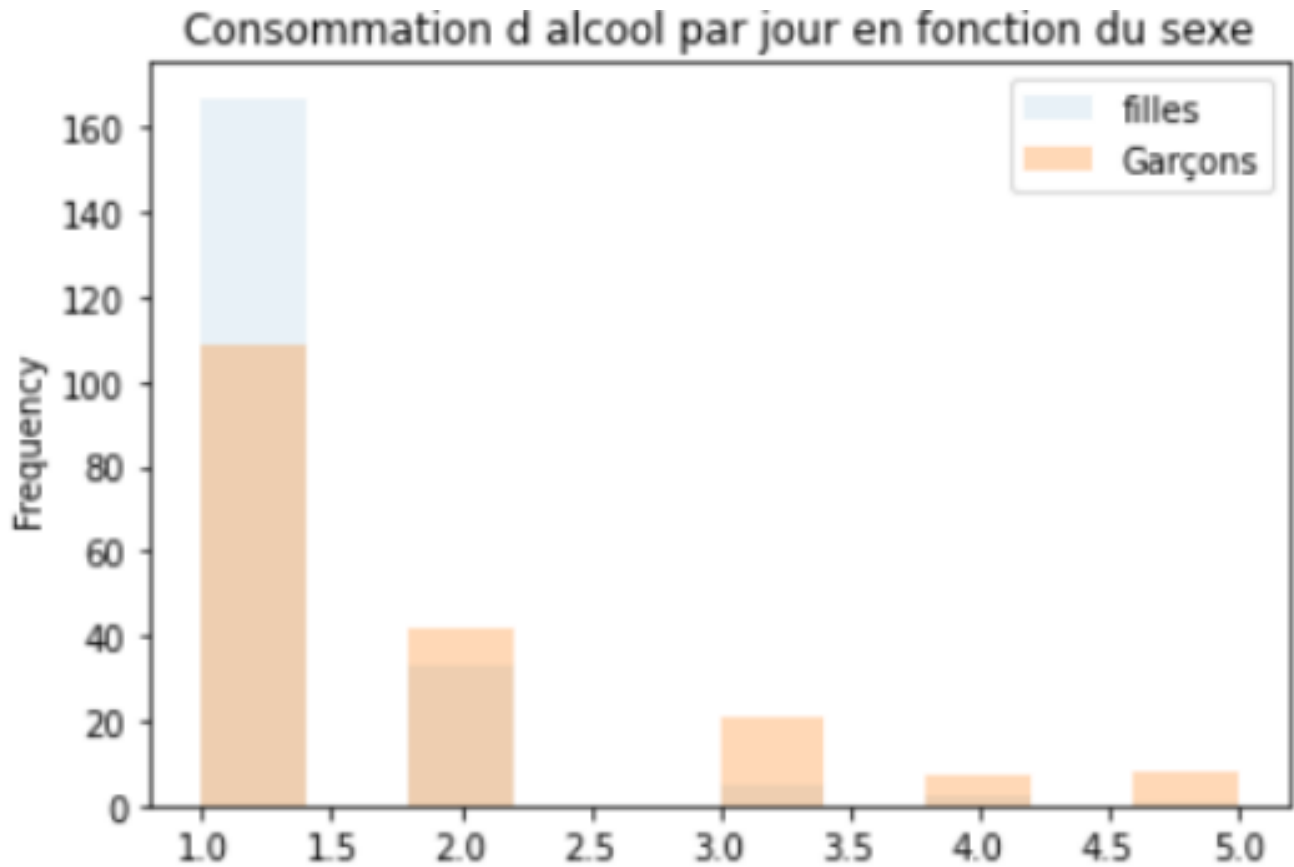


FIGURE 12 – Alcool et sexe.

5.3.2 Résultat

Dans le cadre de notre étude, nous porterons une attention particulière sur les notes des élèves. Ainsi, nous avons regroupé les trois notes correspondant aux colonnes G1, G2 et G3 en une seule colonne « Grade » représentant la moyenne de ces trois notes. Le diagramme en boîte suivant nous permet de visualiser la distribution de notre variable quantitative.

```
import matplotlib.pyplot as plt

plt.boxplot(df.Grade) ; plt.title("Boîte à moustache") ;

plt.gca().xaxis.set_ticklabels(['Note']) ; plt.show()
```

