**Remerciement :**

Nous profitons par le biais de ce rapport pour exprimer nos vifs remerciements à toute personne contribuant de près ou de loin à l’élaboration de ce travail.

Nous tenons à remercier vivement tous nos professeurs. Le Directeur de notre établissement Mr Mohammed HALIM, qui ont contribué à la réalisation de ce modeste projet, qui nous ont encadrés et aidés tout au long de notre parcours. Ainsi nous remercions l’administration de l’École Nationale Supérieure d’Informatique et d’Analyse des Systèmes qui à travers leur programme nous ont fourni des outils de qualité facilitant notre spécialisation.

Un merci bien particulier adressé également à Mme. Houda BENBRAHIM notre professeur de Data mining, pour ces remarques, ses directives, et l’intérêt qu’elle porte à ses étudiants. Nous tenons à lui exprimer nos sincères remerciements pour son suivi et ses orientations.

**Introduction :**

Dans notre école l’ENSIAS, on travaille plusieurs projets dans des groupes de deux à six personnes de façon collaborative et responsable. Le travail de groupe développe bien sûr les compétences sociales, mais il poursuit aussi l’objectif d’intensifier l’apprentissage disciplinaire. Pour cela, nous devons bien choisir notre groupe. Les enseignants laissent aux élèves la liberté de former leurs groupes. En général, les groupes sont formés en basant sur l’amitié qui relie les élèves et non pas leurs compétences.

Le but de ce projet est d’aider les élèves à former des groupes en exploitant intelligemment leurs relevés des notes de la première et la deuxième année et extraire des groupes de personnes ayant plus ou moins les mêmes notes dans les différentes matières. Donc nous obtenons des groupes homogènes de personnes qui peuvent exploiter leurs points fort communs pour réaliser un bon projet.

Dans ce projet on va suivre la méthode CRISP-DM qui signifie « Cross-industry standard process for data mining ». Chaque chapitre va être une étape de cette méthode. Donc nous aurons six chapitres dans ce rapport qui sont : la compréhension métier, la compréhension des données, la préparation des données, la modélisation, l’évaluation et enfin le déploiement.

**Chapitre 1 : La compréhension métier**

Cette première phase consiste à bien comprendre les éléments métiers et problématique qu’on vise à résoudre. On va commencer par determiner les objectifs objectifs stratégiques et opérationneles. Puis, on va évaluer la situatuin actuelle. Finallement, vient la traduction de l’objectif stratégique en concepts de Data mining.

1. **Détermination des objectifs stratégiques et opérationnelles**

Le problème qu’on rencontre dans le travail de groupe dans notre école l’ENSIAS c’est la non homogénéité des groupes. Le seul facteur de la formation de ces groupes c’est la relation entre les membres. Ainsi les outils techniques peuvent ne pas plaire tous les membres d’un groupe. D’où vient l’idée de notre projet, former des groupes homogènes des étudiants ayant les mêmes compétences pour qu’ils puissent travailler sur un projet qui exploite leurs compétences communes. Tous cela, en annalysant leurs résultats et chercher les élèves ayant des résultats similaires dans la même fillière.

1. **Analyse de la situation actuelle**

Actuellement, nous disposons de quelques relevés de notes de la deuxième année de quelques filières : la filière e-Management et Business Intelligence (eMBI), la filière Génie Logiciel (GL) et les notes de la première année. Tous ces relevés de notes sont de la promotion de l’année 2021. Nous devons encore collecter les relevés de notes d’autres filières : la filière Ingénierie e-Logistique (IeL), la filière Ingénierie des Systèmes Embarqués et Mobiles (ISEM) et et la filière Sécurité des Systèmes d'Information (SSI).

Parmi les problèmes que nous avons rencontrés c’est la sensibilité des données. On ne peut pas diffuser telles données personnelles, donc on doit cacher les noms et les prénoms des étudiants pour protéger leurs confidentialités.

1. **Détermination des objectifs du Data Mining**

Puisque l’objectif stratégique est clairement défini, il convient maintenant de le traduire en concepts de Data Mining. On va opter à une technique descriptive car nous sommes en train de présenter une information cachée par le volume des données. Et plus spécifiquement on va travailler un problème de clustering qui sert à regrouper des données non étiquetées présentant des propriétés similaires. Dans notre cas les proprités sont les notes de chaque module.

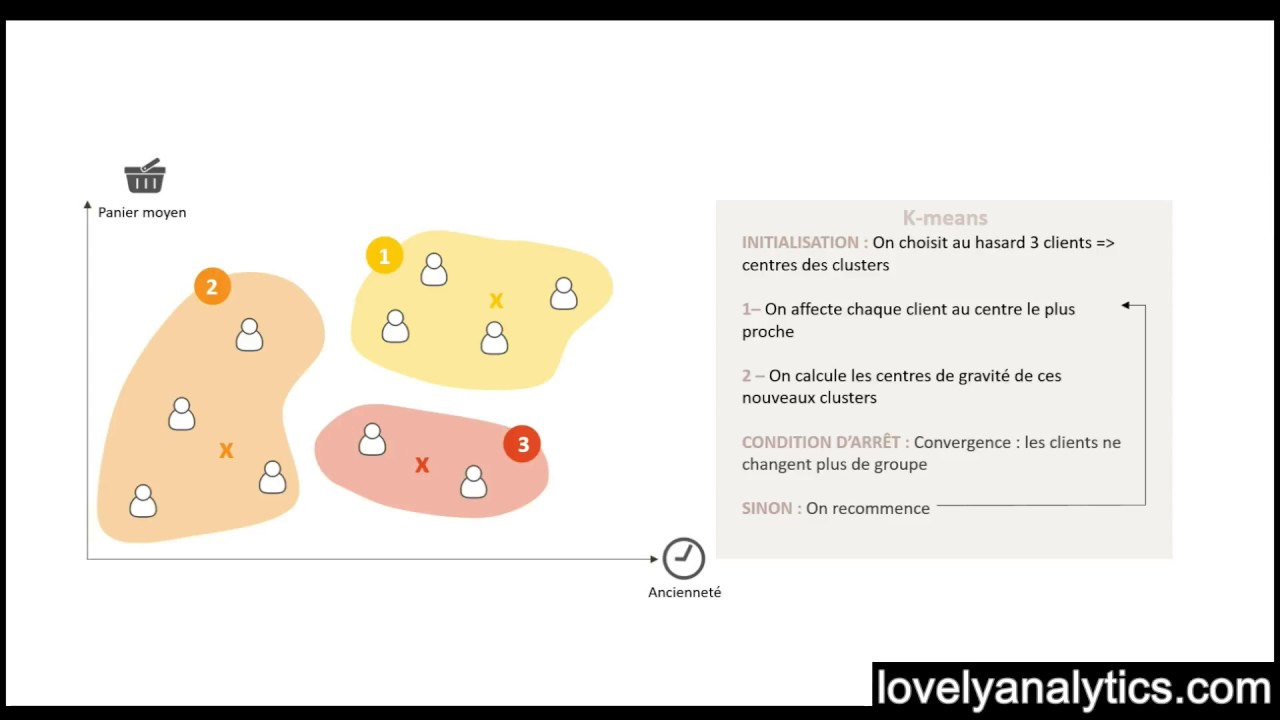
* 1. **Clustering**

Le Clustering ou le partionnement des données est une méthode de classification non supervisée rassemble un ensemble d’algorithmes d’apprentissage pour relever des sous-ensembles des données difficiles à identifier à l’œil nu.

