



ÉCOLE NATIONALE SUPÉRIEURE D'INFORMATIQUE ET D'ANALYSE DES
SYSTÈMES - RABAT

Mémoire de Projet de Data Mining

Filière:

e-Management and Business Intelligence

Prédiction des abandons des cours en ligne

Réalisé par :

BENOMAR Mohammed Taha
SAHLAOUI Botaina
JABILOU Hiba
KASSAOUI Wissal

Encadré par :

Mme. BENBRAHIM Houda



Remerciements

Au terme de ce travail, nous tenons à remercier toute personne qui a aidé à réaliser cette tâche. Nous exprimons surtout nos remerciements Mme Benbrahim Houda qui nous fait le grand honneur d'évaluer ce travail. Nous remercions aussi le corps administratif de l'ENSIAS et surtout le corps enseignant de la filière e-mbi sous la direction de Mr Janati Idrissi Mohamed Abdou.



Résumé

Durant les années précédentes, l'analyse de données se présente comme étant une approche nécessaire pour faire dégager des informations nouvelles à partir des données afin de prendre des décisions. Cependant, l'analyse des données présente des limitations. Ceci a poussé à la découverte du concept de la Data Mining. C'est un concept où de faible hypothèses sont faites sur les lois statistiques suivies, et selon Piateski-Shapiro la Data Mining est "l'extraction d'information originale, auparavant inconnues et potentiellement utiles, à partir de données".

Dans le cadre de ce projet, nous allons aborder un sujet Data Mining où l'objectif est de prédire les étudiants ayant abandonnés les cours MOOC (massive open online course) en se basant sur de différentes informations. Pour ce faire, nous avons utilisé l'outil SPSS modeler qui offre plusieurs algorithmes permettant d'arriver au but. Pour notre cas nous avons choisi l'arbre de décision, Naive Bayes et les réseaux de neurones. En ce qui concerne la méthodologie, puisqu'on a affaire à un projet Data Mining, nous avons suivi tout au long du projet la méthodologie CRISP-DM qui comprend les étapes suivantes: la compréhension du problème, la compréhension de données, la préparation de données, la modélisation, l'évaluation et le déploiement.

Abstract

In previous years, data analysis was presented as a necessary approach to extract new information from data in order to make decisions. However, the data analysis has limitations. This led to the discovery of the concept of Data Mining. It is a concept where low assumptions are made about the statistical laws followed, and according to Piateski-Shapiro Data Mining is "the extraction of original information, previously unknown and potentially useful, from data".

As part of this project, we will address a Data Mining subject where the objective is to predict students who have dropped out of MOOC (massive open online course) courses based on different information. To do this, we used the SPSS modeler tool which offers several algorithms to achieve the goal. For our case we have chosen the decision tree, Naive Bayes and neural networks. As far as the methodology is concerned, since we are dealing with a Data Mining project, we have followed the CRISP-DM methodology throughout the project, which includes the following steps: understanding the problem, understanding the data, preparing data, modeling, evaluation and deployment.

List of Figures

1	Fichiers .csv représentant la Dataset	12
2	Schème de la base de données	12
3	Chargement de la table Courses	17
4	Schema Courses	17
5	Module Duration (in days)	18
6	Length of Modules	18
7	Length of Modules	19
8	Chargement de la table Assessments	19
9	Schema Assessments	20
10	Pondération des modules	20
11	Types of Assessments	21
12	Assessment per module	21
13	Assessment per module and presentation	22
14	Chargement de la table StudentAssssments	22
15	Schema StudentAssssments	23
16	Schema StudentAssssments	23
17	Chargement de la table VLE	24
18	Schema VLE	24
19	Most Common VLE Activities	25
20	Chargement de la table StudentInfo	25
21	Schema StudentInfo	26
22	Boxplot of num_of_prev_attempts	26
23	Boxplot of student_credits	27
24	Modules codes	27
25	Years of course	28
26	Gender	28
27	Number of Previous Attempts	29
28	disability	29
29	Age band	30
30	Most Common Region	30
31	Most Common Education Level	31
32	Socio-Economic Status	31
33	Student Outcome	32
34	Statistic about Student Outcome	32
35	Chargement de la table StudentRegistration	33
36	Schema StudentRegistration	33
37	Students missing from the Results table	34
38	Students Information table missing from the Assessment Results table	35
39	Student Registration Date (Month)	36
40	Student Unregistration Date (Month)	36
41	Registration months varied across module types	37
42	Unregistrations per Module per Month	37
43	Chargement de la table studentVle	38
44	Schema studentVle	38
45	VLE interaction per Module	39