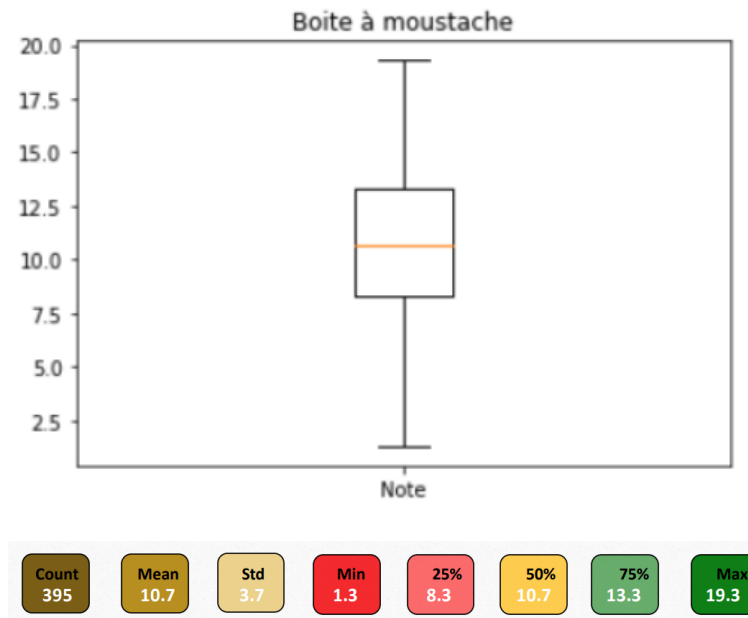


FIGURE 13 – Boîte à moustache des notes

D'après le graphique précédent, la moyenne de note en mathématiques est 10,7. La note minimale est de 1.3 tandis que la note maximale est de 19.3. Parmi les 395 étudiants, 1/4 des élèves ont des notes inférieures ou égales à 8.3, 1/2 ont des notes inférieures ou égales 10.7 et 3/4 ont des notes inférieures ou égales à 13.3. L'écart-type étant de 3.7, les notes sont assez dispersées autour de la moyenne.

Afin d'avoir une vision plus globale des notes, nous avons regroupé les notes en 5 catégories (que nous appellerons également mention) dans une nouvelle colonne « GradeTEST ». Pour ce faire, nous avons arrondi les notes de la colonne « Grade » à l'unité. La catégorie « very_cgood » inclut les notes supérieures ou égales à 16 ensuite nous avons « good » pour les notes de 14 et de 15 puis « quite_well » pour les notes allant de 10 à 13, « average » pour les notes de 7 à 9 et enfin « poor » pour les notes inférieures ou égales à 6.

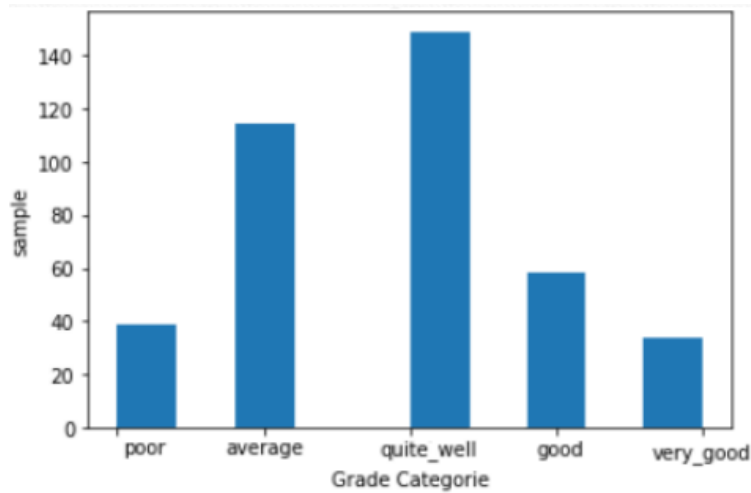


FIGURE 14 – Catégorie des notes

Ce diagramme illustre la répartition des élèves selon la catégorie correspondant à leur note. La catégorie « quite _ well » (assez bien) arrive en tête avec plus de 140 étudiants suivis par la catégorie « average » qui compte plus de 110 étudiants. On compte plus d'élèves avec des notes « good » plutôt que « poor » toutefois on retrouve moins d'étudiants avec des notes catégorisées « very _ good ».

5.3.3 Etablissement

Nous nous demandons si le choix de l'établissement impacte sur les résultats d'un étudiant. Le jeu de données que nous avons sélectionné contient des données d'étudiants de deux écoles distinctes. Afin de répondre à notre problématique, nous allons évaluer le niveau des élèves en mathématiques selon leur établissement.

