

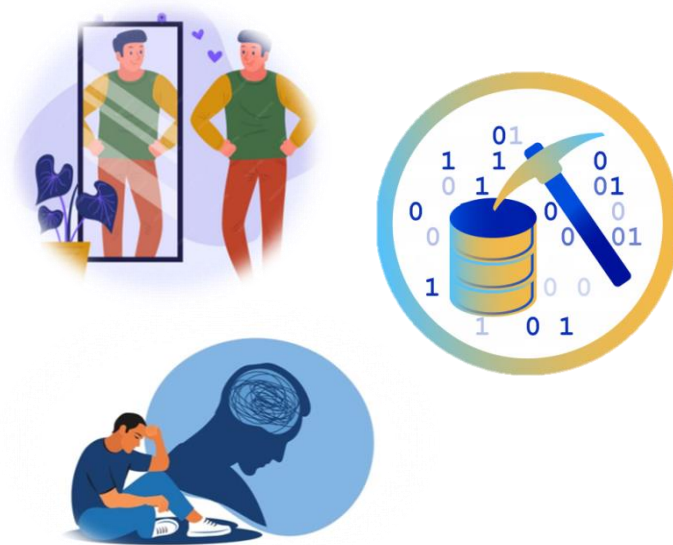
Master : Système d'informations décisionnels et imagerie (SIDI)

Module : data mining & machine learning -----Demi module : DATA MINING

Semestre : S3 année universitaire 2023/2024

Mini projet :

« **Exploration des Facteurs Psychosociaux : Une Approche Data Mining pour la Mesure de l'Estime de Soi et du Degré de Dépression et leur corrélation** »



Encadré par : Prof M. Mohamed Sabiri

Réalisé par : Hamid JABRANE

Introduction :	3
1) Définition de problème :	4
2) Collecte de données :	4
Questionnaire :	4
Teste d'estimation de soi de Morris Rosenberg	4
Teste de dépression d'Aaron Beck	5
3) Nettoyage et transformation des données :	7
Le codage des données :	7
Les valeurs manquantes :	7
Les valeurs aberrantes :	9
Réduire la dimension du problème en trois composantes principales :	10
4) Appliquer les techniques de fouille de données :	15
La régression :	15
La classification :	16
Arbre de décision :	18
Clustering k-means	18
5) Interprétation du module et établissement des conclusions :	20
Régression logistique :	20
Etablir conclusions :	21
6) Gérer la connaissance découverte :	21
Construire un modèle :	21
Conclusion générale :	23
Annexes	23

Introduction :

Ce mini-projet s'inscrit dans de la mise en pratique des compétences et les connaissances acquises avec les méthodes avancées de l'exploration de données. L'objectif est d'appliquer des techniques de Data Mining, dans le domaine psychosocial, pour extraire des insights pertinents à partir de mesures d'estime de soi et de dépression, en mettant en évidence les relations entre ces variables et en explorant des modèles potentiels qui peuvent contribuer à la prise de décision en matière de santé mentale.

Dans le domaine de la psychologie et des sciences sociales, la mesure des phénomènes psychosociaux joue un rôle crucial dans la compréhension de la santé mentale et émotionnelle des individus. Ces mesures fournissent des outils précieux pour évaluer divers aspects de la vie psychologique, tels que l'estime de soi et le degré de dépression. Ces deux concepts sont d'une importance capitale, car ils sont étroitement liés au bien-être général et à la qualité de vie des individus.

Les résultats de cette évaluation aident les professionnels de la santé mentale, tels que les psychologues ou les psychiatres, à estimer le niveau de dépression chez un individu et à orienter le traitement approprié.

Dans ce travail on va utiliser l'outil Orange qu'est un logiciel open-source dédié à l'analyse de données et au data mining. Il offre une interface visuelle conviviale permettant aux utilisateurs de construire et d'explorer des modèles de données facilitant ainsi la compréhension des résultats par les utilisateurs. Avec Orange, les utilisateurs peuvent explorer des ensembles de données, effectuer des analyses statistiques, construire des modèles prédictifs, et visualiser les résultats, le tout à travers une interface graphique intuitive.

1) Définition de problème :

Dans cette étude psycho-sociale on vise à mesurer la corrélation entre trois facteurs :

- ✓ L'indicateur social ;
- ✓ L'estime de soi ;
- ✓ La dépression.

C'est-à-dire trouver la relation ou les relations entre ces trois facteurs.

2) Collecte de données :

On a collecté les données à travers trois instruments de mesure :

Questionnaire : qui contient un ensemble des questions :

- Personnelles de l'élève : Age, Sexe, Niveau scolaire, interne/externe...
- Famille : niveau scolaire des parents, leur situation (mariés, divorcés...) ...

Aperçu sur les éléments de ce questionnaire :

B4	f _x					
B	C	D	E	F	G	H
اسم المستشار	الرقم المخصص لورقة التلميذ :	الجنس (نكر 1 - أنثى 2) :	السن بالسنوات (كتابة العدد فقط) :	المستوى الدراسي :	الوسط (حضري 1 - قروي 2)	نوع التعليم (عمومي 1 - خصوصي الوسيط)
حميد جبران	1	1	15	TC		1

B4	f _x				
I	J	K	L	M	N
هل تعيش مع (الأم والأب 1 - الأم) :	مهنة الأب أو ولي الأمر :	مهنة الأم (أمي 1 - أمهنة الأم) :	أبنة المستوى الدراسي للأب (أمي 1 - أبنة المستوى الدراسي للأب) :	أبنة المستوى الدراسي للأم (أمي 1 - أبنة المستوى الدراسي للأم) :	هل لديك مع (الأم والأب 1 - الأم) :
2	مياوم	ريت بيت	2	1	1

B4	f _x			
O	P	Q	R	
هل لديك في أسرتك شخص مدمن على (مخدرات، كحول، ...) :	هل لديك في أسرتك شخص يعاني من مرض نفسي :	هل لديك في أسرتك شخص يعاني من مرض نفسي :	هل لديك في أسرتك شخص مدمن على (مخدرات، كحول، ...) :	هل لديك في أسرتك شخص مدمن على (مخدرات، كحول، ...) :
1	1	2	2	

Teste d'estimation de soi de Morris Rosenberg Ce test explore différentes dimensions de l'estime de soi en demandant à la personne évaluée de réagir à des affirmations telles que "Je me sens généralement satisfait(e) de moi-même" ou "En général, je tends à penser que je suis un échec". Les participants sont invités à choisir la réponse qui correspond le mieux à leur propre perception. Il comprend dix items, et chaque item propose quatre choix de réponses.

L'échelle de réponse généralement utilisée pour chaque item va de "tout à fait d'accord" à "plutôt d'accord", "plutôt en désaccord" et "tout à fait en désaccord".

Aperçu sur ce teste :

S	T	U	V	W
1. أنا على العموم مقتنع بذاتي -	2. أشعر أحياناً أنه لا قيمة لي -	3. أشعر أحياناً أنني أتوفر على خصال حميدة -	4. لا أستطيع الآخرون القيام بها -	5. لا أستطيع أن أفخر بنفسى -
3	4	4	4	3

X	Y	Z	AA	AB
6. أشعر أحياناً أنني عديم الجدوى -	7. أفضل من بعض الناس -	8. أألم مزيداً من الاحترام لذاتي -	9. إلى الاعتقاد أنني شخص فاضل -	10. رأيي إيجابى عن ذاتى -
3	4	2	4	3

Teste de dépression d'Aaron Beck est en effet composé de 21 items. Cet inventaire est conçu pour évaluer la sévérité des symptômes dépressifs chez une personne. Chacun des 21 items représente une déclaration liée à la vie quotidienne et aux symptômes dépressifs que la personne peut ressentir au cours de sa journée.

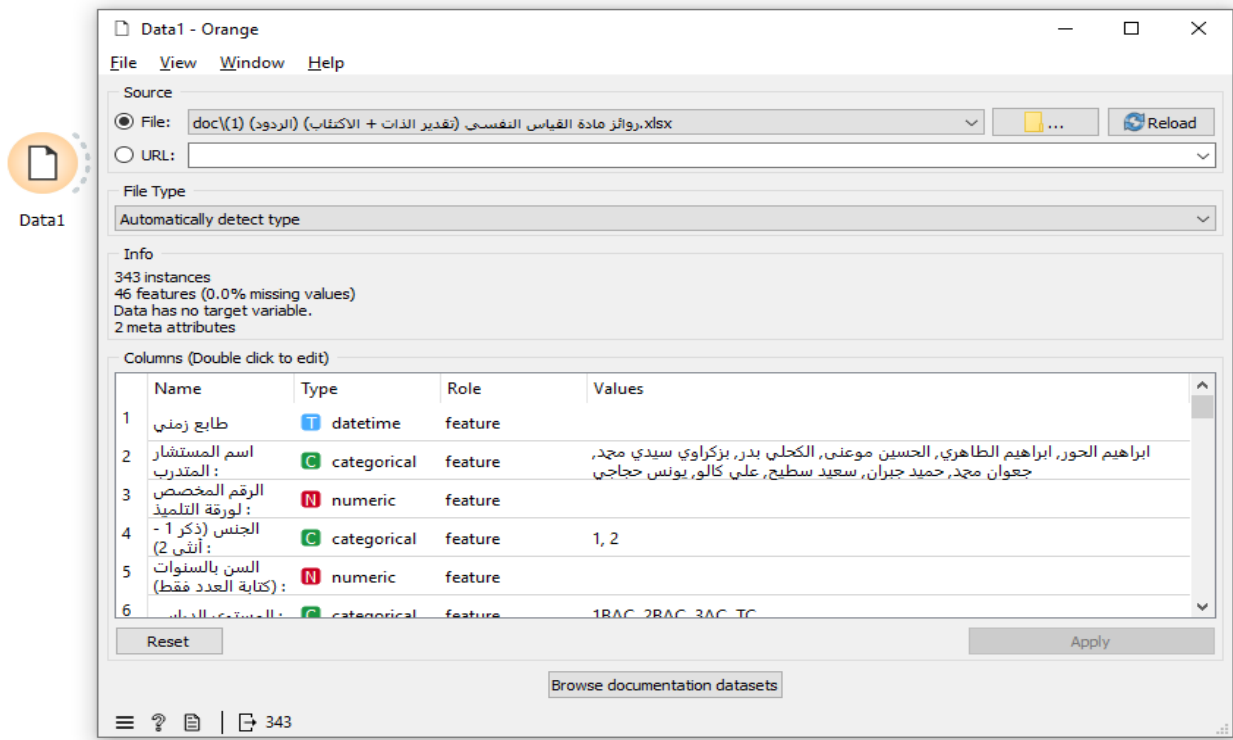
AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM
الشعور بالحزن :	نمط النوم :	الاهتمام :	أفكار الانتحار :	التشاؤم :	فقدان الاستمتاع :	الإرهاق :	حب الذات :	التركيز :	التفكير والاستشارة :	الانتعاش :
1	0	1	0	0	0	1	0	1	0	0

AN	AO	AP	AQ	AR	AS	AT	AU	AV
شهية الطعام :	الانزعاج :	الإحساس بتأنيب الضمير :	مشاعر العقاب :	البكاء :	نقد الذات :	الشعور بالفشل :	التردد :	فقدان الطاقة :
1	0	0	2	0	0	0	0	0

Par exemple, la personne évaluée devrait indiquer dans quelle mesure elle se sent en accord avec des déclarations telles que "Je me sens triste" ou "Je n'éprouve pas de plaisir". Les réponses sont évaluées sur une échelle de 0 à 3 (**0 : pas du tout, 1 : un peu, 2 : modérément, 3 : beaucoup**).