46	Average VLE Interaction	39
47	Most Common VLE Activity	40
48	Qualité assessment	40
49	Qualité VLE	40
50	Qualité StudentVLE	41
51	Première étape du fusionnement	43
52	Deuxième étape du fusionnement	43
53	Processus du fusionnement	44
54	Calcul de fail_rate	44
55	Histogramme de num_of_prev_attempts	45
56	colonnes abandonnées 1	45
57	Corrélation de weighted_score	46
58	Noeud Remplacer	46
59	Données manquantes	47
60	remplacer weighted_score	47
61	remplacer late_rate	48
62	Médiiane de total_assesments	48
63	remplacer total_assesments	49
64	distribution de age_band	49
65	nouvelle distribution de age_band	49
66	distribution de region	50
67	nouvelle distribution de region	50
68	Noeud Calculer	51
69	Dropout -1	51
70	Dropout -2	51
71	nettoyage dans SPSS	52
72	Super Noeud Données manquantes	52
73	Implémentation de la validation croisée avec SPSS	55
74	Choix de la colonne Dropout comme cible et les autres colonnes comme entrée	56
75	Arbre de décision par SPSS	56
76	Arbre de décision	57
77	Réseau bayésien par SPSS	57
78	Réseau bayésien	58
79	Performances respectives de l'arbre de décision et réseaux de neurones	58
80	Performance du réseau bayésien	58

Contents

1 La compréhension métier

1.1	Détermination des objectifs stratégiques	10
1.2	Évaluation de la situation actuelle	10
1.3	Détermination de la problématique du Datamining	10
1.4	Plan du projet	10

2 Compréhension des données

2.1	Collecte des données	12
2.2	Description des données	12
2.3	Exploration des données	16
2.3.1	Fichier Courses	16
2.3.2	Fichier Assessments	19
2.3.3	Fichier StudentAssssments	22
2.3.4	Fichier VLE	23
2.3.5	Fichier StudentInfo	25
2.3.6	Fichier StudentRegistration	33
2.3.7	Fichier studentVle	37
2.4	Vérification de la qualité des données	40

3 Préparation des données

3.1	Intégration des données	43
3.2	Création de nouvelles colonnes	44
3.3	Sélection des données	44
3.4	Nettoyage des données	46
3.4.1	Données manquantes	46
3.4.2	Vérification de la distribution des données	49
3.4.3	Variable cible	50

4 Modélisation

4.1	Sélection des techniques de modélisation	54
4.1.1	Les réseaux neuronaux	54
4.1.2	Réseau bayésien	54
4.1.3	Arbre de décision	55
4.2	Génération d'une conception de test	55
4.3	Construire le modèle	55
4.3.1	Implémentation de l'arbre de décison	56
4.3.2	Implémentation du réseau bayésien	57

4.4 Évaluation de la performance des modèles	58
5 Evaluation et déploiement	59
5.1 Evaluation	60
5.1.1 Évaluation des résultats	60
5.1.2 Déterminer les prochaines étapes	60
5.2 Déploiement	60
6 Conclusion	62

Introduction générale

L'importance et la popularité de l'apprentissage en ligne ont beaucoup augmenté à mesure que le monde entre dans l'ère de l'information. Ce vaste potentiel a fourni l'apprentissage des opportunités pour des millions d'apprenants à travers le monde. COVID-19 a mis l'accent sur le potentiel de l'apprentissage en ligne qui s'est avéré être un moyen prometteur pour une éducation équitable et de qualité ainsi qu'une composante essentielle du système éducatif lorsque les campus ferment.

