**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.

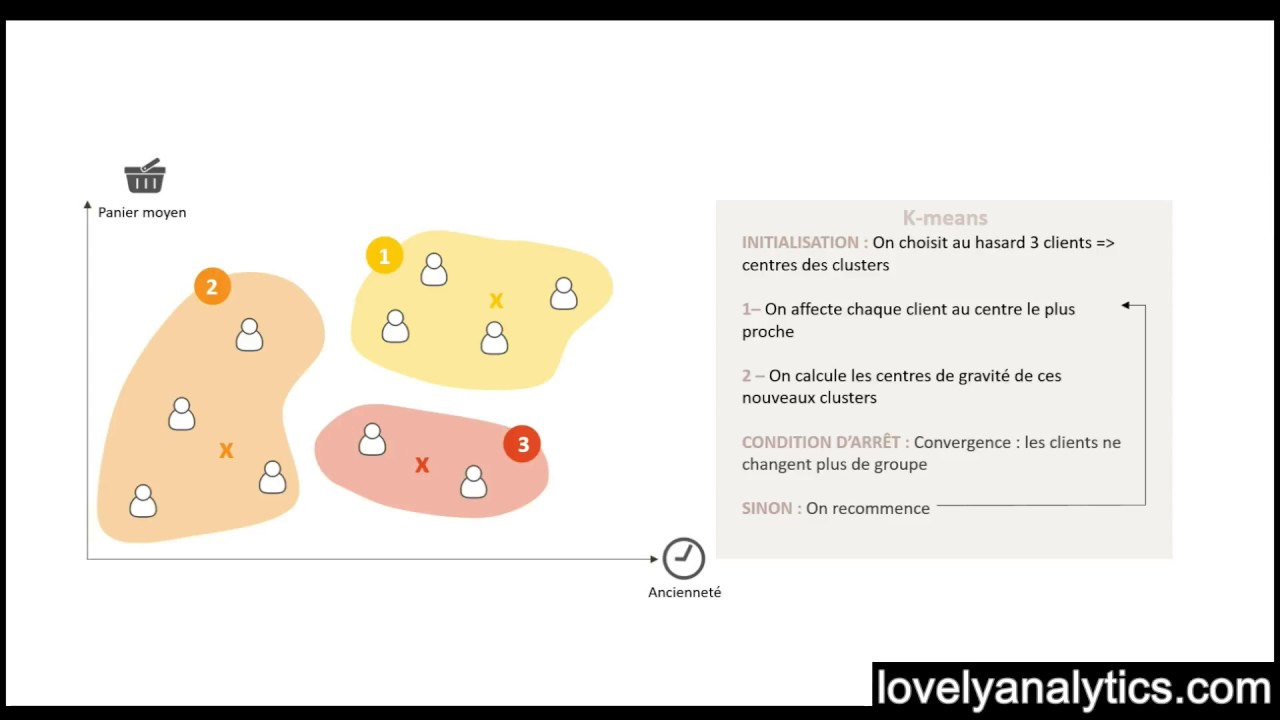
* 1. **Clustering**

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 :



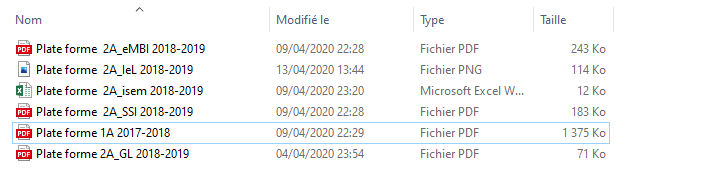
**Figure 1 : Principe algorithmique de K-means**

**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.

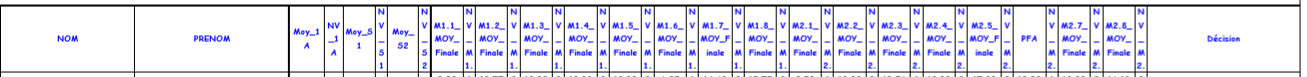


**Figure 2 : Les données collectées**

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.