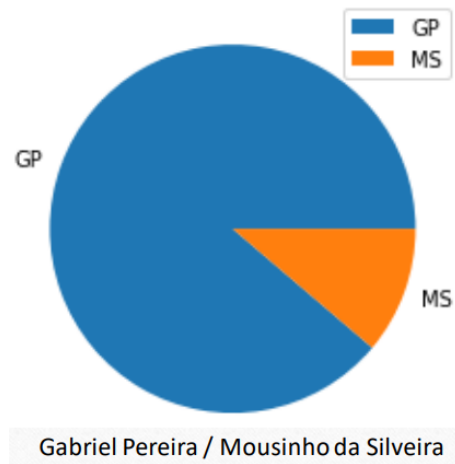


FIGURE 15 – Répartition des élèves selon leur établissement

Le diagramme circulaire ci-dessus nous montre la répartition des élèves entre les écoles Gabriel Pereira (GP) et Mousinho da Silveira (MS). L'effectif des élèves est de 349 à Gabriel Pereira contre seulement 46 élèves soit 7,5 fois moins à Mousinho da Silveira.

```
#tableau des effectifs
z = df.groupby(by=["school", "GradeTEST"], as_index=False).size()
sumgp = z[z.school == "GP"]['size'].sum()
summs = z[z.school == "MS"]['size'].sum()
coef = (z.school == "GP") * sumgp + (z.school == "MS") * summs
z["freq"] = z["size"] / coef
z
```

	school	GradeTEST	size	freq
0	GP	average	102	0.292264
1	GP	good	54	0.154728
2	GP	poor	32	0.091691
3	GP	quite_well	129	0.369628
4	GP	very_good	32	0.091691
5	MS	average	13	0.282609
6	MS	good	5	0.108696
7	MS	poor	6	0.130435
8	MS	quite_well	20	0.434783
9	MS	very_good	2	0.043478

FIGURE 16 – Répartition des élèves selon leur établissement et leur catégorie de note

Ce tableau indique les effectifs et les fréquences du nombre d'élèves classé dans chaque catégorie en fonction de leur note en distinguant les deux établissements de notre dataset. Étant donné la différence significative des effectifs, nous allons nous focaliser sur la fréquence des catégories par élèves (freq) plutôt que sur l'effectif d'élèves dans chaque catégorie (size). Indépendamment de leur établissement, la majorité des étudiants ont une mention « average ». Le taux est de 29% à Gabriel Pereira contre 28% à Mousinho da Silvera ce qui représente une légère différence de -1%. Néanmoins, on relève des différences plus élevées dans les autres catégories. Dans les catégories affectées aux notes les plus hautes, les élèves de Gabriel Pereira ont une fréquence plus importante que ceux de Mousinho da Silvera. Au contraire, dans les catégories affectées aux notes les plus basses, les élèves de Gabriel Pereira ont une fréquence plus basse que ceux de Mousinho da Silvera. Respectivement, on observe des différences de 5% et 4.6% en faveur de l'établissement Gabriel Pereira dans les catégories « very_good » et « good ». D'autre part, on observe respectivement des différences de 6.3% et 3.9% en défaveur de l'établissement Mousinho da Silvera dans les catégories « quite_well » et « poor ».

En conclusion, les étudiants de l'établissement Gabriel Pereira ont obtenu de meilleur résultat que ceux de Mousinho da Silvera. Cette conclusion répond-elle à notre problématique ? En effet, le choix de l'école se fait selon plusieurs critères : ressources pédagogiques, réseau professionnel, corps enseignant de haut niveau, infrastructures, vie associative, environnement... Tout ces paramètres peuvent avoir un effet sur les résultats d'un étudiant. Toutefois, cette analyse est limitée car nous n'avons aucune connaissance sur les caractéristiques des établissements de notre dataset et nous ne savons pas s'il existe notamment une sélection des élèves en fonction de leur résultat ou non. Si un établissement accepte dès le départ uniquement des élèves à partir

d'un certain niveau scolaire alors on ne pourrait pas réellement évaluer l'impact du choix de l'école de cette école par rapport à une autre.