D'autre part, les cours en ligne ouvert massifs (massive open online course MOOC) sont devenus de plus en plus populaires depuis leur lancement en 2008. Un concept clé des MOOC est de fournir des cours en libre accès via Internet qui peuvent échelle à n'importe quel nombre d'étudiants inscrits. Il s'agit de cours sur Internet où les étudiants peuvent apprendre à leur propre rythme et suivre leur propre emploi du temps. Les universités ont également changé leur enseignement traditionnel en classe confinée en accueillant des MOOC.

Compte tenu des différences entre le paradigme d'apprentissage traditionnel et les MOOC, un nouveau programme de recherche axé sur la prédiction et l'explication de l'abandon des étudiants et des faibles taux d'achèvement dans les MOOC a émergé. La nature gratuite des MOOC a eu pour conséquence que de nombreux étudiants abandonnent le cours ou ne sont pas en mesure d'obtenir de bonnes notes. Alors que les plateformes de cours en ligne ouverts et massifs (MOOC) offrent des connaissances d'une manière nouvelle et unique, le nombre très élevé d'abandons est un inconvénient important.

Chapitre 1

1 La compréhension métier

La phase de compréhension métier se concentre sur la compréhension des objectifs et des exigences du projet. On commence tout d'abord par le contexte générale du projet ainsi que la détermination des objectifs métiers, ensuite l'évaluation de la situation actuelle, puis la détermination de la problématique du Datamining et finalement le plan du projet

1.1 Détermination des objectifs stratégiques

Le taux élevé d'abandon des cours en ligne ouverts et massifs (MOOC) a été une préoccupation majeure des chercheurs et des éducateurs au fil des ans. Ainsi, prédire leur probabilité de décrochage serait utile pour maintenir et encourager les activités d'apprentissage des élèves. Le défi consiste à prédire l'échec scolaire des étudiants en se basant sur la collecte d'attributs fournissant des informations sur les caractéristiques pré-académiques des étudiants et leur interaction avec la plateforme d'apprentissage virtuelle.

1.2 Évaluation de la situation actuelle

Les données utilisées dans cette recherche proviennent de l'Open University Learning Analytics Dataset. Il contient des données sur les cours, les étudiants et leurs interactions avec l'environnement d'apprentissage virtuel pour sept cours sélectionnés. Les auteurs du Dataset ont souligné que l'activité quotidienne de l'étudiant peut être facilement rendue anonyme. Selon les auteurs, une information binaire sur un étudiant s'il a été actif un jour donné peut être aussi significative que l'utilisation de données privées sensibles comme le sexe, le handicap et le niveau d'éducation le plus élevé. Les auteurs ont classé les étudiants en fonction de leur utilisation quotidienne de la plate-forme d'apprentissage virtuel. Tous les différents types d'activités de l'environnement d'apprentissage virtuel ont été combinés en une seule mesure. combinés en une seule mesure.

1.3 Détermination de la problématique du Datamining

Puisque l'objectif stratégique est clairement défini, il convient maintenant de le traduire en concepts de Data Mining. L' objectif majeur du projet est de proposer un outil prédictif suffisamment efficace pour prédire le potentiel d'échec ou de réussite des étudiants grâce à l'analyse des données collectées sur la plate-forme d'apprentissage virtuelle. Il s'agit donc d'un problème d'apprentissage supervisé permettant de prédire par un algorithme de classification si l'élève va réussir ou échoué le cours virtuel.

1.4 Plan du projet

Nous allons commencer par la collecte et la préparation des données. L'étape suivante consiste à préparer l'ensemble des données. Ensuite, des modèles sont générés par des algorithmes d'apprentissage. Les résultats de ces modèles sont utilisés pour prédire les échecs.