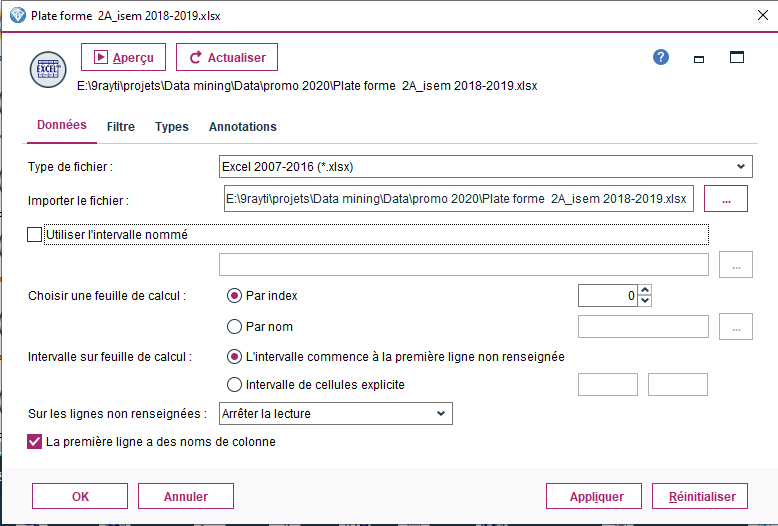
**Figure 3 : les composants de chaque fichier**

|  |  |  |  |
| --- | --- | --- | --- |
| 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**

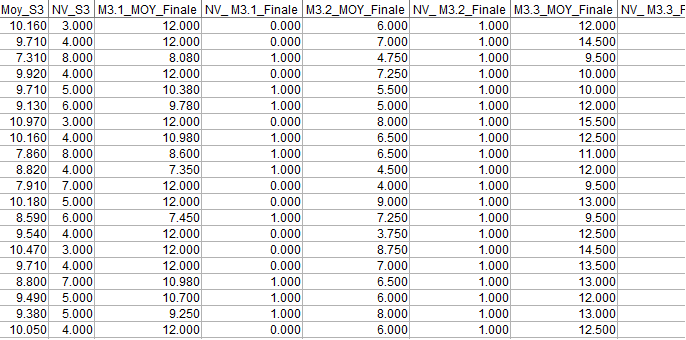
Pour explorer les données il faut tout d’abord transformer nos fichiers en un format adopter par le Spss modeler. Nous avons trois formats : des fichiers excel adopter par le Spss modeler et des fichiers PDF et une image. Pour les fichiers PDF nous avons pu les convertir en fichiers pdf facilement. Par 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**

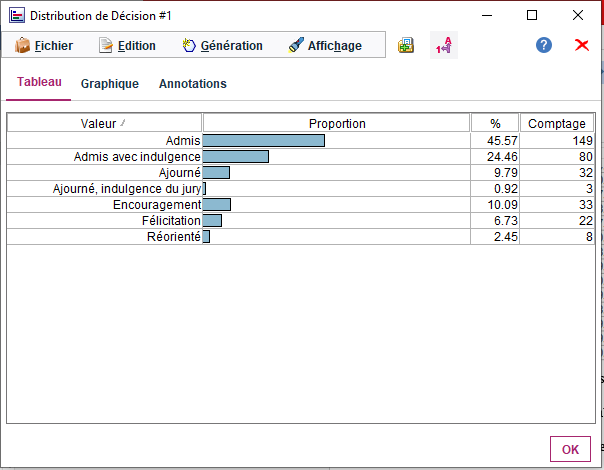
**Figure 5 : Explorer les données à l’aide des tableaux**

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.