Pour atteindre ce but, cette technique vise à maximiser l’inertie entre les sous-ensembles pour bien les différencier et minimiser celle au sein du même ensemble pour qu’il soit plus homogène. Il existe plusieurs méthodes de clustering : méthodes hiérarchiques, méthodes de partitionnement et méthode mixtes.

* 1. **K-means**

Nous avons choisis travailler avec l’algorithme de K-means c’est un algorithme de partionnement qui permet de regrouper les observation de notre data set en K clusters distincts. L’algorithme de K-means est comme suit :



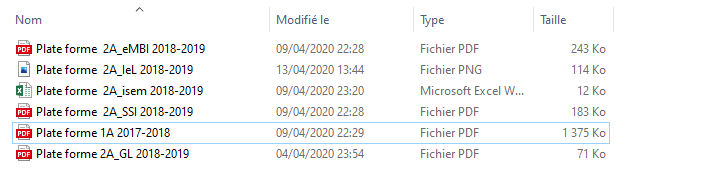
**Figure 1 : Principe algorithmique de K-means**

**Chapitre 2 : La compréhension des données**

La phase de compréhension des données de CRISP-DM implique l’étude des données disponibles pour le Data mining. On doit tout d’abord collecter les données puis faire la description de ces données. La troisième étape est l’exploration des données et comme étape finale vérifier la qualité des données collectées.

1. **Collecte des données**

Nous avons pu collecter les notes de toute la promotion 2020 dans toutes les fillière auprès des élèves de cette promotion. Les notes de la première année qui est une année commune à toutes les filières. Puis les notes des élèves de chaque filière : eMBI, GL, IeL, SSI, et ISEM en deuxième année.

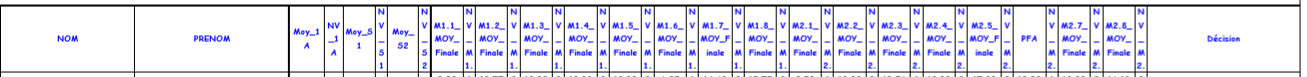


**Figure 2 : Les données collectées**

Nos fichiers sont de différents formats : des fichiers pdf, un fichier png et un fichier Microsoft Excel.

1. **Description des données**

Chaque fichier contient le nom et le prénom de l’étudiant, le moyen dans chaque semestre et chaque module ainsi la décision finale.



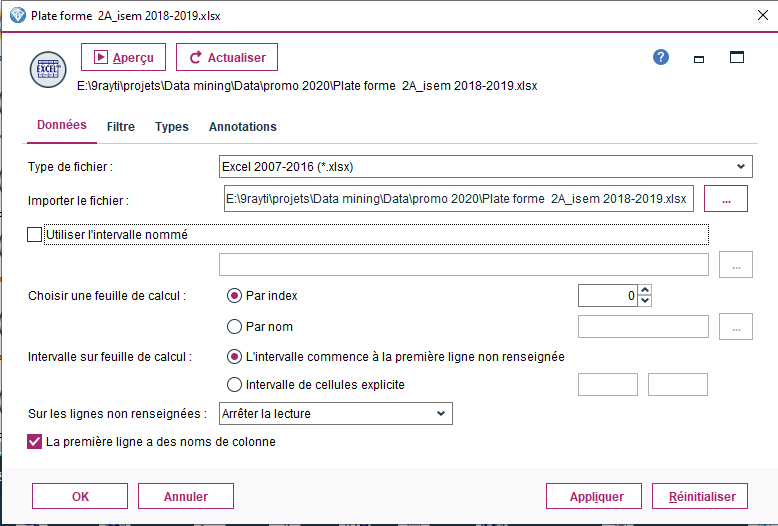
**Figure 3 : les composants de chaque fichier**

|  |  |  |  |
| --- | --- | --- | --- |
| Code | Signification | Type | longueur |
| NOM | Le nom de l’étudiant | Texte | 32 |
| PRENOM | Le prénom de l’étudiant | Texte | 32 |
| Moy\_1A | Le moyen de la première année | Numérique | 5 |
| Moy\_S1 | Le moyen de la première semestre de la première année | Numérique | 5 |
| Moy\_S2 | Le moyen de la deuxième semestre de la première année | Numérique | 5 |
| NV\_1A | Le nombre des modules non validé dans la première année | Numérique | 2 |
| NV\_S1 | Le nombre des modules non validé dans la première semestre de la première année | Numérique | 1 |
| NV\_S2 | Le nombre des modules non validé dans la deuxième semestre de la première année | Numérique | 1 |
| Mx.y\_MOY\_Finale | Le moyen du yème module dans la xème semestre | Numérique | 5 |
| NV\_Mx.y | Le module y de la xème semestre est-il valide ?   * 0 : si le module est validé (Mx.y\_MOY\_Finale >=12) * 1 : si le module n’est pas validé | Numérique | 5 |
| PFA | La note du projet de la fin de l’année | Numérique | 5 |
| Décision | La décision du conseil il prend 6 valeurs :   * Félicitation si Moy\_1A>=15 * Encouragement si 15<Moy\_1A<=14 * Admis si 14<Moy\_1A<=12 et NV\_1A<=4 * Admis avec indulgence si 12<Moy\_1A<=min\_moy ou NV\_1A>4 * Ajourné si min\_moy<Moy\_1A (min\_moy est décidé par le conseil) * Réorienté si l’étudiant n’était pas admis pendant deux ans | Texte | 32 |

**Tableau 1 : Dictionnaire des données**

1. **Exploration des données**

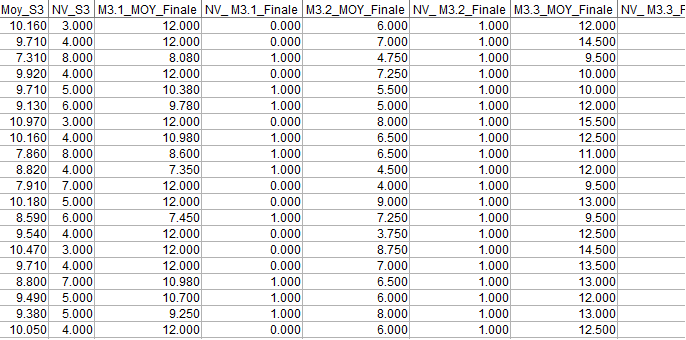
Pour explorer les données il faut tout d’abord transformer nos fichiers en un format adopter par le Spss modeler. Nous avons trois formats : des fichiers excel adopter par le Spss modeler et des fichiers PDF et une image. Pour les fichiers PDF nous avons pu les convertir en fichiers pdf facilement. Par contre la conversion de l’image a été faite manuellement puisque sa qualité était mauvaise.