Chapitre 2

2 Compréhension des données

Cette étape vise à donner un sens aux données après leur collecte , et cela n'est possible qu'en établissant une description et une exploration de données.

2.1 Collecte des données

Nous avons collecté les données grâce à l'Open University Learning Analytics dataset. Ci-dessous sont listés les fichiers .csv utilisés dans le projet :

 assessments.csv	25/09/2015 12:36	Fichier CSV Micro... Format : CSV	9 Ko
 courses.csv	25/09/2015 12:36	Fichier CSV Micro... Format : CSV	1 Ko
 studentAssessment.csv	25/09/2015 12:36	Fichier CSV Micro... Format : CSV	5 557 Ko
 studentInfo(1).csv	24/05/2022 00:32	Fichier CSV Micro... Format : CSV	2 631 Ko
 studentInfo(2).csv	24/05/2022 00:43	Fichier CSV Micro... Format : CSV	2 672 Ko
 studentInfo.csv	25/09/2015 12:36	Fichier CSV Micro... Format : CSV	3 381 Ko
 studentRegistration.csv	25/09/2015 12:36	Fichier CSV Micro... Format : CSV	1 084 Ko
 studentsregistration2.csv	30/05/2022 14:52	Fichier CSV Micro... Format : CSV	1 284 Ko
 studentVle.csv	25/09/2015 12:36	Fichier CSV Micro... Format : CSV	443 200 Ko
 vle.csv	25/09/2015 12:36	Fichier CSV Micro... Format : CSV	255 Ko

Figure 1: Fichiers .csv représentant la Dataset

2.2 Description des données

Le schéma suivant illustre la structure globale de notre base de données :

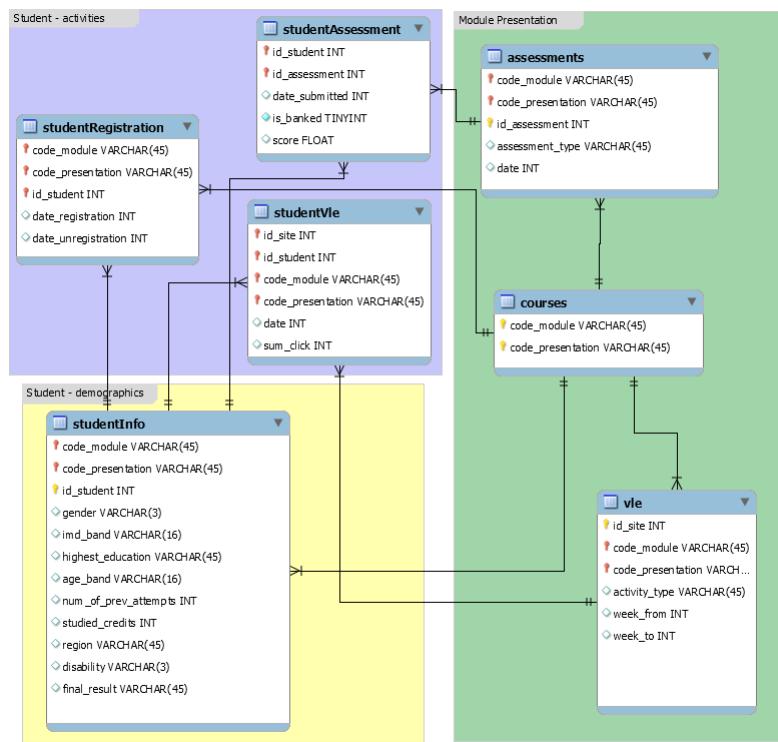


Figure 2: Schème de la base de données

- **course.csv** : ce fichier contient la liste des modules disponibles et leurs présentations. Les colonnes du fichier sont :

- **code_module** : c'est le code du module qui sert comme identifiant.
- **code_presentation** : c'est le code de la présentation des cours, il s'agit de l'année : 'B' si la présentation du cours commence en Février et 'J' si elle commence en Octobre.
- **length** : il s'agit de la durée de la présentation du module en jours.