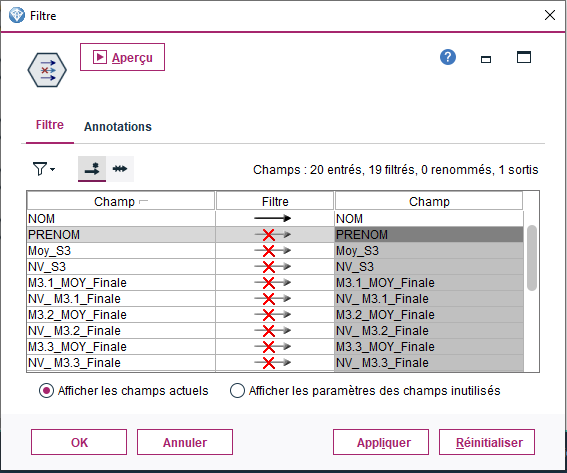


**Figure 6 : Pourcentage de réussite**

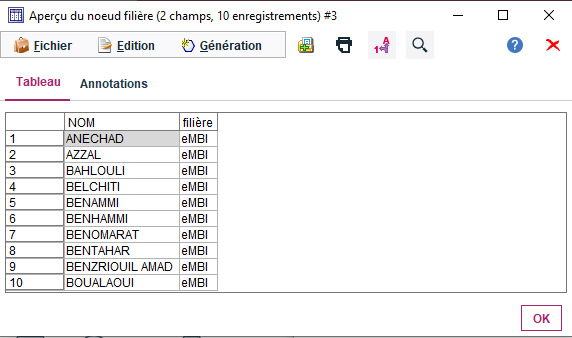
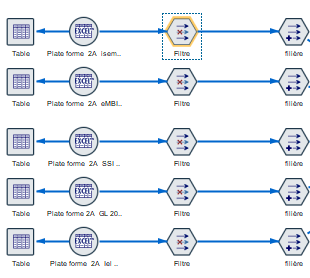
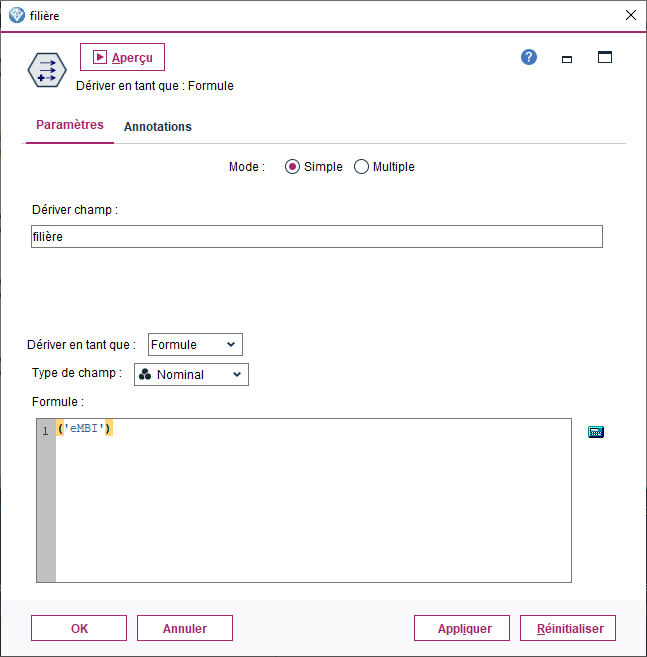
D’après le résultat du nœud Distribution, la plupart des étudiants sont des admis. Ainsi nous avons constaté que nous avons seulement 22 étudiants qui réussirent avec félicitation , c’est-à-dire ils ont une note supérieure à 15.

* 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 ».