Après que les élèves (la population cible) ont rempli ces instruments, nous avons initié la phase de saisie des données dans un classeur Excel.

Les informations sur la taille des données :



Alors on a 343 observations en total et 48 items 46 (features) et 2 méta-données

Visualisation des données brutes :

Diagram showing the data flow from **Data1** to **Data Table**.

	المستوى الدراسي	الجنس (ذكر 1 - أنثى 2)	السن بالسنوات (كتابة العدد فقط)	المهنة	الرقم المخصص : لورقة التلميذ
201	1	1	1	ربة بيت	2
202	1	1	1	ربة بيت	4
203	1	1	1	ربة بيت	3
204	1	1	1	رئيسة مدرسة الف...	5
205	1	1	1	ربة بيت	1
206	1	1	1	ربة بيت	3
207	1	1	1	ربة بيت	1
208	1	1	1	فلاحة	1
209	1	1	1	ربة بيت	3
210	1	1	1	ربة بيت	2
211	1	1	1	ربة بيت	5
212	1	1	1	ربة بيت	2
213	1	1	1	ربة بيت	4
214	1	1	1	ربة بيت	4
215	1	1	1	ربة بيت	3
216	1	1	1	ربة بيت	5
217	1	1	1	ربة بيت	5
218	1	1	1	ربة بيت	4
219	1	1	1	متوفاة	3
220	1	1	1	ربة بيت	2

D'après cette visualisation on constate que la base de données est trop grande ?

3) Nettoyage et transformation des données :

Le codage des données :

Genre : masculin → 1 féminin → 2 type catégorielle

Age : c'est nombre discret type numérique

Niveau scolaire : les élèves de 3^{ème} année collégiale (3) – Tronc commun (4)

Milieu : Urbain (1) et Rural (2)

Enseignement : Public (1) et Privé (2)

Nombre des frères ou sœurs : nombre discret

Profession de père et celle de mère

Les autres questions sur la situation familiale on a deux choix : Oui (1) et Non(2)

Pour le teste d'estime de soi : chaque item peut avoir une réponse selon les déclarations :

Déclarations (items) positives : "tout à fait d'accord (4) " à "plutôt d'accord (3) ", "plutôt en désaccord (2) " et "tout à fait en désaccord (1) "

Déclarations (items) négatives : "tout à fait d'accord (1) " à "plutôt d'accord (2) ", "plutôt en désaccord (3) " et "tout à fait en désaccord (4) "

Pour le teste de dépression chaque item peut avoir une réponse suivant l'échelle :

(0 : pas du tout, 1 : un peu, 2 : modérément, 3 : beaucoup).

Les valeurs manquantes :

Pour les valeurs manquantes on va faire un impute, qui nous permet de : supprimer les observations contenant les valeurs manquantes, ou de les attribuer une valeur aléatoire, ou de les remplacer par la moyenne de la colonne, ou en se basant sur un modèle d'arborescence ou bien de les remplacer par la valeur la plus fréquente. Puisque nous avons une quantité suffisante des données on va faire la suppression

Impute - Orange

File View Window Help

Default Method

- ☐ Don't impute
- ☐ Average/Most frequent
- ☐ As a distinct value
- ☐ Fixed values; numeric variables: 0, time: 1970-01-01 01:00:00
- ☐ Model-based imputer (simple tree)
- ☐ Random values
- ☒ Remove instances with unknown values

Individual Attribute Settings

Filter...

- ☒ المستوى الدراسي :
- ☒ الوسط (حضري 1 - قروي 2)
- ☒ نوع التعليم (عمومي 1 - خصوصي 2)
- ☒ مهنة الأم :
- ☒ المستوى الدراسي للأب (أمي 1 - ابتدائي 2)
- ☒ المستوى الدراسي للأم (أمية 1 - ابتدائي 2)
- ☒ هل تعيش مع (الأب والأب 1 - الأم لوحدها 2)
- ☒ الوضعية العائلية (الأب والأم مجتمعان 1 - الأم والأب 2)
- ☒ محل السكني أثناء الدراسة (بيت الأسرة 1 - بيت 2)
- ☒ هل لديك في أسرتك شخص يعاني من مرض 1

☒ Apply Automatically

343 | - 332

Voilà le résultat obtenu après cette suppression : (on a supprimé 9 observations)

Data Table (1) - Orange

File Edit View Window Help

Info

332 instances (no missing data)
46 features
No target variable.
2 meta attributes

Variables

☒ Send Automatically

332 | 332 | 332

المعرف	وسط	نوع التعليم (عمومي 1 - خصوصي 2)	مهنة الأم :	المستوى الدراسي للأب (أمي 1 - ابتدائي 2)	المستوى الدراسي للأم (أمية 1 - ابتدائي 2)	هل تعيش مع (الأب والأب 1 - الأم لوحدها 2)	الوضعية العائلية (الأب والأم مجتمعان 1 - الأم والأب 2)	محل السكني أثناء الدراسة (بيت الأسرة 1 - بيت 2)	هل لديك في أسرتك شخص يعاني من مرض 1
327	1	1	ربة بيت	5	5				
328	1	1	ربة بيت	3	1				
329	1	1	ربة بيت	2	3				
330	1	1	ربة بيت	5	2				
331	1	1	ربة بيت	1	3				

La distribution des données :

Distributions - Orange

File Edit View Window Help

Variable

Filter...

- ☒ 8 - أأرغب في أن أنال مزيدا من الاحترام لذا 8 -
- ☒ أنا أميل إلى الاعتقاد أنني شخص فاضل 9 -
- ☒ رأيي إيجابي عن ذاتي 10 -
- ☒ الشعور بالحب :

☐ Sort categories by frequency

Distribution

Fitted distribution: Normal

☒ Apply Automatically

332 | 332 | 36

Frequency

10

8

6

4

2

0

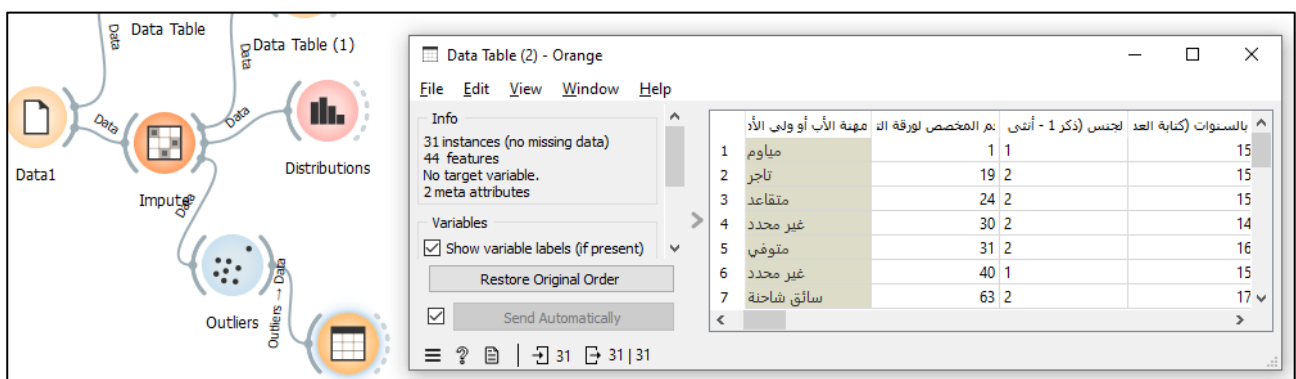
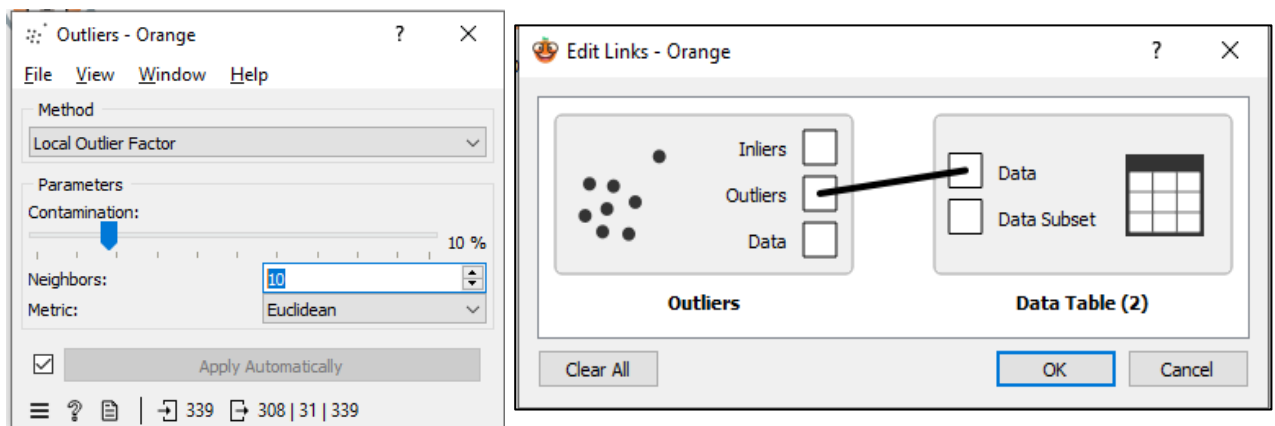
0 100 200 300