- **assessments.csv** : ce fichier contient les informations concernant les évaluations au cours du module. Chaque présentation contient un certain nombre d'évaluations suivis par l'examen final. Les colonnes sont les suivantes :
 - **code_module** : code d'identification du module auquel appartient l'évaluation.
 - **code_presentation** : code d'identification de la présentation à laquelle appartient l'évaluation.
 - **id_assessment** : identifiant de l'évaluation .
 - **assessment_type** : il existe trois types d'évaluations : Tutor Marked Assessment (TMA), Computer Marked Assessment (CMA) and Final Exam (Exam).
 - **date** : donne des informations sur la date de soumission finale de l'évaluation calculée en nombre de jours depuis le début de la présentation du module. La date de début de la présentation porte le numéro 0 (zéro).
 - **weight** : poids de l'évaluation en pourcentage. En règle générale, les examens sont traités séparément et ont un poids de 100% ; la somme de toutes les autres évaluations est de 100 %.
- **vle.csv** : Le fichier csv contient des informations sur les matériaux disponibles dans le VLE (Virtual Learning Environment). Il s'agit généralement de pages html, de fichiers pdf, etc. Les étudiants ont accès à ces documents en ligne et leurs interactions avec les documents sont enregistrées. Le fichier vle.csv contient les colonnes suivantes :

- **id_site** : identifiant du matériel .

- **code_module** : un code identifiant du module .
 - **code_presentation** : code identifiant la présentation.
 - **activity_type** : le rôle associé au matériel du module.
 - **week_from** : la semaine à partir de laquelle il est prévu d'utiliser le matériel.
 - **week_to** : semaine jusqu'à laquelle il est prévu d'utiliser le matériel.
-
- **studentInfo.csv:** Ce fichier contient des informations démographiques sur les étudiants ainsi que leurs résultats. Le fichier contient les colonnes suivantes:
- **code_module:** un code d'identification d'un module auquel l'étudiant est inscrit.
 - **code_presentation:** le code d'identification de la présentation au cours de laquelle l'étudiant est inscrit sur le module.
 - **id_student:** identifiant unique pour l'élève.
 - **gender:** sexe de l'élève.
 - **region:** identifie la région géographique où l'étudiant a vécu pendant la présentation du module.
 - **highest_education:** niveau d'éducation le plus élevé des étudiants à l'entrée de la présentation du module.
 - **imd_band:** spécifie la bande d'indice de dépravation multiple du lieu où l'étudiant a vécu pendant la présentation du module.
 - **age_band:** tranche d'âge de l'élève.
 - **num_of_prev_attempts:** le nombre de fois que l'étudiant a tenté ce module.

- **studied_credits:** le nombre total de crédits pour les modules que l'étudiant étudie actuellement.
 - **disability:** indique si l'étudiant est handicapé.
 - **final_result:** résultat final de l'étudiant dans la présentation du module.
- **studentRegistration.csv:** Ce fichier contient des informations sur l'heure d'inscription de l'étudiant à la présentation du module. Pour les étudiants qui se sont désinscrits, la date de désinscription est également enregistrée. Le fichier contient cinq colonnes:
- **code_module:** code d'identification d'un module.
 - **code_presentation:** code d'identification de la présentation.
 - **id_student:** identifiant unique de l'étudiant.
 - **date_registration:** il s'agit du nombre de jours mesurés par rapport au début de la présentation du module (par exemple, la valeur négative -30 signifie que l'étudiant s'est inscrit à la présentation du module 30 jours avant son début).
 - **date_unregistration:** date de désinscription de l'étudiant à la présentation du module, c'est le nombre de jours mesurés par rapport au début de la présentation du module. Les étudiants qui ont terminé le cours ont ce champ vide. Les étudiants qui se sont désinscrits ont Retrait comme valeur de la colonne final_result dans le fichier studentInfo.csv.
- **studentAssessment.csv:** Ce fichier contient les résultats des évaluations des étudiants. Si l'étudiant ne soumet pas l'évaluation, aucun résultat n'est enregistré. Les soumissions d'examen final sont manquantes, si le résultat des évaluations n'est pas stocké dans le système. Ce fichier contient les colonnes suivantes:
- **id_assessemement:** identifiant de l'évaluation.
 - **id_student:** identifiant unique de l'étudiant.