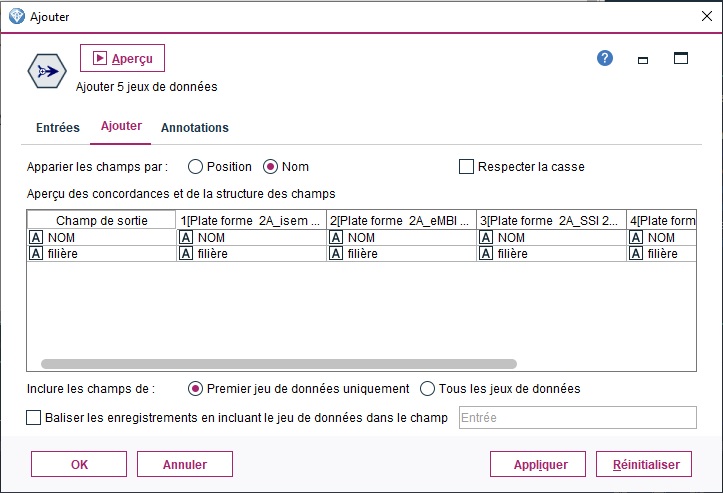
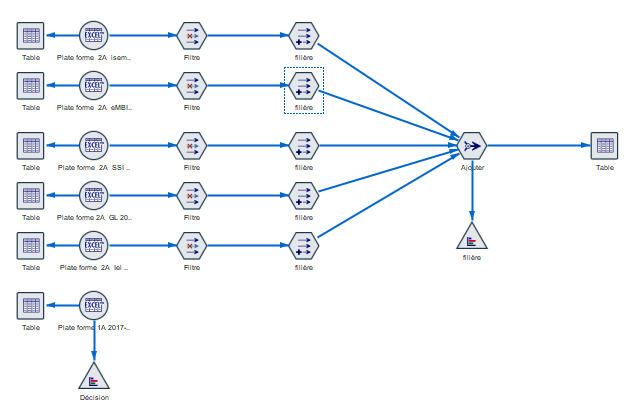
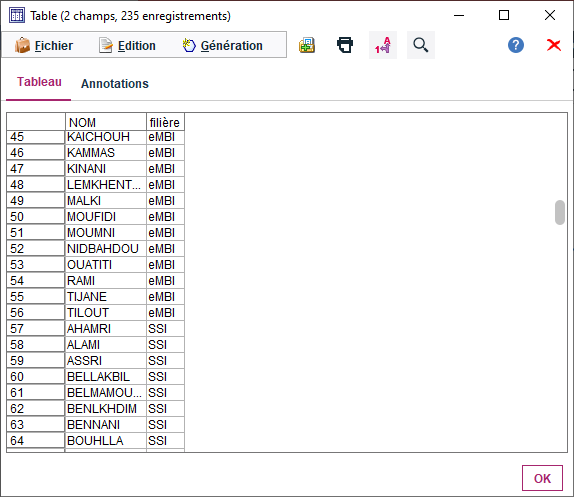


**Figure 7 : Extraire le champ des noms**

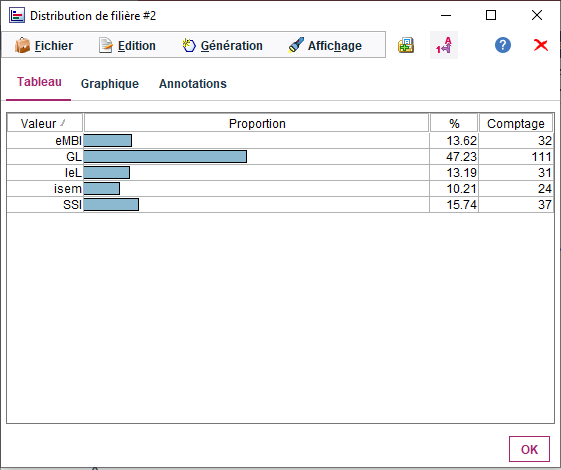


**Figure 8 : Ajouté la colonne de la filière**

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 ».