5.3.4 Education des parents

Nous nous demandons si le niveau d'éducation des parents a un impact sur les résultats d'un étudiant. Pour cette analyse, nous allons utiliser les colonnes contenant les niveaux d'éducation scolaire de la mère (Medu) et du père (Fedu). On rappelle que ces colonnes prennent des valeurs numériques allant de 0 à 4. On a : 0 - none, 1 - "primary education" (4th grade), 2 - "5th to 9th grade", 3 - "secondary education" et 4 - "higher education". Concrètement, plus le numéro est élevé, plus l'éducation est grande.

```
df.groupby(['Medu']).sum().plot(kind='pie', y='Grade')
```

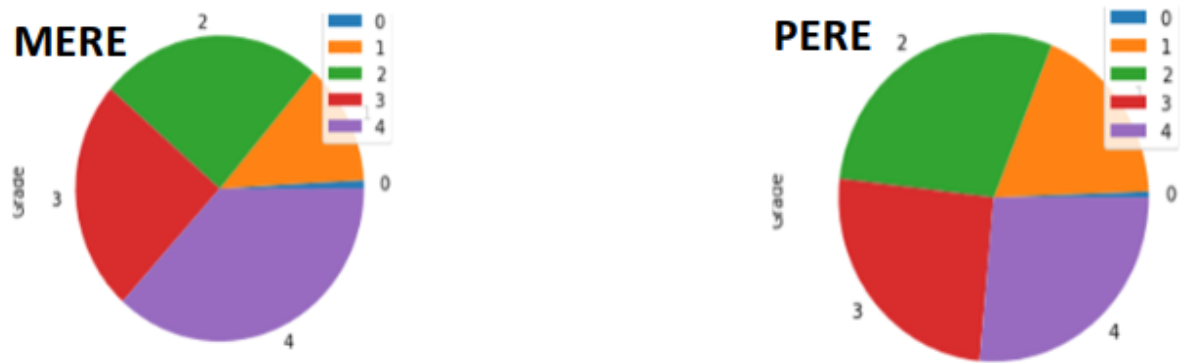


FIGURE 17 – Le niveau d'éducation scolaire des parents

Les diagrammes circulaires nous montrent la répartition des élèves selon le niveau scolaire de la mère et du père séparément. On remarque que le nombre d'élèves ayant une mère ou un père avec aucun niveau d'éducation scolaire est quasiment inexistant. Les observations sont similaires : les deux camemberts ont le même top 5. On retrouve davantage d'élèves ayant des parents avec un niveau d'éducation 4, puis 3, 2, 1 et enfin 0. Sur le camembert gauche, on remarque que beaucoup plus ont des mères ayant fait de longues études (niveau 4) en comparaison aux pères. On remarque également que la répartition des pères ayant fait des études de la 5ème année à la 9ème années (niveau 2) est la plus importante avant les pères ayant fait des études jusqu'en 4ème année.

```
#tableau des effectifs
z = df.groupby(by=["school", "GradeTEST"], as_index=False).size()
sumgp = z[z.school == "GP"]['size'].sum()
summs = z[z.school == "MS"]['size'].sum()
coef = (z.school == "GP") * sumgp + (z.school == "MS") * summs
z["freq"] = z["size"] / coef
z
```


	Medu	GradeTEST	size
0	0	average	1
1	0	good	2
2	1	average	23
3	1	good	6
4	1	poor	10
5	1	quite_well	19
6	1	very_good	1
7	2	average	31
8	2	good	9
9	2	poor	10
10	2	quite_well	48
11	2	very_good	5
12	3	average	30
13	3	good	13
14	3	poor	12
15	3	quite_well	35
16	3	very_good	9
17	4	average	30
18	4	good	29
19	4	poor	6
20	4	quite_well	47
21	4	very_good	19

	Fedu	GradeTEST	size
0	0	good	1
1	0	quite_well	1
2	1	average	35
3	1	good	8
4	1	poor	12
5	1	quite_well	23
6	1	very_good	4
7	2	average	28
8	2	good	20
9	2	poor	11
10	2	quite_well	49
11	2	very_good	7
12	3	average	32
13	3	good	13
14	3	poor	8
15	3	quite_well	37
16	3	very_good	10
17	4	average	20
18	4	good	17
19	4	poor	7
20	4	quite_well	39
21	4	very_good	13

FIGURE 18 – Catégories selon le niveau d'éducation scolaire des parents

Les tableaux représentent les mentions des élèves selon le niveau d'éducation scolaire des parents. En vert, nous pouvons voir les catégories qui comptent le plus d'élèves selon les niveaux 0, 1, 2, 3 et 4. De la même manière, en rouge sont représentés les catégories avec le moins d'élèves. Le jaune signifie que les effectifs sont identiques.