- **date_submitted:** la date de soumission de l'étudiant, mesurée en nombre de jours depuis le début de la présentation du module.
 - **is_banked:** un indicateur d'état indiquant que le résultat de l'évaluation a été transféré d'une présentation précédente.
 - **score:** la note de l'évaluation. La plage va de 0 à 100. Le score inférieur à 40 est interprété comme un échec. Les notes sont comprises entre 0 et 100.
- **studentVle.csv:** Le fichier contient des informations sur les interactions de chaque étudiant avec le matériel du VLE. Ce fichier contient les colonnes suivantes:
- **code_module:** identifiant du module.
 - **code_presentation:** code identifiant de la présentation du module.
 - **id_student:** identifiant unique de l'étudiant.
 - **id_site:** identifiant du matériel VLE.
 - **date:** la date d'interaction de l'étudiant avec le matériel mesurée en nombre de jours depuis le début de la présentation du module.
 - **sum_click:** le nombre de fois qu'un élève interagit avec le matériel au cours de la journée.

2.3 Exploration des données

Durant cette phase, nous avons explorer les 7 fichiers csv qui représentent notre source de donnée, à savoir *courses*, *assessments*, *vle*, *studentInfo*, *studentRegistration*, *studentAssessment* et *studentVle*.

2.3.1 Fichier Courses

Ce fichier contient des informations sur tous les modules disponibles et leurs présentations.

	code_module	code_presentation	module_presentation_length
1	AAA	2013J	268
2	AAA	2014J	269
3	BBB	2013J	268
4	BBB	2014J	262
5	BBB	2013B	240
6	BBB	2014B	234
7	CCC	2014J	269
8	CCC	2014B	241
9	DDD	2013J	261
10	DDD	2014J	262

Figure 3: Chargement de la table Courses

Tout d'abord, on a commencer par faire plusieurs manipulations sur les données du fichier **Courses** pour bien les comprendre. Comme suit:

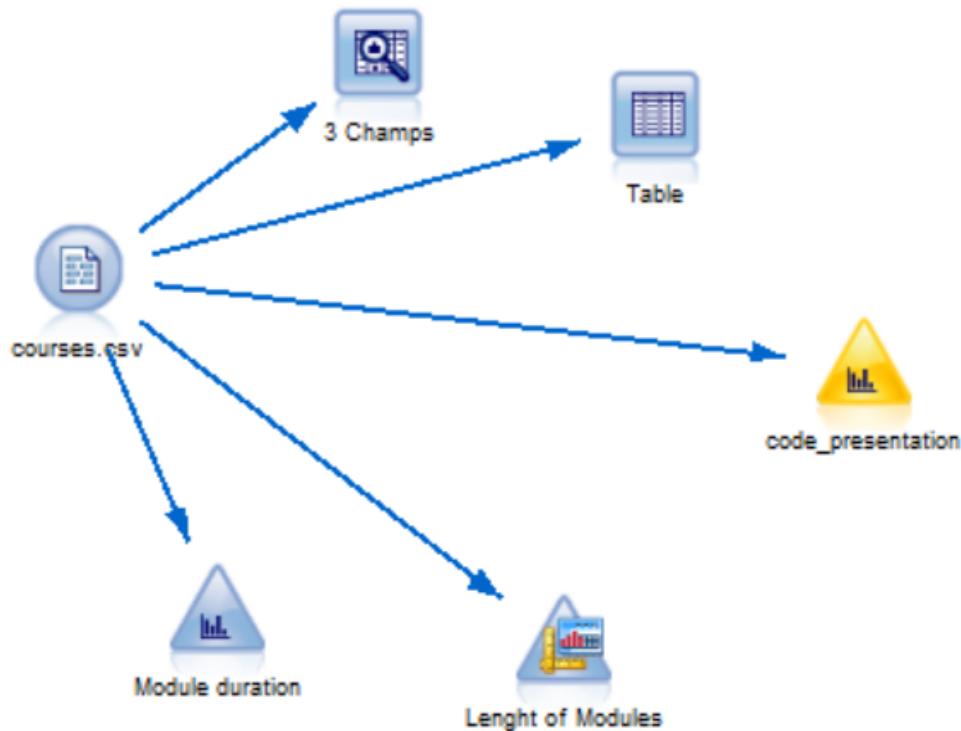


Figure 4: Schema Courses

On a réaliser une visualisation de la longueur des présentation. On remarque que la plupart des cours durent environ 8 mois chacun.

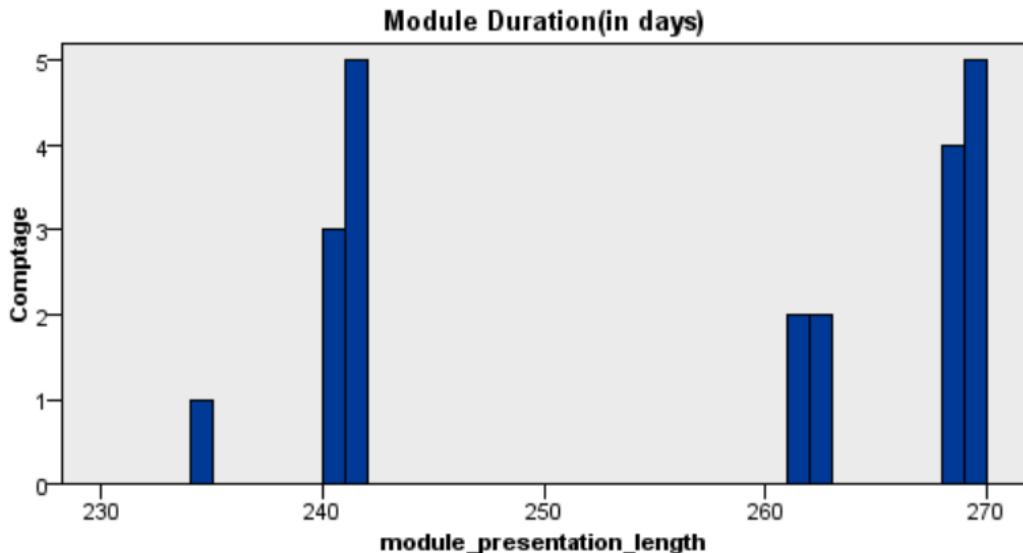


Figure 5: Module Duration (in days)

Après, on a visualiser la longueur des présentations pour chaque module.
Nous pouvons voir que les modules sont de longueurs différentes pour chaque prise.

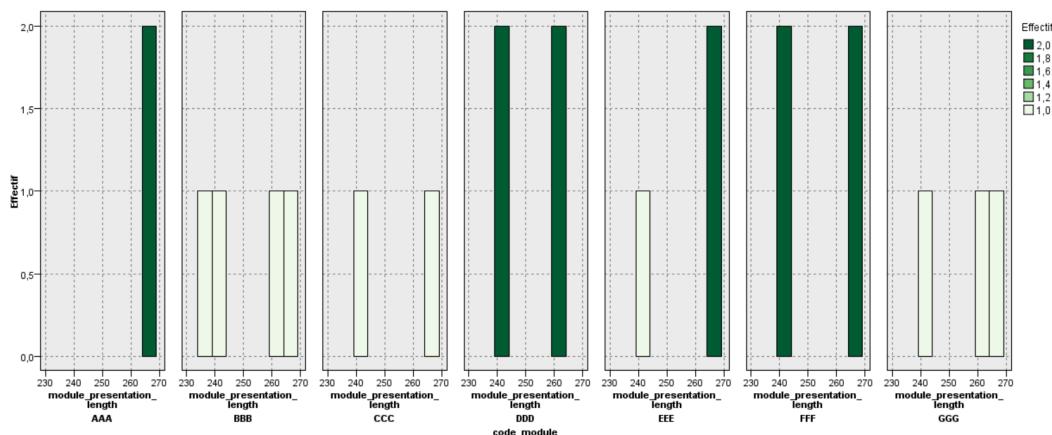


Figure 6: Length of Modules

Ici, on réalise une visualisation de la longueur des modules en tenant compte du code de chaque présentation.