**Figure 4 : Création des nœuds sources**

Nous avons placé des nœuds sources excel et nous avons édité les nœuds en important les fichiers excel qui contiennent les données.

* 1. **Exploradion à l’aide des tableau**

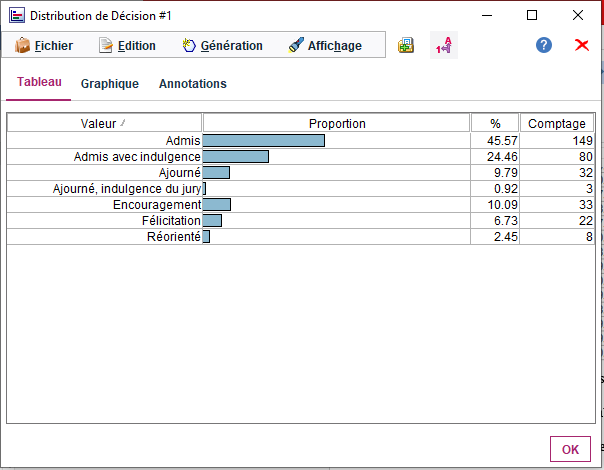
**Figure 5 : Explorer les données à l’aide des tableaux**

Nous avons connecté un nœud tableau avec chaque nœud source pour visualiser les données.

* 1. **Visualisation du pourcentage de réussite**

Il sera utile de connaitre le pourcentage des gens qui ont réussi avec félicitation et le pourcentage des admis et aussi les autres catégories donc nous avons choisi d’explorer cette donnée.



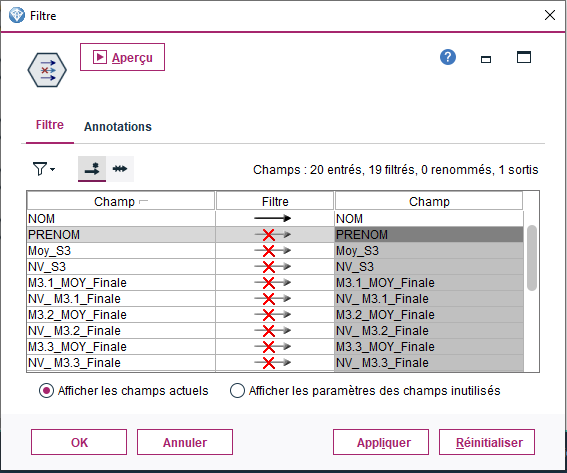


**Figure 6 : Pourcentage de réussite**

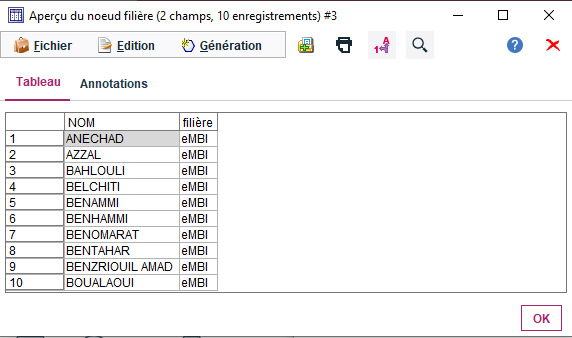
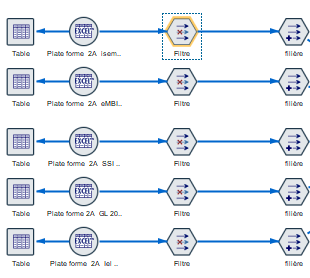
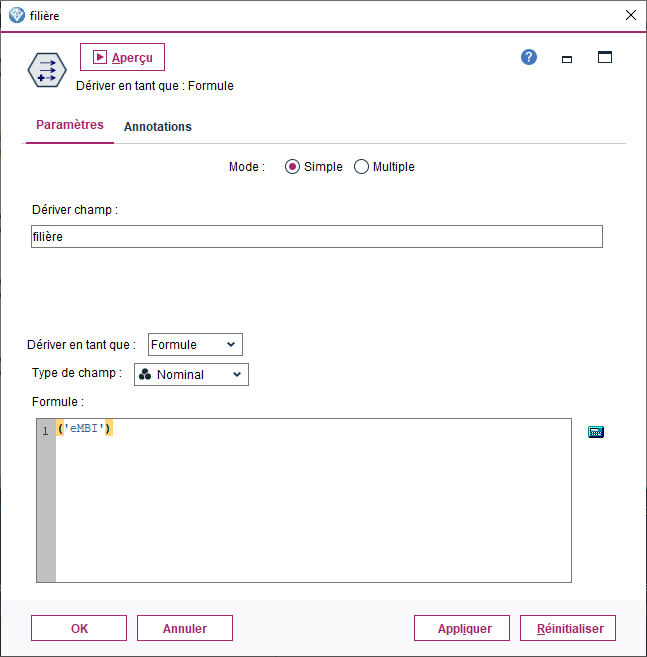
D’après le résultat du nœud Distribution, la plupart des étudiants sont des admis. Ainsi nous avons constaté que nous avons seulement 22 étudiants qui réussirent avec félicitation , c’est-à-dire ils ont une note supérieure à 15.

* 1. **Visualisation du pourcentage des filières**

Il sera utile aussi de visualiser le pourcentage et le nombre des étudiants dans chaque filière. Pour savoir le nombre des groupes quand on aura. Pour cela nous avons ajouté un champ qui contient la filière de chaque étudiant en utilisant le nœud « calculer ». Mais avant nous allons extraire le champ « NOM », puisque dans le fichier des étudiants « GL » nous disposons seulement des noms des étudiants, avec le nœud « filter ».

