Noms : *Mohamed SELMI, Amir BEN MALLEM, Said MOHAMMED SEGHIR. GROUPE :14*

Description du jeu de données HCV

I.Introduction :

Notre sujet consiste en un problème de classification, il faudra diagnostiquer les maladies du foie à partir de différents échantillons d'analyses à l'aide de l'apprentissage automatique. Notre application devra classer le patient dans l'une des catégories suivantes : Donneur de sang, Hépatite C en incluant les complications (simple hépatite C, fibrose, cirrhose).

Le jeu de données mis à disposition est : https://archive.ics.uci.edu/ml/datasets/HCV+data

II.Description des données :

Nous avons au total 615 observations.

II.1 Variables Explicatives :

	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT	Âge
Min	14.9	11.3	0.9	10.6	0.8	1.42	1.43	8	4.5	44.8	19
Max	82.2	416.6	325.3	324	254	16.41	9.67	1079.1	650.9	90	77
Moy	41.62	68.28	28.45	34.78	11.39	8.19	5.36	81.28	39.53	72.04	47.40
V.abs	1	18	1	0	0	0	10	0	0	1	0

- *Valeurs Manquantes* : Nous avons remplacé les vides par la moyenne de la colonne appartenant à la même catégorie.

II.2 : Variable expliquée :

La colonne <catégorie> est la cible.

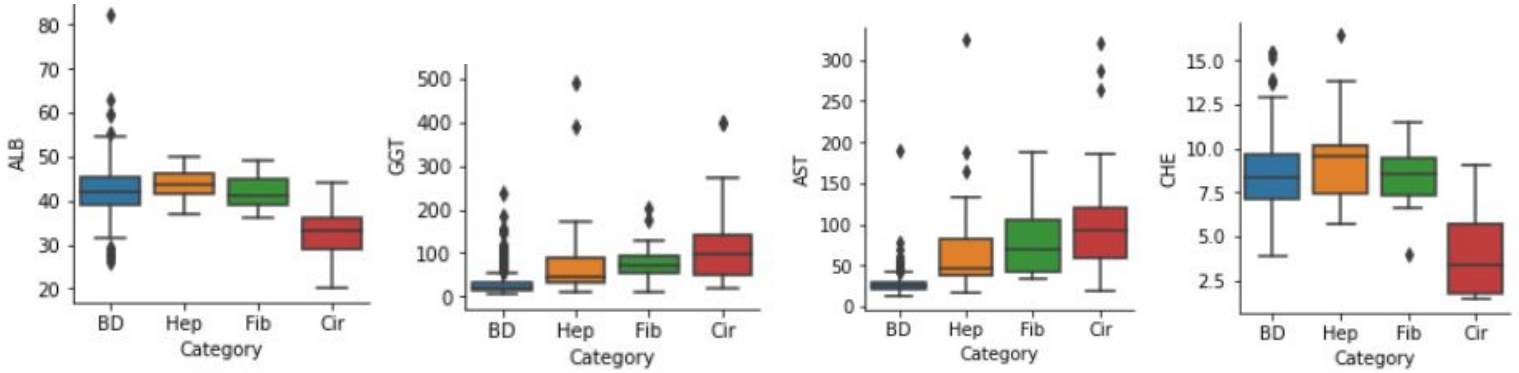
Value	Count	Frequency (%)
0=Blood Donor	533	86.7%
3=Cirrhosis	30	4.9%
1=Hepatitis	24	3.9%
2=Fibrosis	21	3.4%
0s=suspect Blood Donor	7	1.1%

Nous constatons une répartition inéquitable des données du dataset entre les catégories.

III. Analyse descriptive des données :

III.1 : Boîtes à moustache :

Un échantillon des boîtes à moustache, ici sur les variables **ALB**, **GGT**, **AST** et **CHE** en fonction de chaque catégorie.



-Mots clés : BD = Blood donor, Hep = Hépatite, Fib = Fibrose, Cir = cirrhose.

On remarque que le taux de ALB (Albumine) et CHE (Cholinestérase) inférieur ou supérieur à la moyenne est un signe de maladie hépatique comme une hépatite ou une cirrhose.