Les observations sont similaires entre « Medu » et « Fedu ». Le niveau "0" correspondant à un niveau d'éducation scolaire nulle est très peu représenté. On retrouve des élèves avec une mention « good » des deux côtés.

Toutefois, ce petit effectif ne nous permet pas de tirer une conclusion. Les élèves dont les parents possèdent un niveau d'éducation scolaire de niveau "1" ont majoritairement les mentions « average » et « quite_well ». Ces derniers sont plus nombreux à avoir la mention « poor » plutôt que les mentions « very_good » et « good ». Les élèves dont les parents possèdent un niveau d'éducation scolaire de niveau "2" sont plus nombreux que les précédents. Les observations sont similaires exceptions faite le passage de la mention maximum de « average » à « quite_well ». A partir du niveau "2", les élèves sont les plus nombreux à avoir des notes au-dessus de la moyenne. Le niveau "3" est le seul où nous remarquons une différence des extremums entre les parents. En se penchant davantage sur les effectifs, on voit que cette différence n'est pas très significative car les écarts ne sont pas très importants. Pour les pères, le minimum passe désormais de la mention « very_good » à la mention « poor ». Pour les mères, les extremums restent les mêmes que précédemment. Cependant, on relève que nous avons désormais plus d'élèves avec une mention « good » que « poor » des deux côtés. Dernièrement, les extremums évoluent au niveau "4" avec un maximum d'élèves ayant une mention « quite_well » et un minimum ayant une mention « poor ».

En conclusion, la mention d'un élève a tendance à augmenter en fonction du niveau d'éducation scolaire de ses parents.

5.3.5 Temps de transport

Nous nous demandons si le temps de transport effectué par un étudiant afin de se rendre à son établissement a un impact sur ses notes.

```
import matplotlib.pyplot as plt
%matplotlib inline

plt.hist(df['traveltime'])
plt.title('Temps de transport par étudiant')
plt.xlabel('Temps de transport')
plt.ylabel('effectif')
plt.show()
```

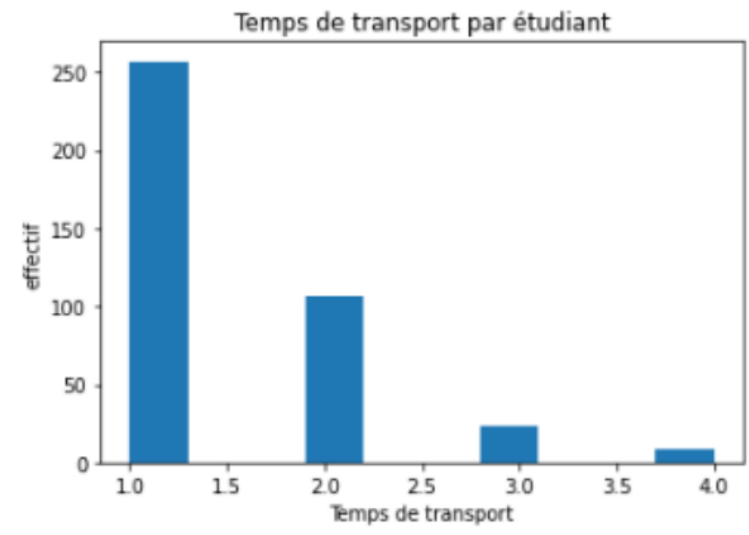


FIGURE 19 – Temps de transport par étudiant

Ce diagramme en bâton représente l'effectif des étudiants selon leur temps de trajets. Ici, nous avons fait appel à la colonne « traveltime ». Les variables sont quantitatives et prennent des valeurs qui évoluent graduellement selon le temps de trajet sur un intervalle de 1 à 4 avec : 1 - "<15 min.", 2 - "15 à 30 min.",

3 - "30 min. à 1h", ou "4 - >1h"). L'immense majorité des étudiants mettent moins de 14 minutes de chez eux jusqu'à leur établissement. Ensuite, plus le temps de trajet est long, plus l'effectif se réduit.