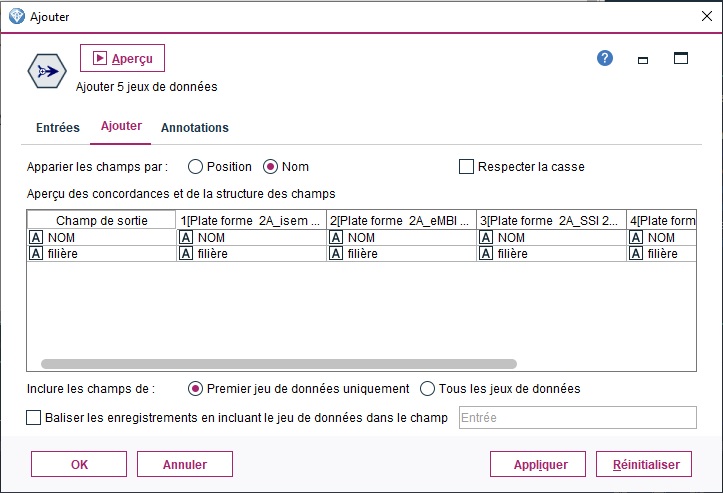
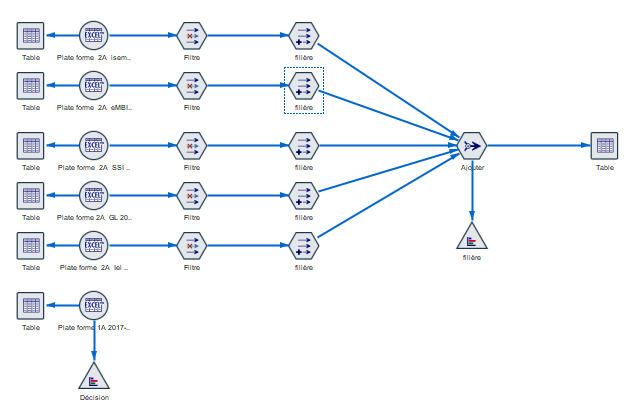
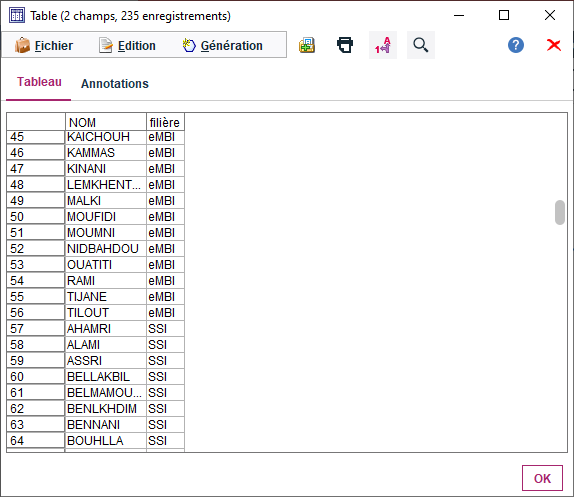


**Figure 7 : Extraire le champ des noms**

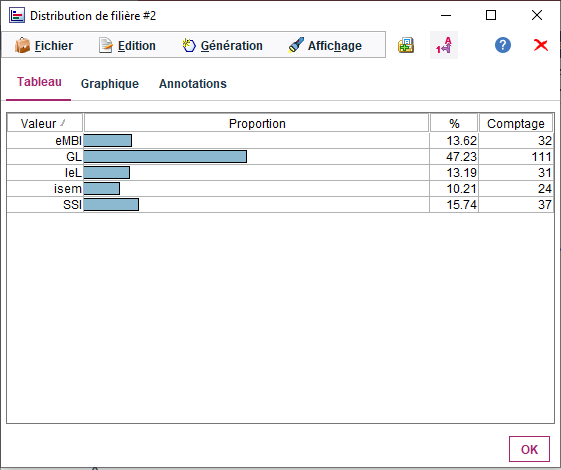


**Figure 8 : Ajouté la colonne de la filière**

Puis on va faire une union entre les nœuds résultants et le nœud qui contient les notes de la première année avec le nœud « ajouter ».



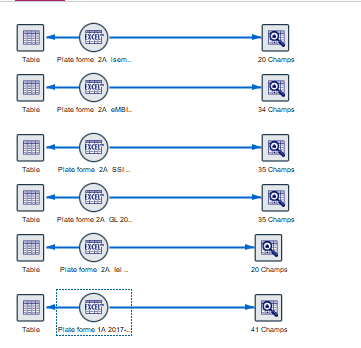
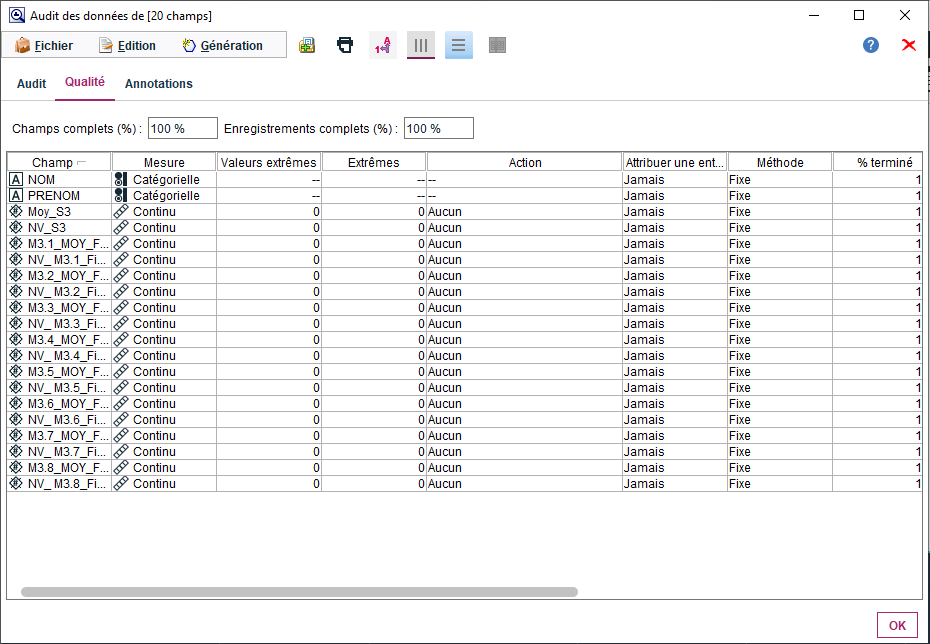
**Figure 9 : Union des nœuds**