$\mu=174.51, \sigma=99.62$

الرقم المخصص لورقة التلميذ :

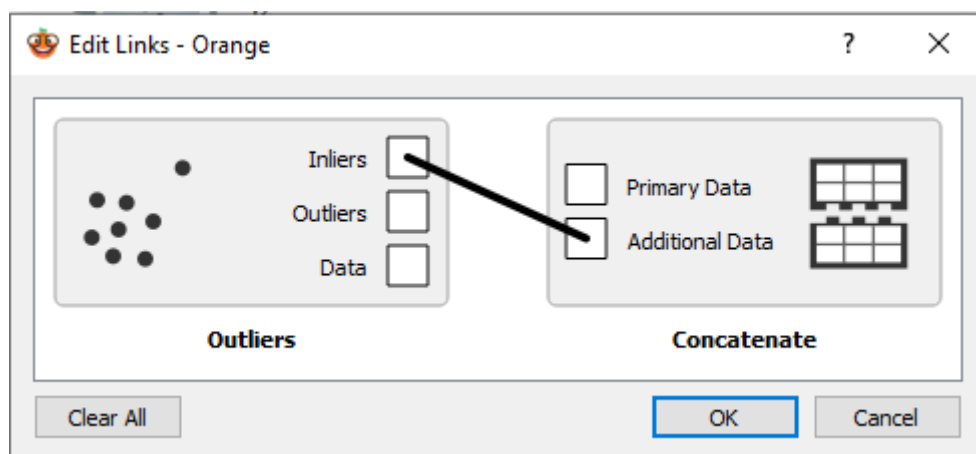
Les valeurs aberrantes :

Pour cela, sur orange, on va ajouter (Outiliers) :



On a détecté 31 lignes contenant des valeurs aberrantes.

Pour supprimer ces valeurs aberrantes : on ajoute « Concatenate »



Alors les valeurs aberrantes ont été supprimées : il nous reste maintenant 308 lignes



Réduire la dimension du problème en trois composantes principales :

La dimension 1 : indicateur social

Sur Excel on a fait une équation qui calcule cette indice $\rightarrow 1/(\text{somme des valeurs}) \times 10$

La dimension 2 : Niveau Estime de soi

Calculer la somme de tous les items du test : selon la somme on aura trois niveaux :

Si somme < 20 alors niveau = faible

Si somme ≥ 20 et somme < 30 alors niveau = moyen

Si somme ≥ 30 alors niveau = élevé

Image explicative :

=SI(D2<20;"Faible";SI(D2<30;"Moyen";"Elevé"))				
A	B	C	D	E
Genre	Age	Indicateur_Sociale	Test_Estime_De_Soi	EstimeDeSoi
1	15	15,17	34	Elevé
1	16	7,70	18	Faible
1	15	14,33	36	Elevé
2	15	11,00	34	Elevé
3	15	18,50	34	Elevé

La dimension 3 : Degré de Dépression

Calculer la somme de tous les items du test : selon la somme on aura trois niveaux :

Si somme <10 alors degré = état normal

Si somme >=10 et somme <19 alors degré=Dépression légère

Si somme >=19 et somme<29 alors degré=Dépression modérée

Si somme>=29 alors degré=Dépression sévère

Image explicative de la fonction utilisée :

=SI(E3<10;"Etat normale";SI(E3<19;"Dépression légère";SI(E3<29;"Dépression modéréé";"Dépression sévère")))					
	D	E	F	G	H
e	Indicateur_Sociale	Test_Depression	DegréDépression		
	15,17	7	Etat normale		
	7,70	28	Dépression modéréé		
	14,33	1	Etat normale		

Plus les informations élémentaires : Genre et Age

Après ces modifications on a obtenu une base de données simple à manipuler :

Importation de la nouvelle base de données dans l'outil Orange :

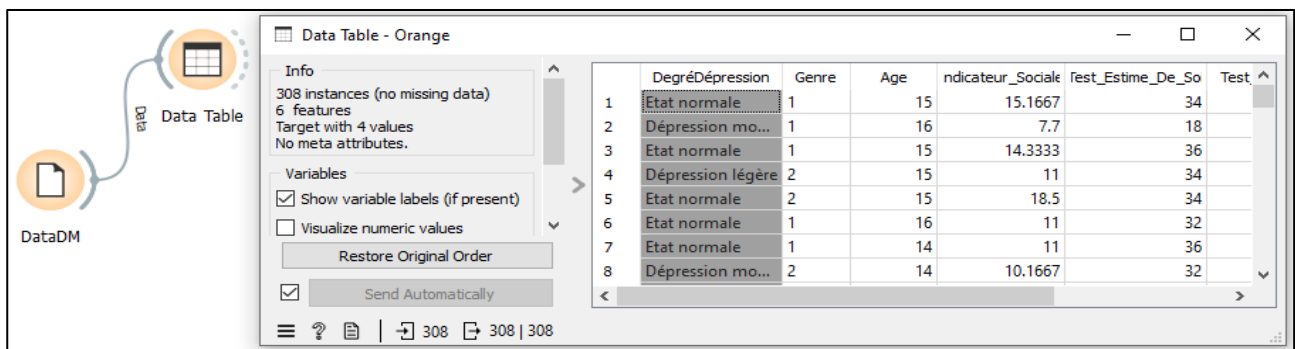
The screenshot shows the Orange Data Mining software interface. The 'DataDM' widget is selected, and its configuration panel is visible. The 'Source' section shows the file 'data mining\DataDM.xlsx' loaded. The 'File Type' is set to 'Automatically detect type'. The 'Info' section displays: 308 instances, 7 features (no missing values), Data has no target variable, and 0 meta attributes. The 'Columns' section lists 7 variables:

Name	Type	Role	Values
1 Genre	Categorical	feature	1, 2
2 Age	Numeric	feature	
3 Indicateur_Soci...	Numeric	feature	
4 Test_Estime_De...	Numeric	feature	
5 Test_Depression	Numeric	feature	
6 EstimeDeSoi	Categorical	target	Elevé, Faible, Moyen
7 DegréDépress...	Categorical	target	Dépression légère, Dépression modéréé, Dépression sévère, Etat normale

Buttons for 'Reset' and 'Apply' are visible at the bottom of the configuration panel. A 'Browse documentation datasets' button is also present.

Remarque : Alors on a 7 éléments : 5 features et 2 targets (ici on Orange il n'accepte pas les 2), donc va les traiter chacun à part.

Visualisation des données :



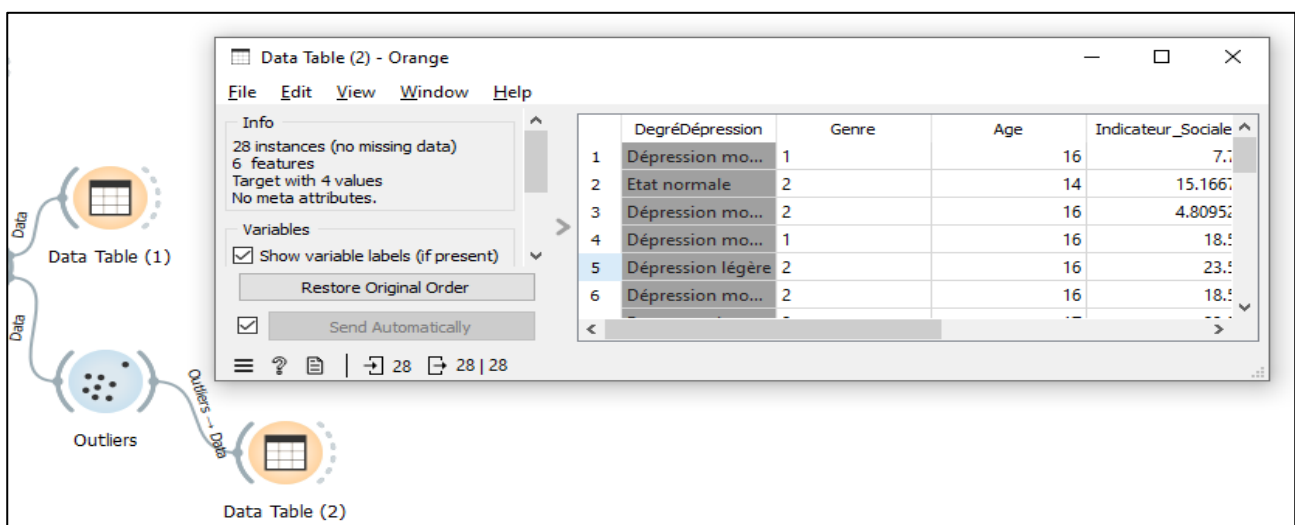
The screenshot shows the Orange Data Mining software interface. On the left, a workflow consists of a 'DataDM' widget connected to a 'Data Table' widget. The 'Data Table' widget's properties panel is open, displaying the following information:

- Info:** 308 instances (no missing data), 6 features, Target with 4 values, No meta attributes.
- Variables:** ☒ Show variable labels (if present), ☐ Visualize numeric values.
- Buttons:** Restore Original Order, Send Automatically (checked).
- Status:** 308 | 308 | 308

The data preview table is as follows:

	DegréDépression	Genre	Age	Indicateur_Sociale	Test_Estime_De_So	Test
1	Etat normale	1	15	15.1667	34	
2	Dépression mo...	1	16	7.7	18	
3	Etat normale	1	15	14.3333	36	
4	Dépression légère	2	15	11	34	
5	Etat normale	2	15	18.5	34	
6	Etat normale	1	16	11	32	
7	Etat normale	1	14	11	36	
8	Dépression mo...	2	14	10.1667	32	

Alors dans cette base de données on n'a pas les valeurs manquantes, cependant pour les valeurs aberrantes (28 valeurs) :