On remarque que plus le taux de CGT et AST sont élevés, plus le foie est touché, plus le niveau de lésions hépatiques est élevé.

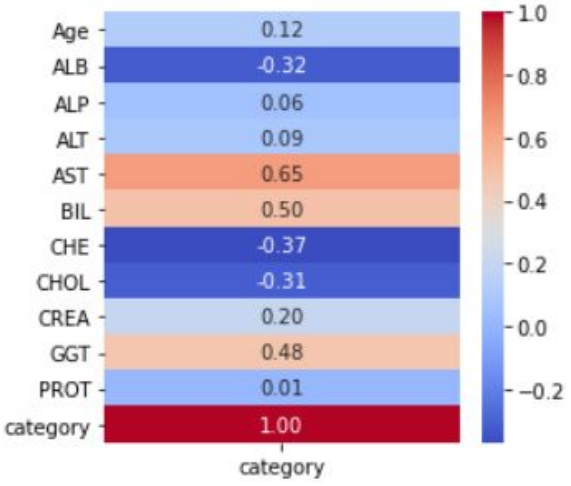
III.2 : Corrélation linéaire entre les variables explicatives :

	Age	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
Age	1.000000	-0.170138	0.165691	-0.042102	0.073111	0.035758	-0.080787	0.146670	-0.012079	0.141081	-0.122002
ALB	-0.170138	1.000000	-0.099209	0.092611	-0.180774	-0.248960	0.381006	0.166669	-0.021206	-0.087750	0.480602
ALP	0.165691	-0.099209	1.000000	0.071101	0.024504	0.064349	0.048188	0.135030	0.162796	0.421882	-0.006812
ALT	-0.042102	0.092611	0.071101	1.000000	0.215161	-0.032033	0.173361	0.062697	-0.034862	0.169919	0.204289
AST	0.073111	-0.180774	0.024504	0.215161	1.000000	0.322628	-0.234705	-0.220186	-0.013165	0.496529	0.071277
BIL	0.035758	-0.248960	0.064349	-0.032033	0.322628	1.000000	-0.342546	-0.185584	0.031017	0.238781	-0.065814
CHE	-0.080787	0.381006	0.048188	0.173361	-0.234705	-0.342546	1.000000	0.425648	-0.012087	-0.091646	0.285723
CHOL	0.146670	0.166669	0.135030	0.062697	-0.220186	-0.185584	0.425648	1.000000	-0.055956	0.013399	0.161668
CREA	-0.012079	-0.021206	0.162796	-0.034862	-0.013165	0.031017	-0.012087	-0.055956	1.000000	0.123133	-0.061982
GGT	0.141081	-0.087750	0.421882	0.169919	0.496529	0.238781	-0.091646	0.013399	0.123133	1.000000	0.073540
PROT	-0.122002	0.480602	-0.006812	0.204289	0.071277	-0.065814	0.285723	0.161668	-0.061982	0.073540	1.000000

On remarque qu'il n y a pas de relation linéaire forte, positive ou négative, entre les variables.

III.3 Lien entre la variable expliquée et les variables explicatives

Ci-dessous, nous avons la matrice de corrélation entre la catégorie et les différentes variables explicatives :



- Les variables à corrélation positive les plus importantes sont l'AST, la BIL et la GGT.
- Les variables à corrélation négative les plus importantes sont notamment le CHE et l'ALB.

On peut constater que l'augmentation du taux de certaines variables impactent beaucoup plus la catégorie de la personne que la baisse d'autres variables, ce qui est confirmé par les documents scientifiques traitants de l'Hépatite qui stipulent que l'augmentation de l'AST, la GGT et la BIL sont parmi les facteurs les plus importants de détection de la maladie.

IV. Conclusion :

Après l'étude statistique et la description des données, nous avons remplacé les variables manquantes et les donneurs suspects ont été exclus du jeu de données de final. L'énoncé demandant la classification en 4 catégories (personnes saines et personnes atteintes d'hépatite avec complication), le nombre très faible de personnes suspectes et les données quelquefois aberrantes de leurs analyses nous ont poussé à les supprimer du jeu de données final.
Il y a également des variables explicatives qui influent de façon plus importante sur la variable expliquée que d'autres.