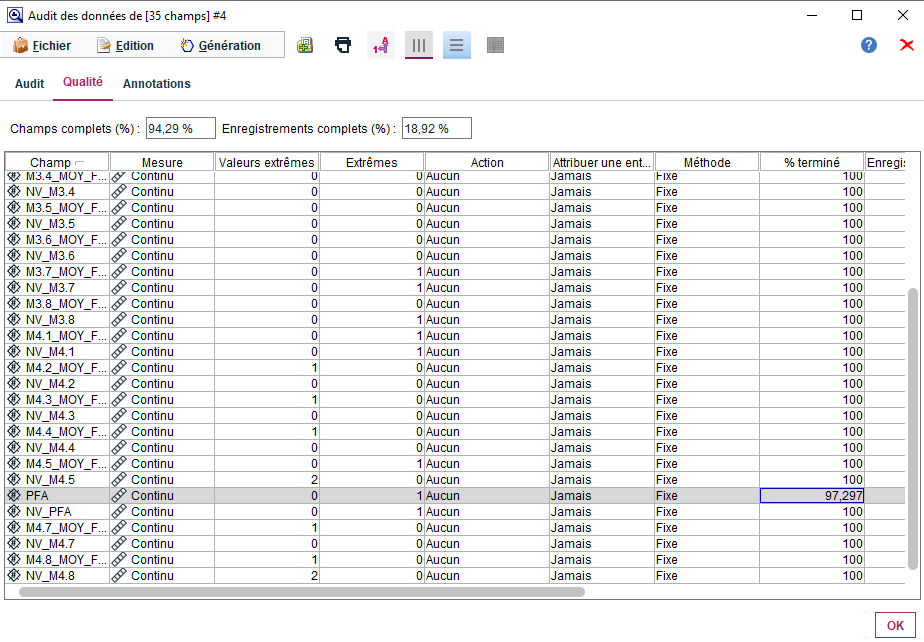
**Figure 10 : Pourcentage des filières**

Nous avons remarqué que l’effectif des étudiants en « GL » est le plus grand. Par contre l’effectif en « Isem » est le plus petit. Ainsi on peut constater que les filières « eMBI », « Isem » contient un nombre pair des élèves donc on n’aura que des binômes. Par contre dans les autres filières, il y a un nombre impair des étudiants donc il y aura des monômes.

1. **Vérification de la qualité des données**



**Figure 11 : Qualité des données**

Pour la qualité des données tous les fichiers ont des champs complets sauf le fichiers qui contient les notes de la filière SSI.

**Figure 12 : Qualité des données de nœud notes SSI**

Il y’ a un manque des valeurs dans le champ status (30 valeurs manquantes) ainsi dans le champ PFA il y a une seule valeur manquante.

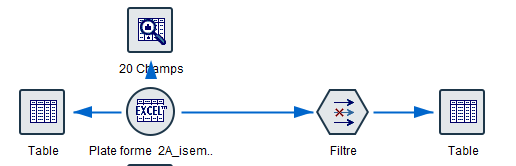
**Chapitre 3 : Préparation des données**

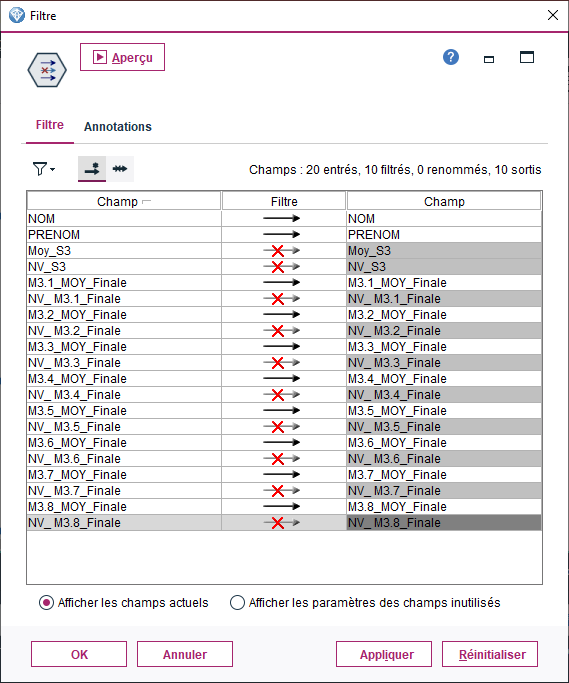
La préparation des données est l’un des aspects les plus importante et les plus coûteux en temps du Data Mining. Dans cette section, on va voir la sélection des données, le nettoyage des donnés et l’intégration des données.

1. **Sélection des données**

Après avoir effectué la collecte initiale de données on doit maintenant choisir les données pertinentes pour nos objectifs de Data Mining.

Les données que nous avons besoin sont : les noms et les prénoms des étudiants et leurs notes.



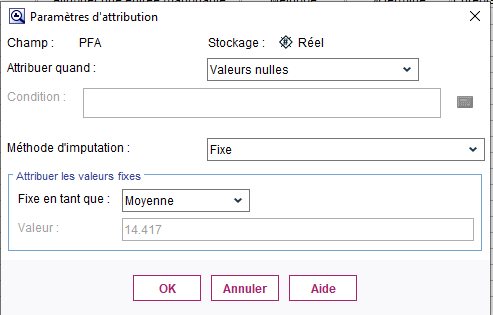


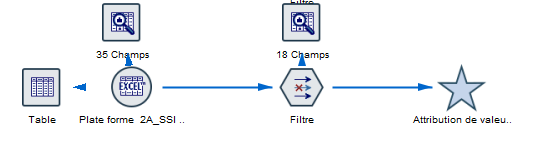
**Figure 13 : Sélection des données**

Pour tous les fichiers, nous avons sélectionné les noms, les prénoms et les notes des étudiants dans les huit modules et nous avons éliminé les autres attributs en utilisant le nœud « Filter ».

1. **Nettoyage des données**
   1. **Données manquantes**

Comme nous avons déjà montré dans le chapitre précédant il y a une seule valeur manquante. C’est la note du projet de la fin d’année (PFA) d’un étudiant dans la filière SSI. On va donc la remplacer avec la valeur moyenne.