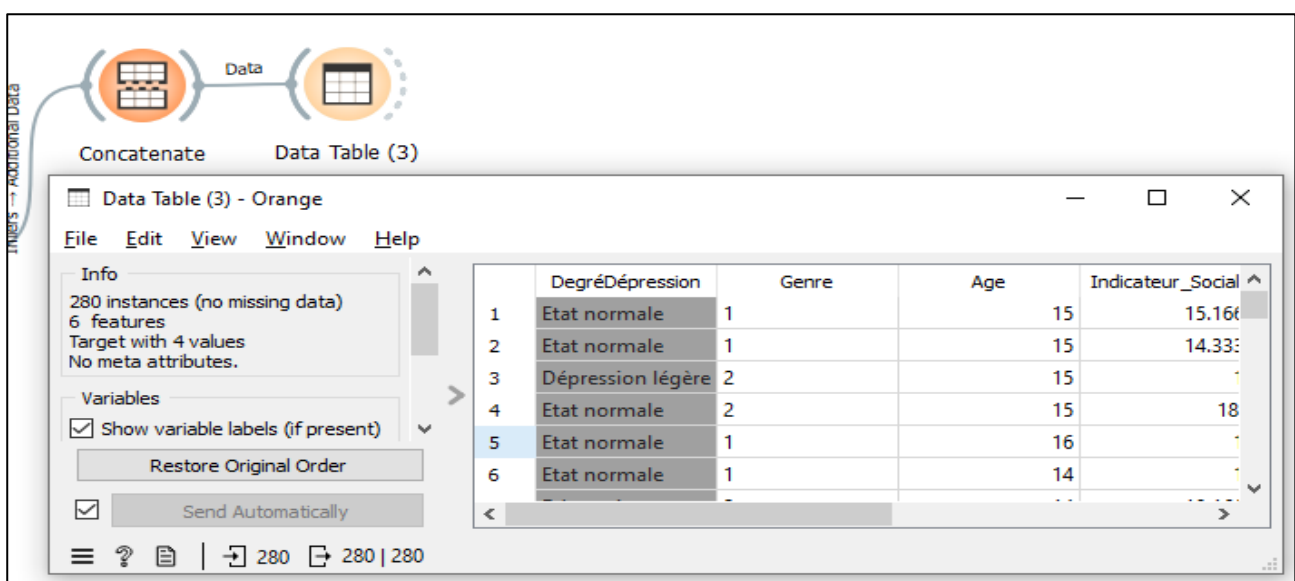
The screenshot shows a workflow in Orange Data Mining. It starts with 'Data Table (1)', followed by an 'Outliers' widget, and then 'Data Table (2)'. The 'Data Table (2)' widget's properties panel is open, showing:

- Info:** 28 instances (no missing data), 6 features, Target with 4 values, No meta attributes.
- Variables:** ☒ Show variable labels (if present).
- Buttons:** Restore Original Order, Send Automatically (checked).
- Status:** 28 | 28 | 28

The data preview table is as follows:

	DegréDépression	Genre	Age	Indicateur_Sociale
1	Dépression mo...	1	16	7.7
2	Etat normale	2	14	15.1667
3	Dépression mo...	2	16	4.80952
4	Dépression mo...	1	16	18.5
5	Dépression légère	2	16	23.1667
6	Dépression mo...	2	16	18.5

On va les supprimer avec « Concatenate » :



The screenshot shows a workflow in Orange Data Mining. It starts with 'Concatenate' (receiving input from 'Additional Data'), followed by 'Data Table (3)'. The 'Data Table (3)' widget's properties panel is open, showing:

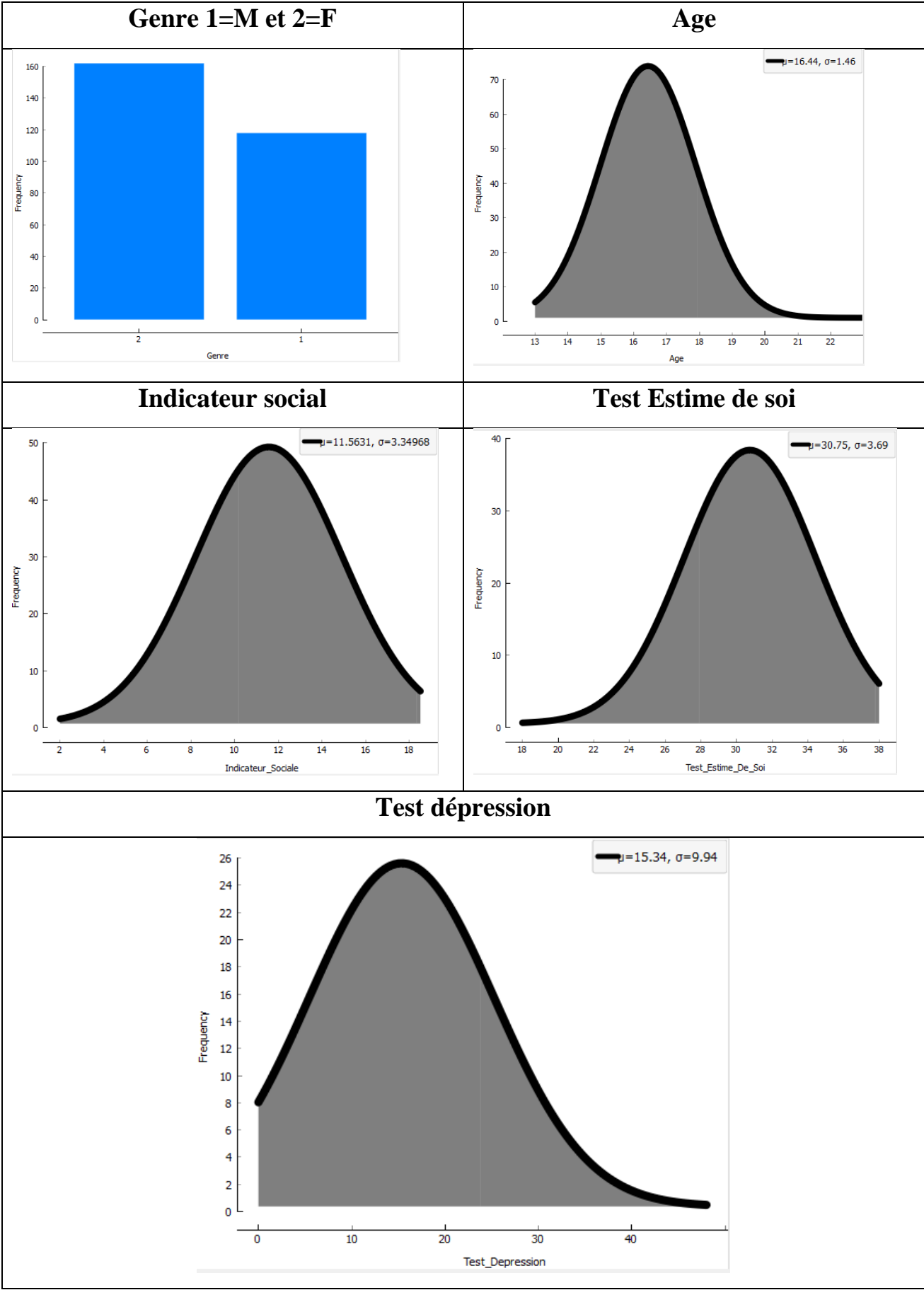
- Info:** 280 instances (no missing data), 6 features, Target with 4 values, No meta attributes.
- Variables:** ☒ Show variable labels (if present).
- Buttons:** Restore Original Order, Send Automatically (checked).
- Status:** 280 | 280 | 280

The data preview table is as follows:

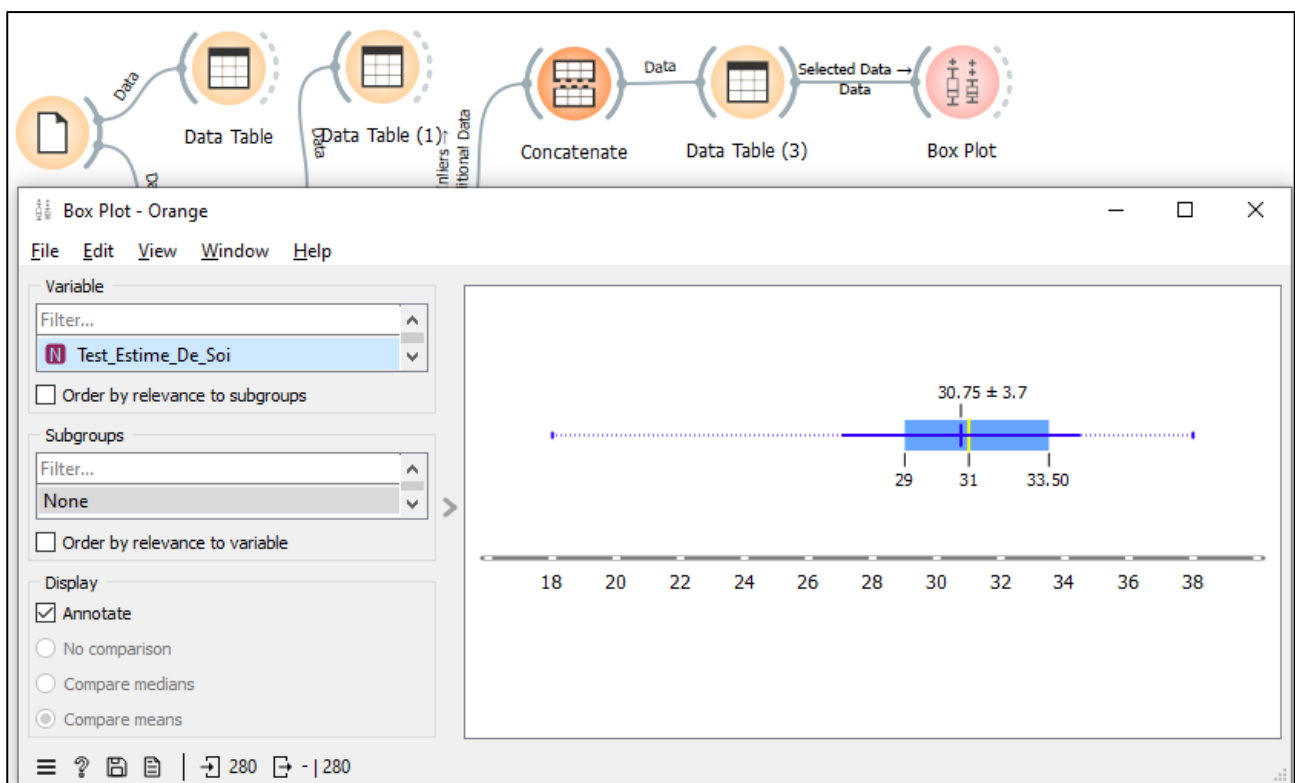
	DegréDépression	Genre	Age	Indicateur_Sociale
1	Etat normale	1	15	15.1667
2	Etat normale	1	15	14.3333
3	Dépression légère	2	15	11
4	Etat normale	2	15	18.5
5	Etat normale	1	16	11
6	Etat normale	1	14	11

Alors il nous reste maintenant **280 observations** dans la totalité de la base de données.