5.3.6 Géographie

Nous nous demandons si la zone d'habitation d'un étudiant a un impact sur ses notes.

```
b = df.loc[(df["address"]=="U"), ["address"]]
U_effectif = b.count()
print(U_effectif)
```

```
address    307
dtype: int64
```

```
b = df.loc[(df["address"]=="R"), ["address"]]
R_effectif = b.count()
print(R_effectif)
```

```
address     88
dtype: int64
```



FIGURE 20 – Répartition des étudiants selon leur zone géographique

Le diagramme circulaire nous montre la répartition des étudiants selon leur zone d'habitation, soit en zone rurale (R) ou alors en zone urbaine (U). On constate que beaucoup plus d'élèves habitent en zones urbaine. L'effectif d'élève habitant en zones urbaine est de 307 contre 88 pour ceux en zones rurale. En utilisant la fonction `.count()` sur python, nous avons pu relever des chiffres précis afin établir des statistiques. Ainsi, parmi les étudiants habitant en zones urbaine, 60% ont obtenu une note au-dessus de la moyenne contre 50% pour les étudiants habitant en zones rurale. Cet écart de 10% montre que les étudiants résidant en zones urbaine ont de meilleurs résultats.

5.3.7 Couple

Nous nous demandons si la situation amoureuse d'un étudiant a un impact sur ses notes.

```
df.loc[(df["romantic"]=="no") & (df["Grade"]<=10), ["romantic"]].count()
```

```
romantic    115  
dtype: int64
```

```
df.loc[(df["romantic"]=="no") & (df["Grade"]>=10), ["romantic"]].count()
```

```
romantic    158  
dtype: int64
```

```
df.loc[(df["romantic"]=="yes") & (df["Grade"]<=10), ["romantic"]].count()
```

```
romantic     64  
dtype: int64
```

```
df.loc[(df["romantic"]=="yes") & (df["Grade"]>=10), ["romantic"]].count()
```

```
romantic     73  
dtype: int64
```

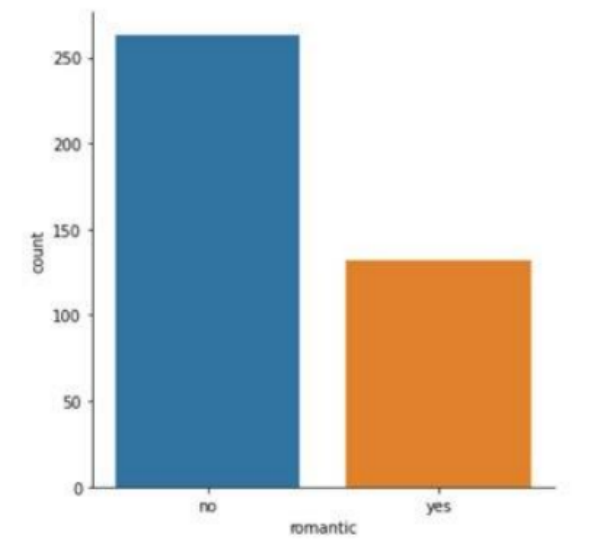
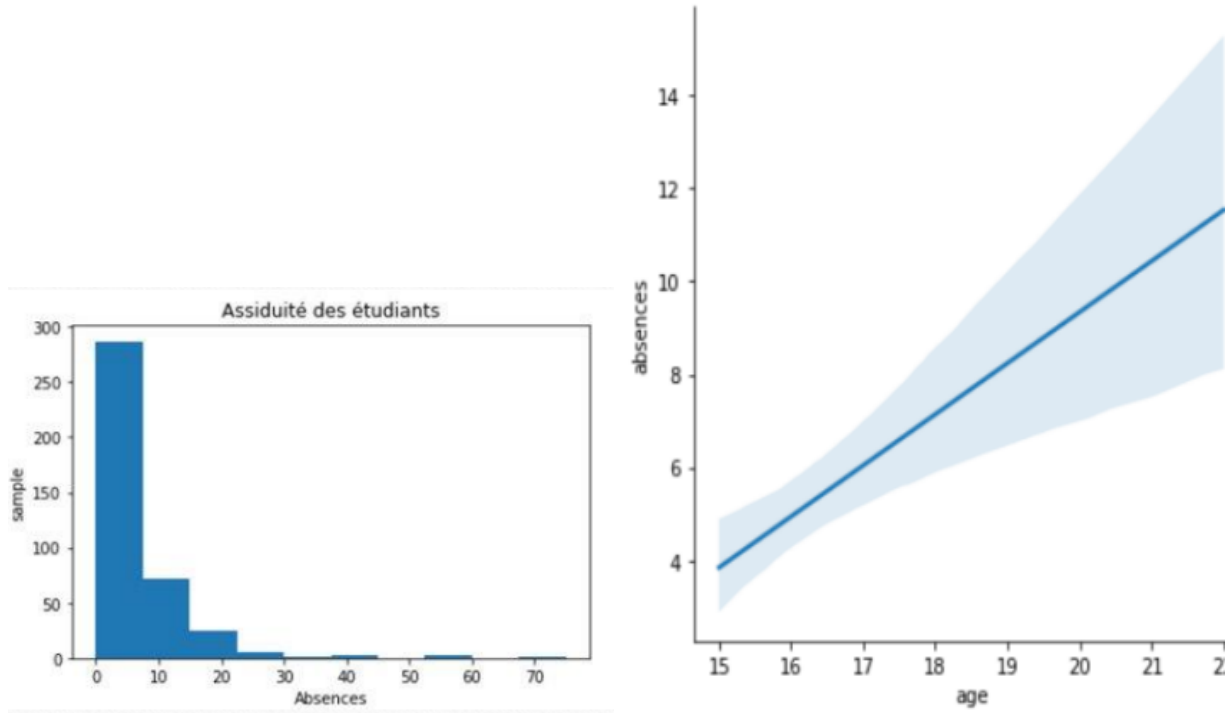