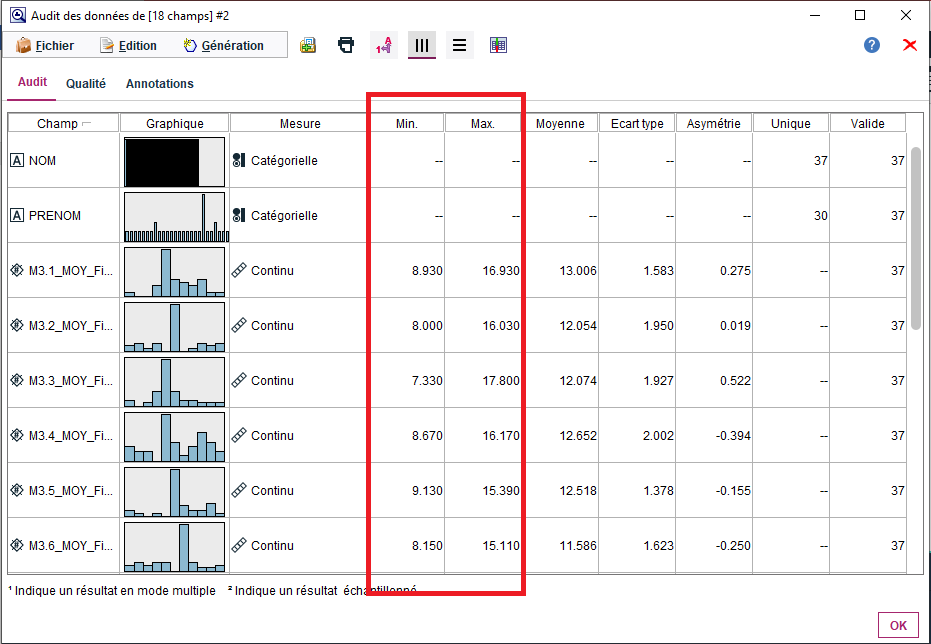


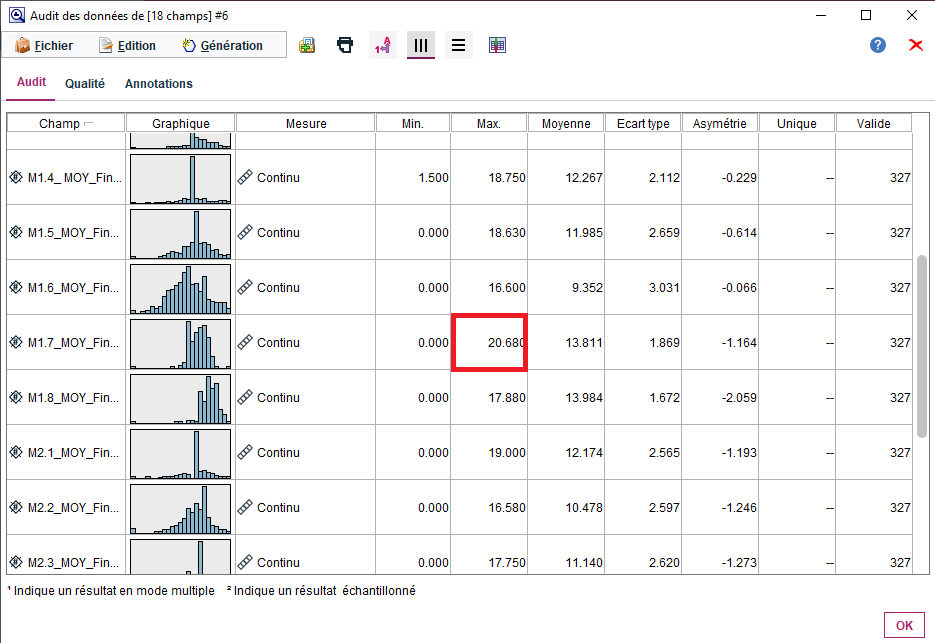


**Figure 14 : Insertion dune valeur manquante**

* 1. **Erreurs dans les données**

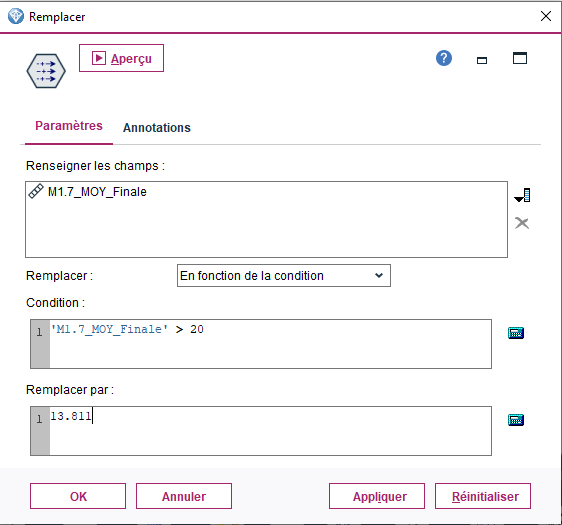
Nous avons vérifier manuellement les valeurs de telles manière que la valeur d’une note soit inclus entre 0 et 20.

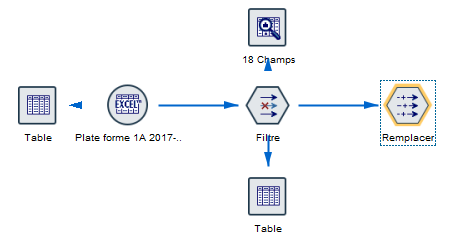
 **Figure 15 : vérification des données**

****

**Figure 16 : Détection d’une valeur erronée**

Nous avons trouvé une valeur anormale au niveau du module 1.7 dans les notes de la première année, la note est 20.68 ce qui est impossible.



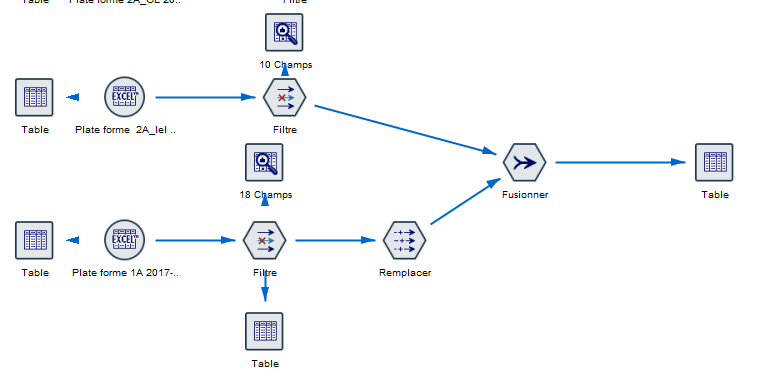


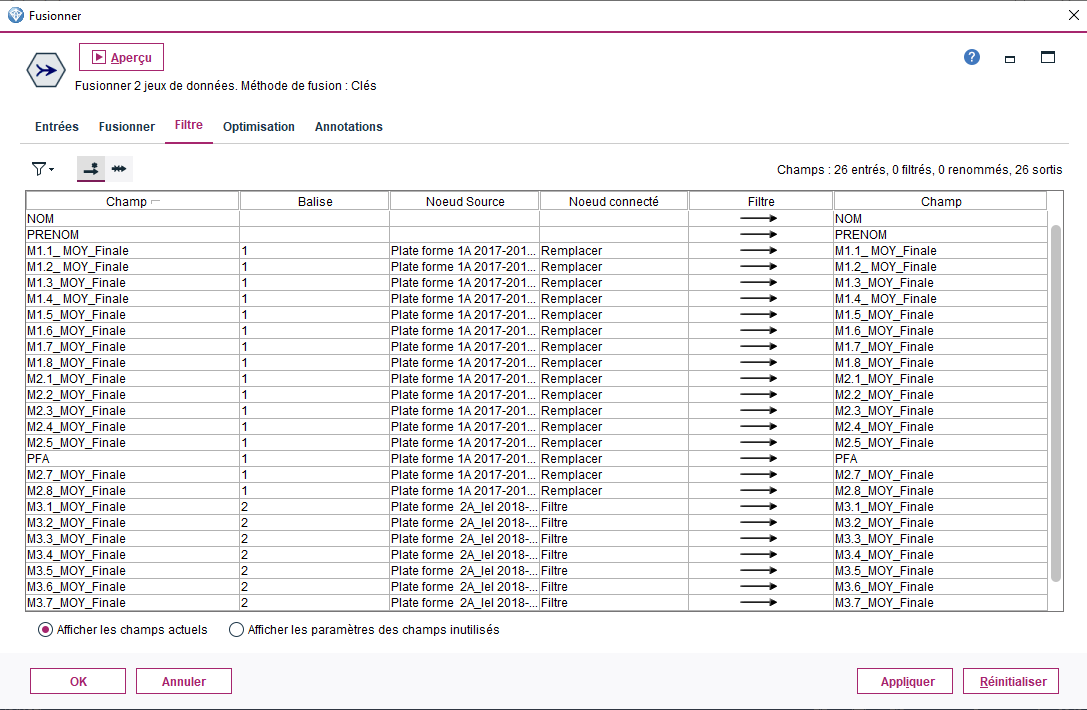
**Figure 17 : Remplacer la valeur erronée par la valeur moyenne**