Distributions des données en affichant la fréquence de chaque élément, la moyenne et l'écart type:



On peut également consulter les autres mesures statistiques (mode, médiane, max, min) sur les variables en utilisant la composante « **Box plot** » : **boite à moustaches**



Les caractéristiques de l'ensemble de données :

The screenshot shows the 'Feature Statistics - Orange' window. It displays a table of statistical measures for various features. The table has columns for Name, Mean, Mode, Median, Dispersion, Min., Max., and Missing. The features listed are Age, Indicateur_Soci..., Test_Estime_De..., Test_Depression, Genre, EstimeDeSoi, and DegréDépression. The 'Test_Estime_De...' feature is highlighted. The window also includes a 'Color' dropdown set to 'DegréDépression' and a 'Send Automatically' checkbox.

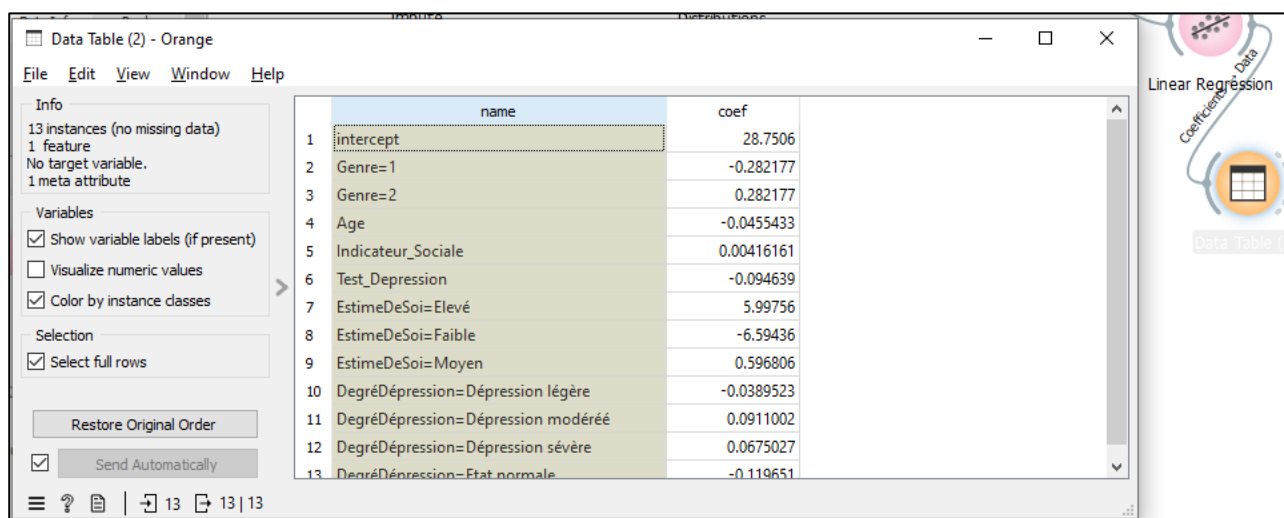
Name	tribut	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
Age		16.44	16	16	0.09	13	22	0 (0 %)
Indicateur_Soci...		11.5631	12.7857	11	0.289688	2.16667	18.5	0 (0 %)
Test_Estime_De...		30.75	30	31	0.12	18	38	0 (0 %)
Test_Depression		15.34	13	14	0.65	0	47	0 (0 %)
Genre			2		0.681			0 (0 %)
EstimeDeSoi			Elevé		0.697			0 (0 %)
DegréDépression			Dépression légère		1.3			0 (0 %)

Grace à ce résultat on va définir les caractéristiques de nos données (les variables importantes) : indicateur sociale – test d'estime de soi et test de dépression

4) Appliquer les techniques de fouille de données :

La régression :

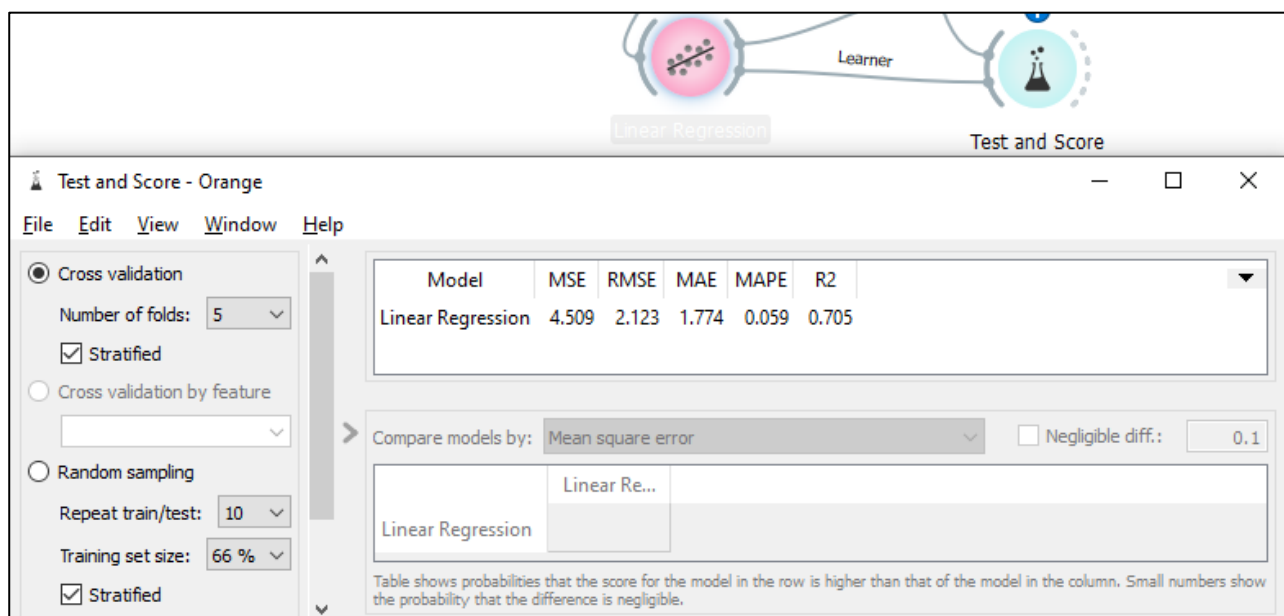
On va considérer le target : valeur de Test_Estime_de_soi :



The screenshot shows the 'Data Table (2) - Orange' window. The 'Info' panel on the left indicates 13 instances, 1 feature, and no target variable. The main table lists regression coefficients for various features. The 'name' column lists features, and the 'coef' column shows their corresponding coefficients.

	name	coef
1	intercept	28.7506
2	Genre=1	-0.282177
3	Genre=2	0.282177
4	Age	-0.0455433
5	Indicateur_Sociale	0.00416161
6	Test_Depression	-0.094639
7	EstimeDeSoi=Elevé	5.99756
8	EstimeDeSoi=Faible	-6.59436
9	EstimeDeSoi=Moyen	0.596806
10	DegréDépression=Dépression légère	-0.0389523
11	DegréDépression=Dépression modérée	0.0911002
12	DegréDépression=Dépression sévère	0.0675027
13	DegréDépression=Etat normale	-0.119651

Alors on remarque que le coefficient de la régression linéaire n'est nul, donc on peut constater qu'il y a une relation significative entre les variables indépendantes et les variables dépendantes.



The screenshot shows the 'Test and Score - Orange' window. The 'Cross validation' section on the left is configured with 5 folds, stratified sampling, and a training set size of 66%. The main table displays performance metrics for the 'Linear Regression' model.

Model	MSE	RMSE	MAE	MAPE	R2
Linear Regression	4.509	2.123	1.774	0.059	0.705

Below the table, the 'Compare models by' section is set to 'Mean square error' with a 'Negligible diff.' threshold of 0.1. A comparison table shows the probability that the score for the model in the row is higher than that of the model in the column.

Linear Regression	Linear Re...
Linear Regression	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

MSE de 4.5, ce qui indique l'ampleur moyenne des erreurs quadratiques.

Votre **RMSE** est de 2.123, fournissant une mesure de l'erreur qui peut être interprétée dans les mêmes unités que la variable dépendante.

La **MAE** est de 1.77, représentant l'erreur moyenne absolue.

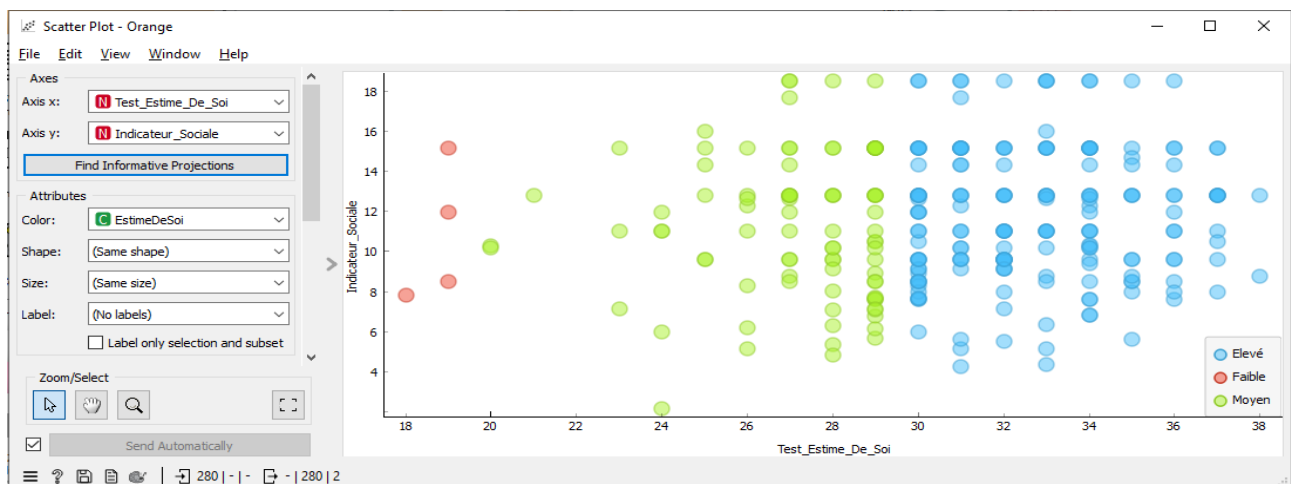
Le **MAPE** est de 0.059, indiquant l'erreur en pourcentage moyen.