FIGURE 21 – Répartition des étudiants selon leur situation amoureuse

Le diagramme ci-dessus représente la répartition des étudiants selon leur situation amoureuse. Le nombre d'étudiant célibataire (no) est environ deux fois plus grand que le nombre d'étudiants en couple (yes). De la même manière que précédemment, nous allons utiliser la fonction `.count()` afin d'établir des statistiques. Ainsi, l'effectif d'étudiant en couple est de 163 contre 263 pour les célibataires. Parmi les célibataires, 60% ont des notes au-dessus de la moyenne contre 55% pour les étudiants en couple. Cet écart de 5% nous montre que les célibataires ont obtenu des résultats supérieurs aux autres.

5.3.8 Age et assiduité



Le premier diagramme à gauche représente l'histogramme des absences en cours en fonction du nombre d'élèves étudiés. Il en ressort que plus de la moitié des étudiants ont entre 0 et 10h absences. Par ailleurs, le second diagramme à droite nous présente le spectre obtenu par l'âge en fonction des heures d'absence et vient apporter plus de précision sur le premier. On peut donc conclure que le nombre d'heures d'absence augmente en fonction de l'âge.

exemple, on peut voir qu'à 15 ans le nombre d'absences est de 4 en moyenne tandis qu'à 22 ans il est de 11 en moyenne.

6 Conclusion

En conclusion, plusieurs facteurs impactent plus ou moins les résultats d'un étudiant tel que l'établissement, le niveau d'éducation des parents, le temps de transport, la zone d'habitation, la relation amoureuse. L'assiduité d'un étudiant est aussi un facteur à prendre en compte et elle est impactée par l'âge, les sorties et la consommation d'alcool. Plus la consommation d'alcool est forte, plus les résultats ont tendance à diminuer. Cela impacte davantage les hommes car ces derniers consomment plus d'alcool que les femmes. Plus les étudiants ont tendance à sortir, plus leur consommation d'alcool croît et leurs notes auront tendance à diminuer.

Dans le cadre de notre projet, nous avons utilisé deux méthodes de modélisation. L'arbre de décision s'est avéré être la modélisation la plus précise car il nous informe sur la prédiction des différents comportements possibles or la matrice de corrélation indique simplement les relations. Toutefois, la matrice de corrélation nous offre une bonne vue d'ensemble sur nos données quantitatives en nous indiquant les relations pertinentes.

Notre étude est limitée par le nombre d'individus de notre dataset qui pourrait être plus grand afin d'avoir des résultats plus précis. Notre dataset étant uniquement constitué de deux établissements portugais, dont l'un compte 7,5 fois plus d'élèves que l'autre, nous sommes limités par la géographie de notre étude et par l'effectif. Enfin, une analyse plus approfondie pourrait inclure la base de données "Portugais" afin d'avoir un ensemble global comportant une matière scientifique et une matière littéraire.

7 Source

Alcohol Effects On Study - AMAN CHAUHAN

Kaggle <https://www.kaggle.com/datasets/whenamancodes/alcohol-effects-on-study>