On va remplacer cette valeur par la moyenne qui est 13.811.

1. **Intégration des données**

Il existe deux méthodes principales pour l’intégration des données : la fusion et l’ajout. Dans notre cas on va faire la fusion des données. Nous avons fusionné les ensembles de données de la première année avec les données de la deuxième année de chaque filière.





**Figure 18 : Fusionner les données de la première année et de la deuxième année de la filière Iel**

Par exemple dans la figure 18 on a la fusion des deux ensembles des données : les données de la première année et les données de la deuxième année des étudiants de la filière Iel.

**Chapitre 4 : Modélisation**

La quatrième étape est la modélisation qui est au cœur de tout projet d'apprentissage automatique. Cette étape est responsable des résultats qui devraient satisfaire ou aider à atteindre les objectifs du projet.

Bien que ce soit la partie glamour du projet, c'est aussi la plus courte dans le temps, car si tout ce qui précède est fait correctement, il y a peu à ajuster. Dans le cas où les résultats sont améliorables, la méthodologie est définie pour revenir à la préparation des données et améliorer les données disponibles.

Dans ce chapitre, on va voir la sélection des techniques de modélisation, la géneration d’une conception de test et la construction du modèle.

1. **Sélection des techniques de modélisation**

Il y’ a plusieurs types de modélisation. Le choix du modèle le plus adéquat sera généralement basé sur les critères suivants :

* **Les types de données disponibles pour l’exploration :**

Les champs inéressants sont numérique.

* **Les objectifs de data mining :**

Notre problème est un problème de classification non supervisé. Nous avons plusieurs méthodes de clustering : méthode hiérarchiques, méthodes de partionnement, méthodes mixtes et analyse floue.

Nous avons choisi de travailler avec l’algorithme de K-means : c’est une méthode hiérarchique. Il permet de regrouper en K clusters distincts les observations du data set.

Nous avons choisi K-means pour ces raisons :

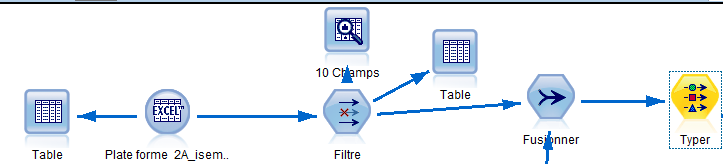
* Simple et robuste
* On définit nous même le nombre de clusters
* Efficace : O(n.t.k) avec n : le nombre des objets, t : nombre des itérations et k : le nombre des clusters.

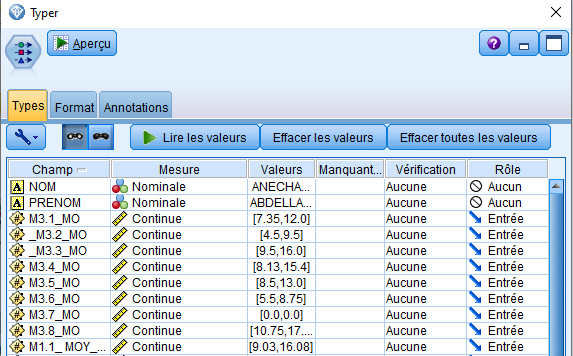
1. **Géneration d’une conception de test**

La dernière étape avant la création du modèle consiste à considérer la façon dont les resultats du modèle seront testés. Puisqu’on est dans le cas d’un modèle non supervisé le critère de « qualité d’ajustement » de notre modèle sera la taille des clusters. En effet nous voulons minimiser les groupes avec une seule personne. Ainsi il sera comme critère la facilité d’interpretation et le temps de traitement nécessaire.

1. **Construire le modèle**

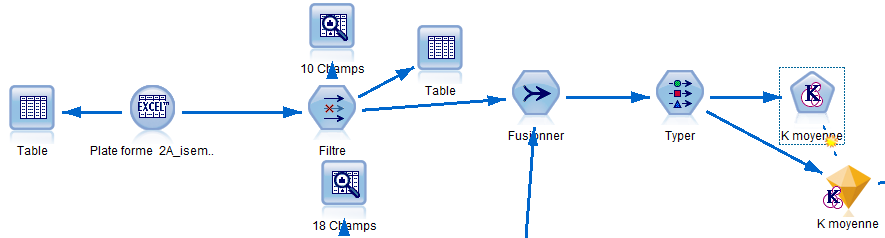
Nous avons commencer par ajouter le noeaud « typer » pour les définir les données entrants du modèle.