Nous avons un **R²** de 0.705, suggérant que le modèle explique environ 70.5% de la variance totale. Cela signifie que le modèle est exploitable à 70%

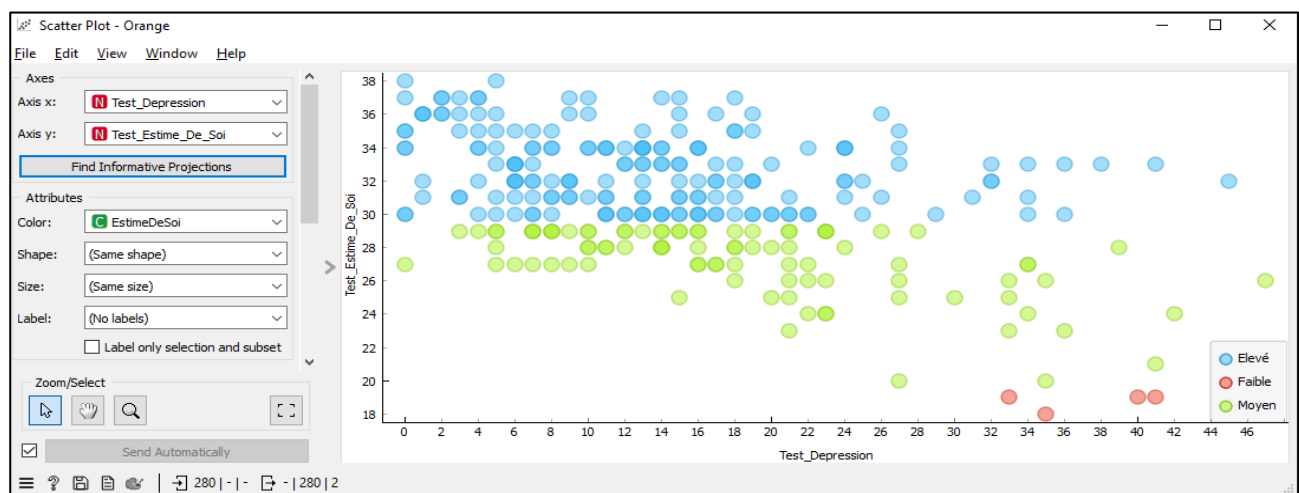
La classification :

Pour l'estime de soi on a trois classes

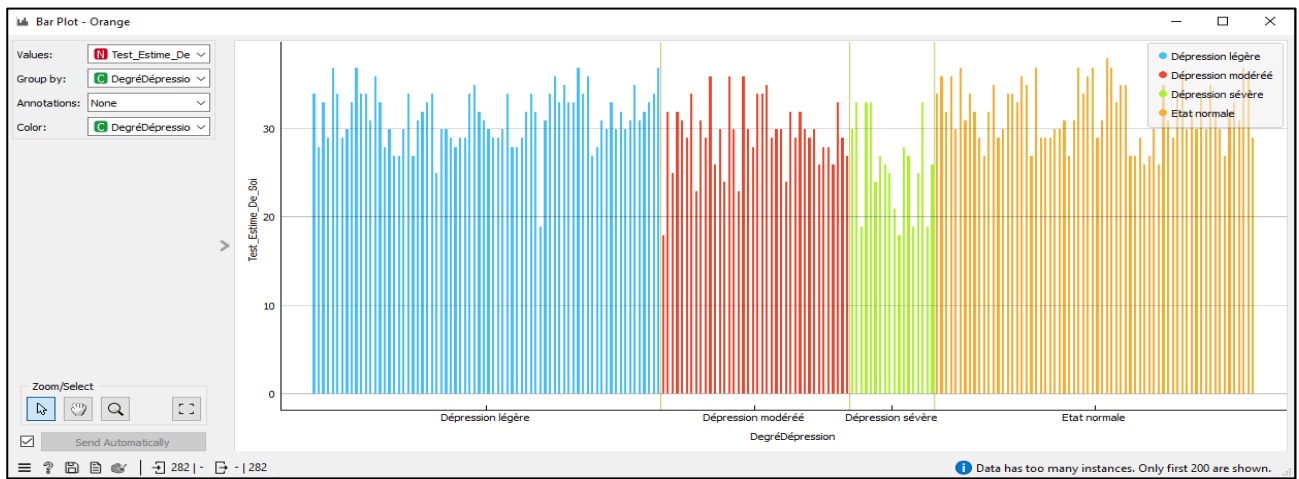
****Par rapport à l'indicateur sociale**



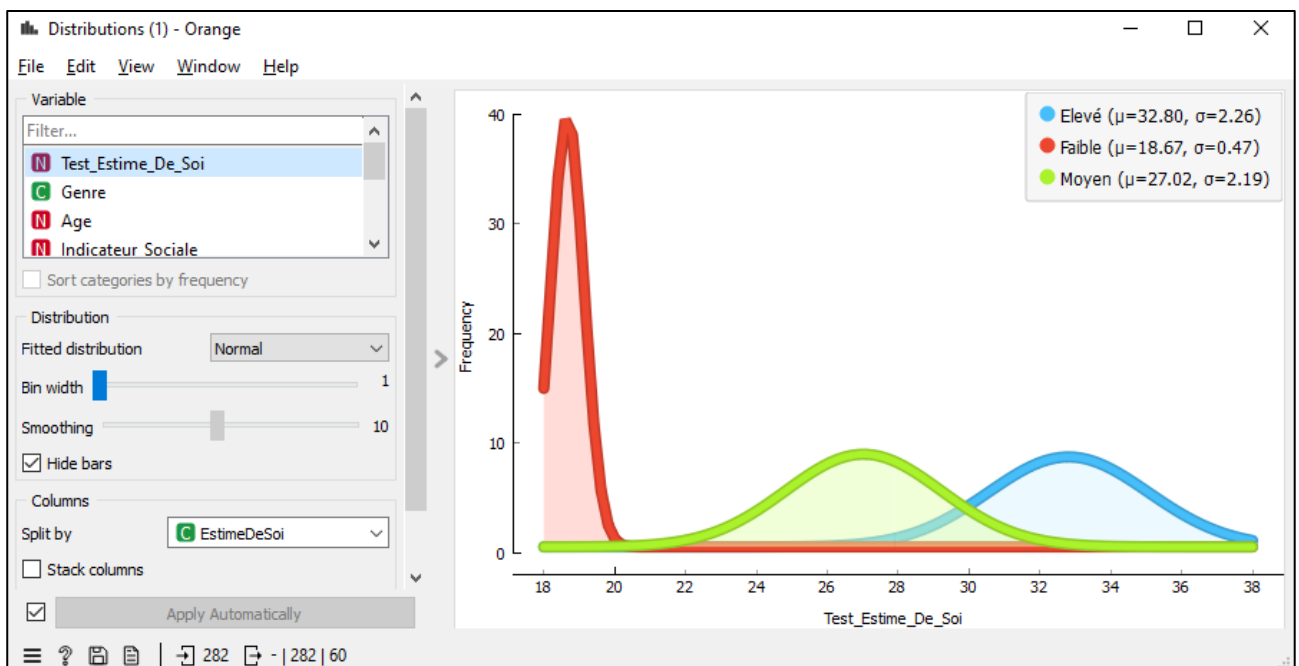
****Par rapport au test dépression**



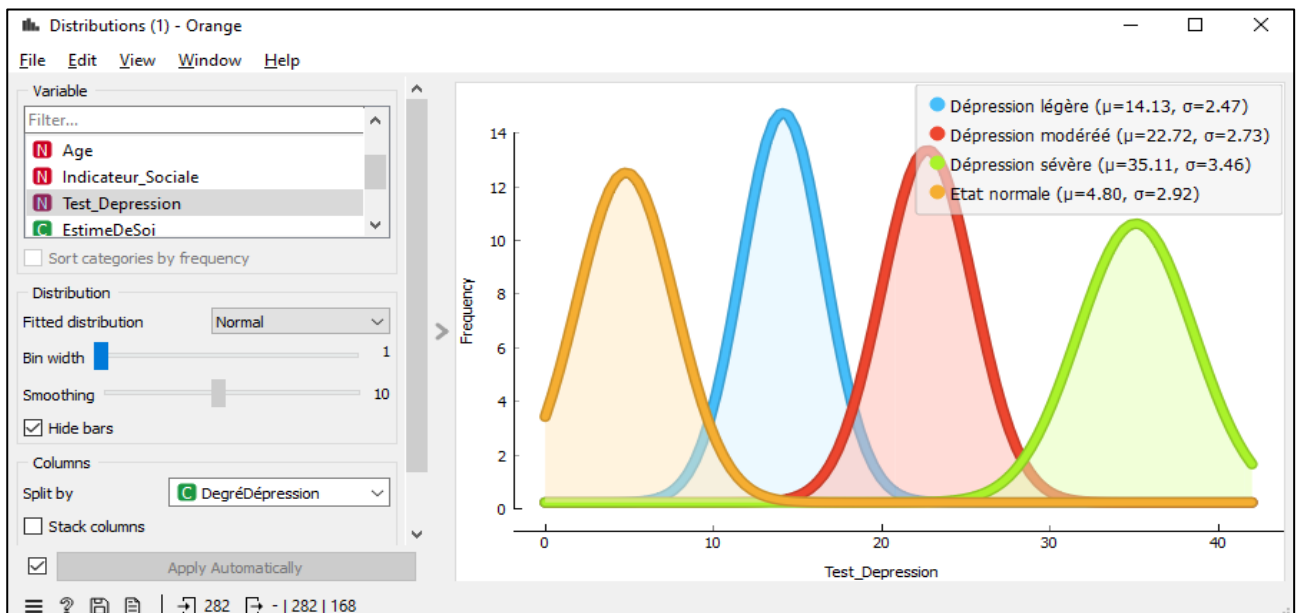
Pour le degré de la dépression on a quatre classes :