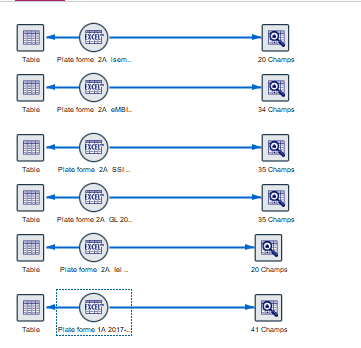
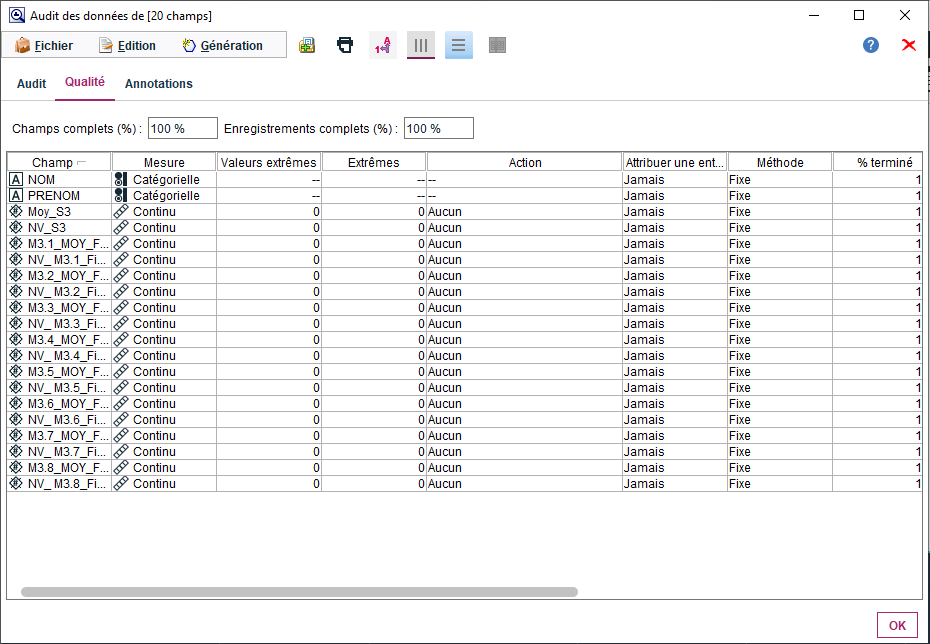
**Figure 9 : Union des nœuds**



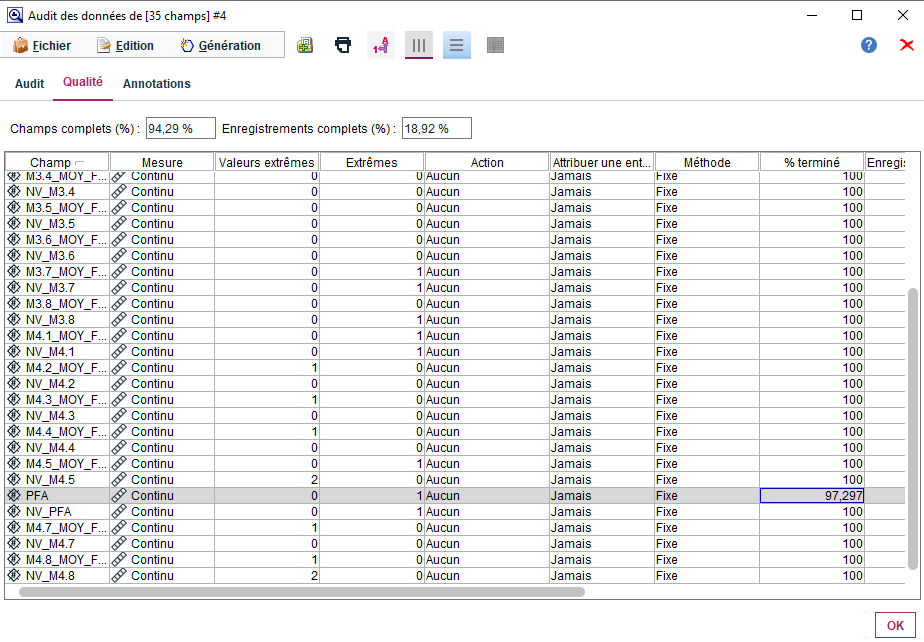
**Figure 10 : Pourcentage des filières**

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.