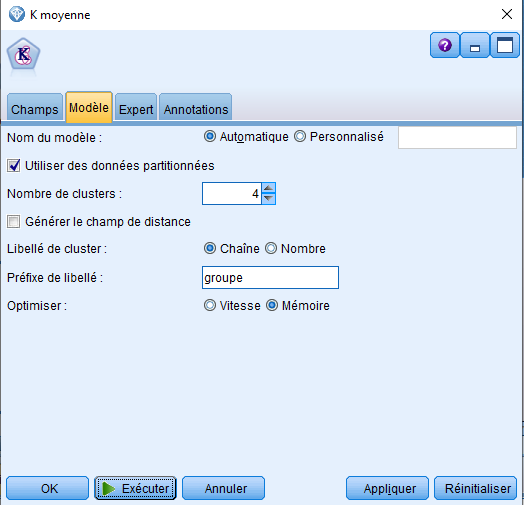




**Figure 19 : Définir les inputs du modèle**

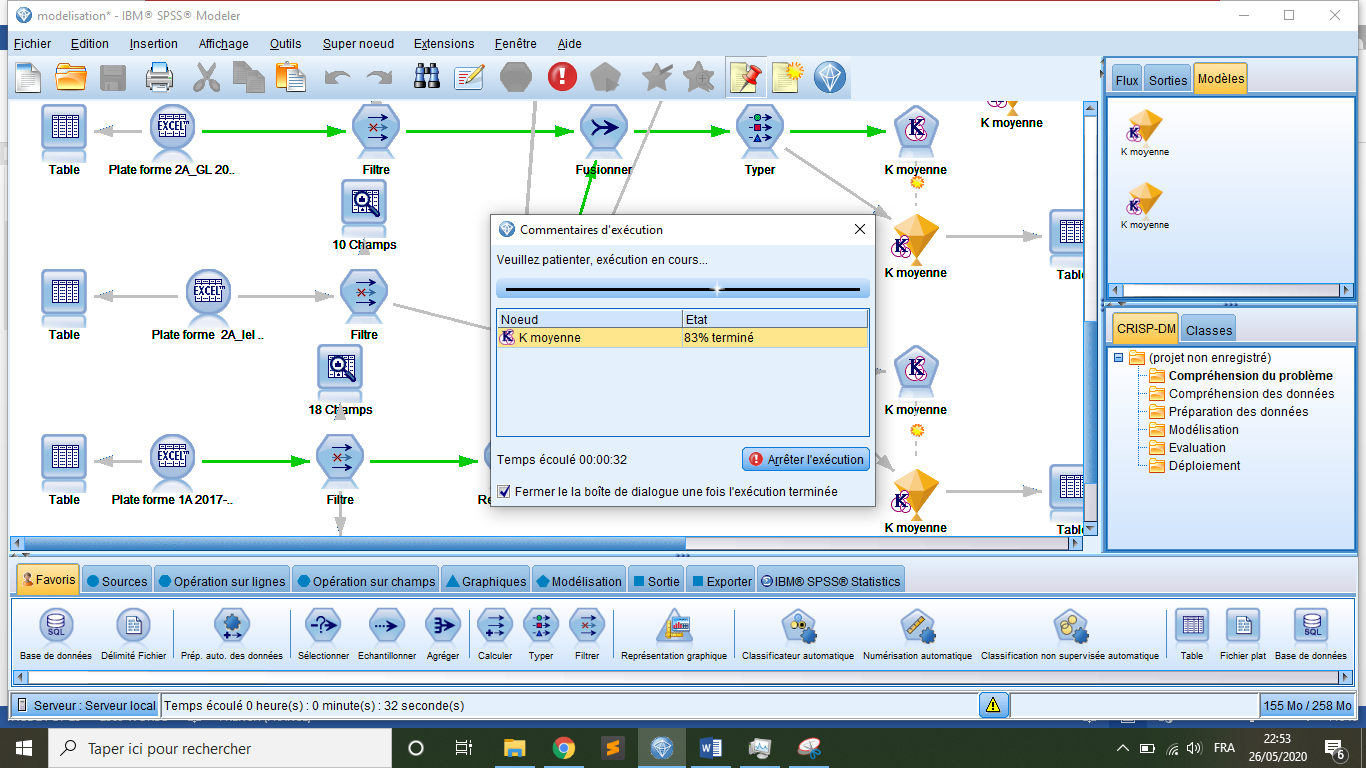
Nous avons définit tous les champs comme des champs « entrée » sauf nom et prénom que nous avons défini leur rôle comme « aucun ».





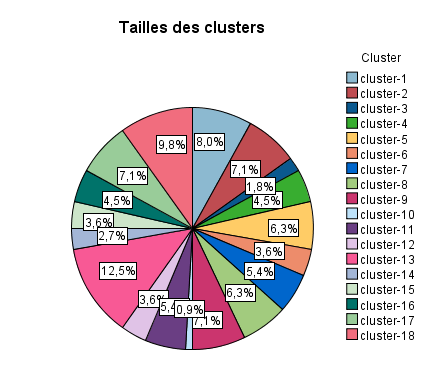
**Figure 20 : Géneration du modèle**

Nous avons fixé le nombre de clusters pour la fillière isem sur 4. Nous aurons donc 4 groupes des étudiants chaque groupe est constitué des étdudiants ayant des notes similaires.



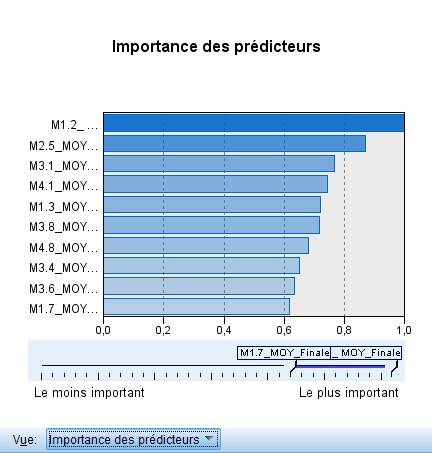
**Figure 21 : Temps d’exécution**

Pour le temps d’exécution, il est raisonnable par exemple dans le cas de la fillière GL le temps d’exécution est 33 secondes



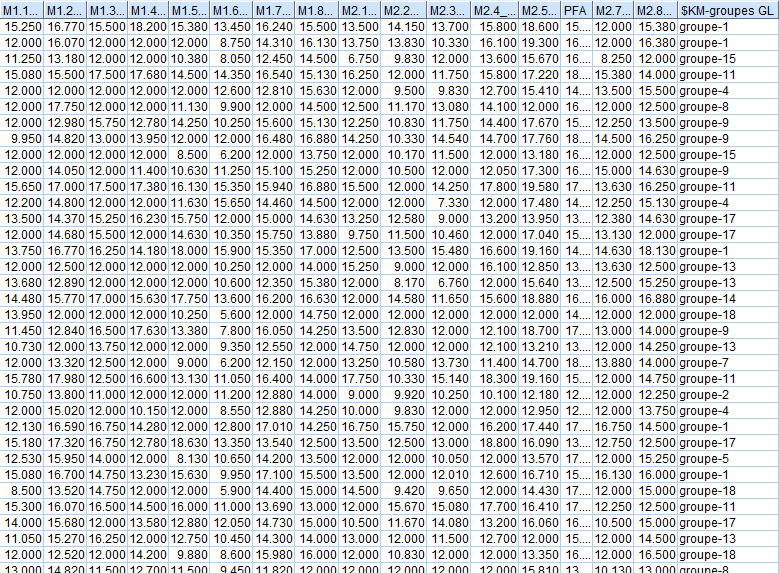
**Figure 22 : Taille des clusters**

Nous avons obtenu des clusters avec différents tailles. On a un seul cluster de taille 1. Donc nous avons atteint notre objectif.



**Figure 23 : Importance des prédicteurs**

Dans le cas des étudiants de la fillière Génie Logiciel (GL) le prédicteur le plus important est la note du module 1.2.



**Figure 24 : Résultat**