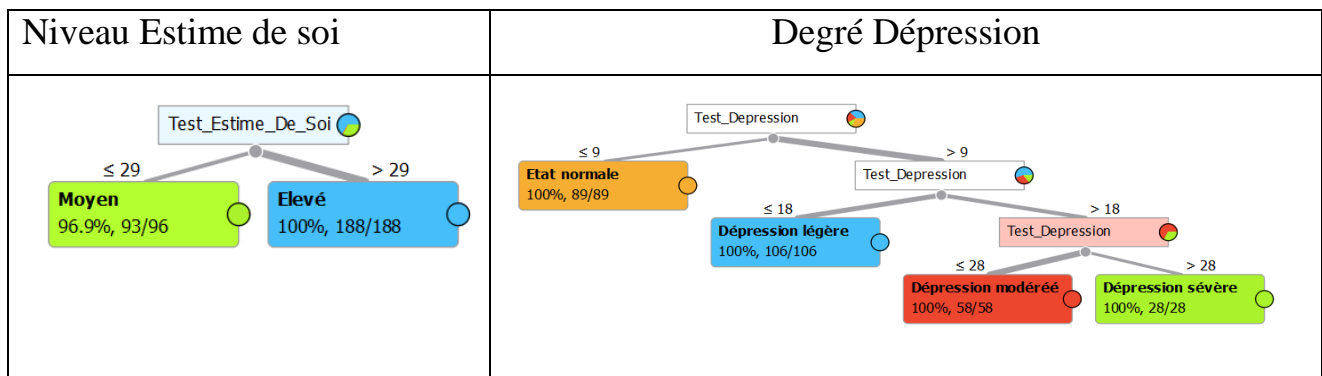
Distribution des données selon les calasses d'estime de soi :



Distribution des données selon les calasses de dépression :



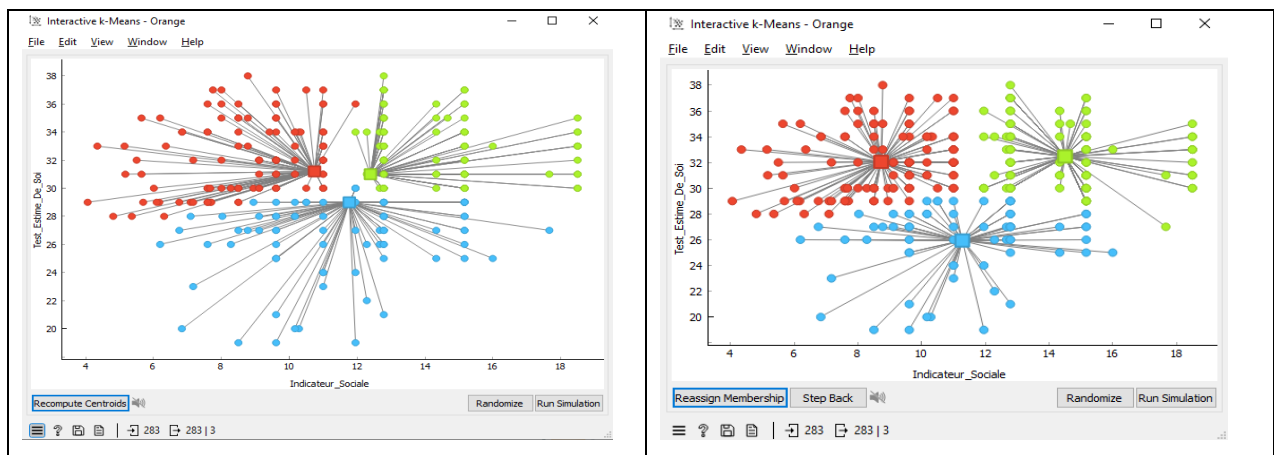
Arbre de décision :



Clustering k-means

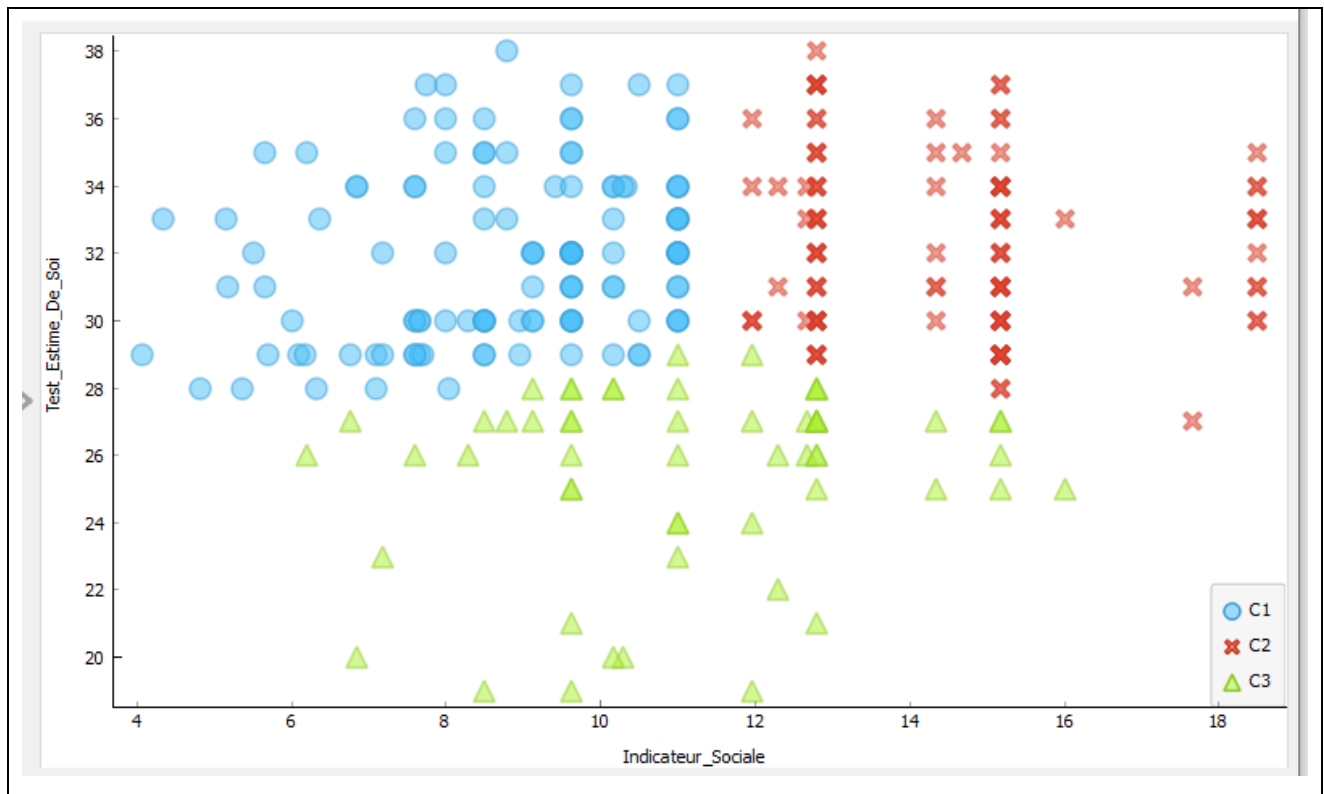
1) Indicateur sociale et teste d'estime de soi :

** On fixe le nombre des clusters en 3 :



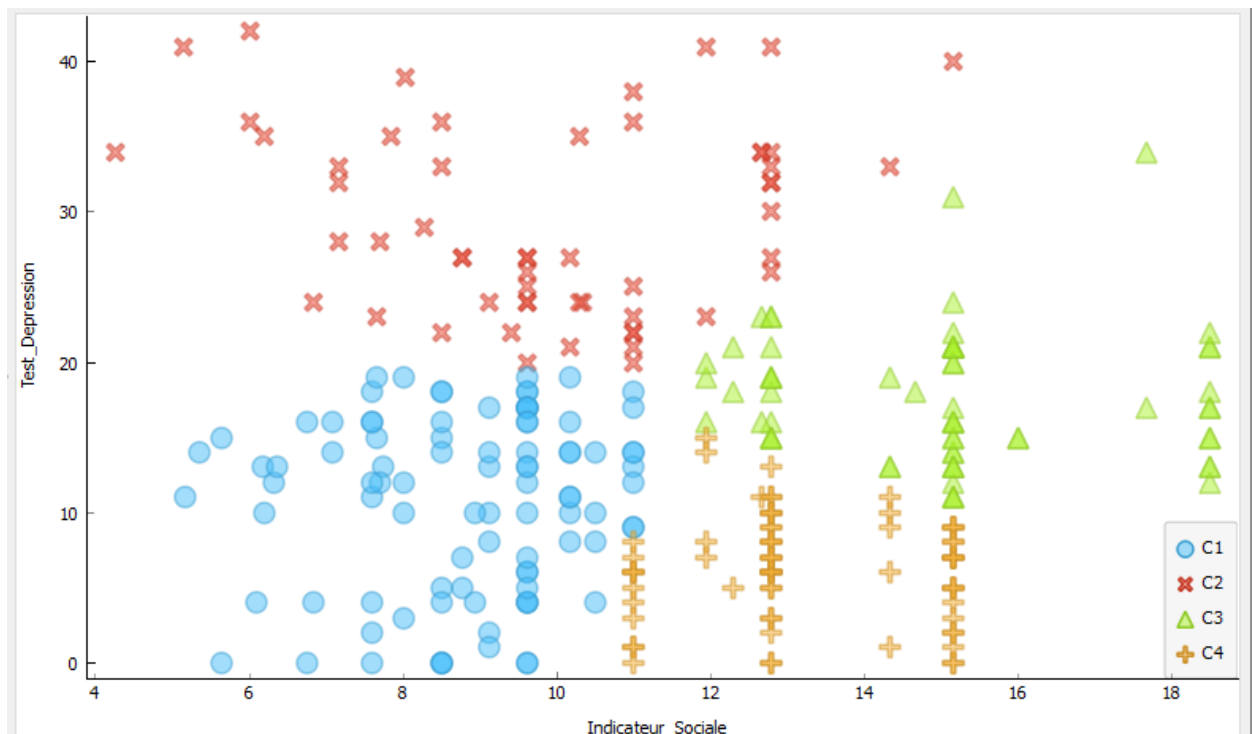
** Dans ce cas on utilisera l'outil K-means pour estimer le nombre optimal de clusters dans les données en entrée :

1	Genre	C	categorical	skip
2	Age	N	numeric	skip
3	Indicateur_Sociale	N	numeric	feature
4	Test_Estime_De_Soi	N	numeric	feature
5	Test_Depression	N	numeric	skip
6	EstimeDeSoi	C	categorical	skip
7	DegréDépression	C	categorical	skip

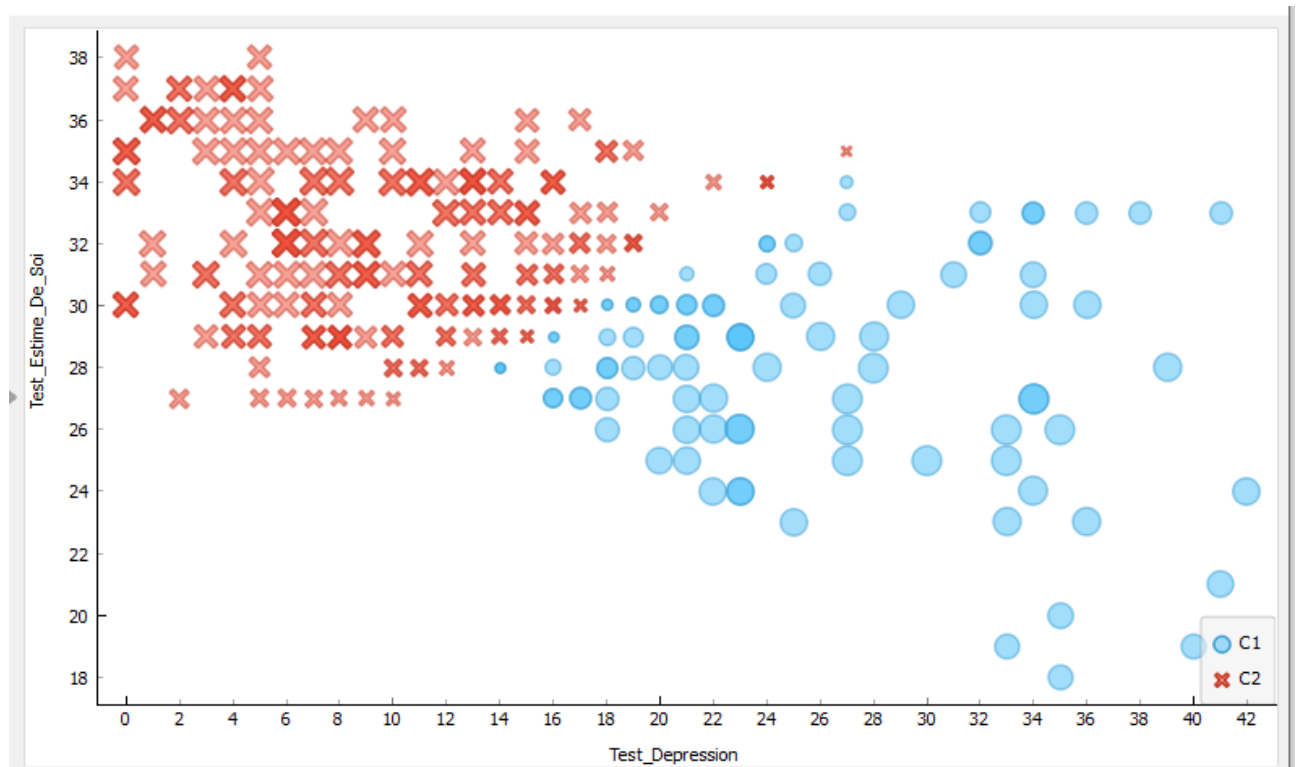


Alors il a détecté qu'il y a effectivement 3 clusters à partir, seulement des données de l'indicateur sociale et les valeurs de test d'estime de soi.

2) Indicateur sociale et test de dépression :



3) Test estime de soi et test de dépression :



5) Interprétation du module et établissement des conclusions :

Régression logistique :

Les métriques de la performance du modèle :

Pour l'indication de classe Niveau d'estime de soi à partir de : Genre, Age, Indicateur sociale et test d'estime de soi :

Evaluation results for target (None, show average over classes) ▾						
Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.999	0.989	0.984	0.979	0.989	0.977

AUC = 0.99 indique une excellente capacité discriminative de distinguer entre les classes ;

CA : 0.98 indique que le modèle a correctement classé 98% des échantillons ;

Le F1 Score est une mesure qui combine à la fois la précision (Precision) et le rappel (Recall) du modèle. Une valeur de 0.98 indique un équilibre entre la précision et le rappel, ce qui est souvent souhaitable ;

Précision = 0.97 indique une haute précision dans la prédiction de la classe positive ;

Recall = 0.98 indique que le modèle capture une grande partie des vrais positifs ;

MCC de 0.977 indique une forte corrélation entre les prédictions du modèle et les véritables étiquettes de classe.

Matrice de confusion :

		Predicted			Σ
		Elevé	Faible	Moyen	
Actual	Elevé	188	0	0	188
	Faible	0	0	3	3
	Moyen	0	0	93	93
Σ		188	0	96	284

Interprétation des résultats :

Concernant le niveau élevé et le niveau moyen la prédiction est exacte à 100%

Pour le niveau faible, il a classé les trois éléments dans la classe Moyen.

Etablir conclusions :

- 1) Dans la relation entre l'estime de soi et le degré de la dépression on a une corrélation négative c'est-à-dire que lorsque l'estime de soi est élevé le degré de dépression devienne faible à l'inverse ;
- 2) L'indicateur social est corrélé positivement avec l'estime de soi, mais il est corrélé négativement avec le degré de dépression

6) Gérer la connaissance découverte :

Construire un modèle :

La classification en utilisant SVM :

Sur python vous allez trouver le code dans les fichiers joints :

Prédiction de niveau d'estime de soi :

Prédiction Estime De Soi

Age:



Indicateur Sociale:



Test Estime De Soi:



Faire des prédictions

Prédiction Estime De Soi: Moyen

Pour le degré de la dépression :

Informations sur le nouveau fichier Excel ' N° Genre Age Indicateur_Sociale Test_Depression DegreDepression

0 1 1 15 15.166667 7 Etat normale 1 2 1 16 7.700000 28 Dépression modéréé 2 3 1 15 14.333333 1 Etat normale 3 4 2 15 11.000000 14 Dépression légère 4 5 2 15 18.500000 8 Etat normale 269 271 2 17 7.166667 33 Dépression sévère 270 272 2 17 12.785714 34 Dépression sévère 271 273 2 16 8.000000 19 Dépression modéréé 272 274 1 17 6.757576 16 Dépression légère 273 275 2 17 12.785714 8 Etat normale

[274 rows x 6 columns]' :

Aperçu des premières lignes du fichier Excel :

	N°	Genre	Age	Indicateur_Sociale	Test_Depression	DegreDepression
0	1	1	15	15.1667	7	Etat normale
1	2	1	16	7.7	28	Dépression modéréé
2	3	1	15	14.3333	1	Etat normale
3	4	2	15	11	14	Dépression légère
4	5	2	15	18.5	8	Etat normale

Nombre total de lignes : 274

Nombre total de colonnes : 6

Prédiction :

Prédiction Degré de dépression

Age:



Indicateur Sociale:



Test Depression:



Faire des prédictions

Prédiction Degré Dépression: Dépression modérée

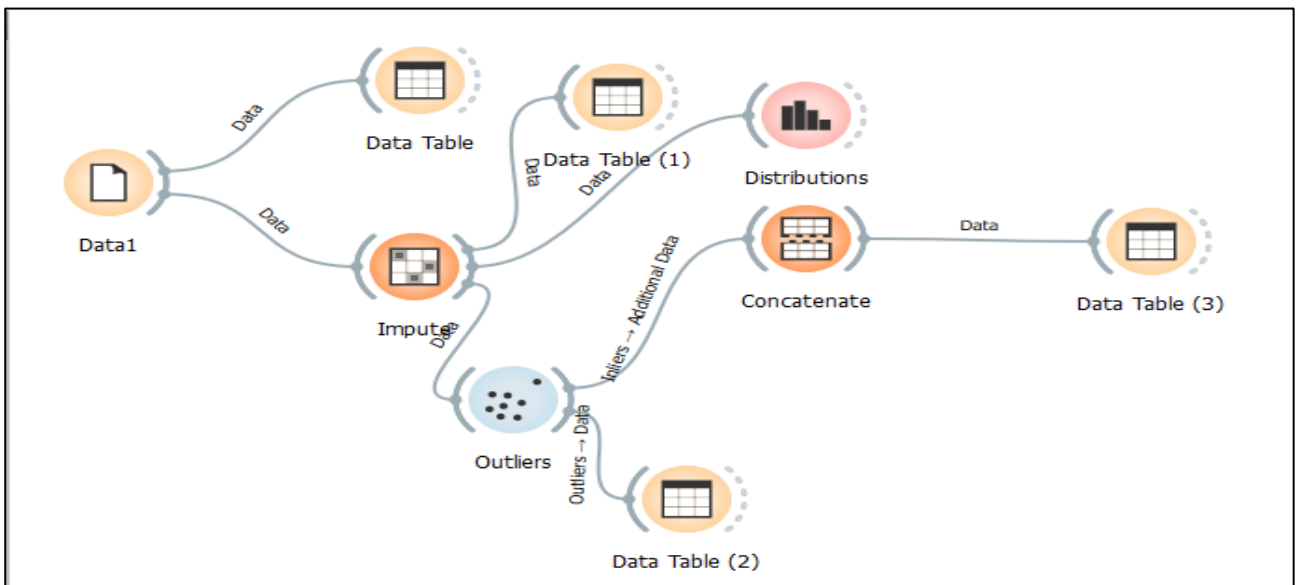
Conclusion générale :

À travers ce travail pratique, nous aspirons à enrichir nos compétences en Data Mining et Machine Learning en les appliquant à des problématiques pertinentes de la santé mentale, soulignant ainsi le rôle essentiel de ces disciplines dans l'avancement de la recherche et de la compréhension dans des domaines aussi cruciaux que la psychologie.

Un grand merci à notre professeur Monsieur Mohamed Sabiri.

Annexes

Annexe (1) : prétraitement de la première base de données + nettoyage



Annexe (2) : prétraitement + nettoyage+ transformation + application des techniques de Data Mining sur la nouvelle base de données obtenue après la réduction des dimensions :