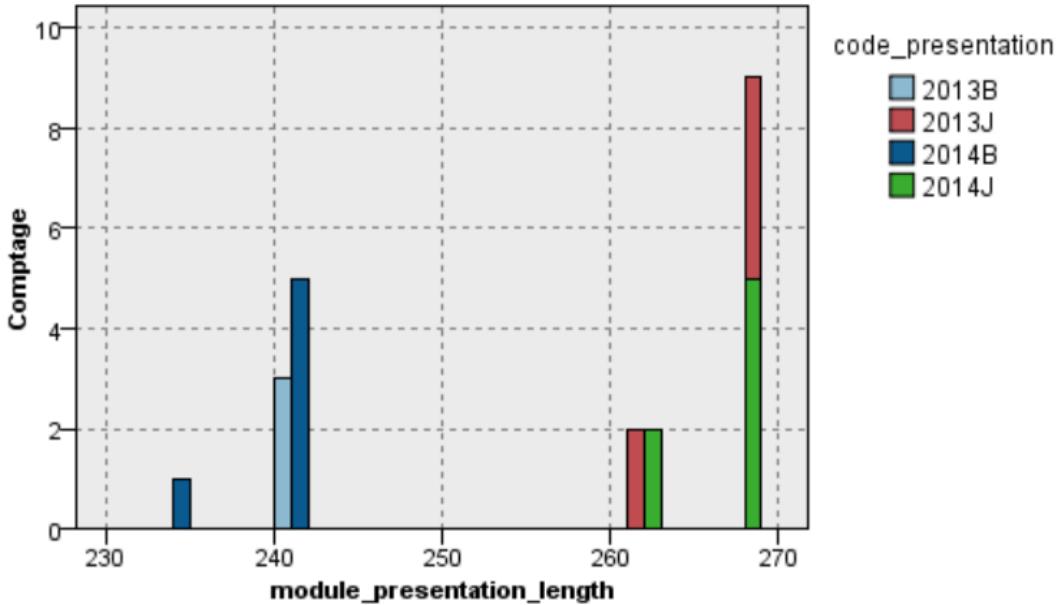


Figure 7: Length of Modules

Ce graphique vérifie en outre que la durée de la présentation du cours varie d'une année à l'autre, quoique légèrement. Cela indique que la durée de présentation du cours offre peu de valeur à l'analyse car elle varie selon les années et les modules. Par conséquent, il peut être nécessaire de le jeter.

2.3.2 Fichier Assessments

Ce fichier contient des informations sur les évaluations dans les présentations de module. Chaque présentation comporte un certain nombre d'évaluations suivies de l'examen final.

	code_module	code_presentation	id_assessment	assessment_type	date	weight
1	AAA	2013J	1752	TMA	19	10
2	AAA	2013J	1753	TMA	54	20
3	AAA	2013J	1754	TMA	117	20
4	AAA	2013J	1755	TMA	166	20
5	AAA	2013J	1756	TMA	215	30
6	AAA	2013J	1757	Exam	\$n...	100
7	AAA	2014J	1758	TMA	19	10
8	AAA	2014J	1759	TMA	54	20
9	AAA	2014J	1760	TMA	117	20
10	AAA	2014J	1761	TMA	166	20

Figure 8: Chargement de la table Assessments

Tout d'abord, on a commencer par faire plusieurs manipulations sur les données du fichier **Assessments** pour bien les comprendre. Comme suit:

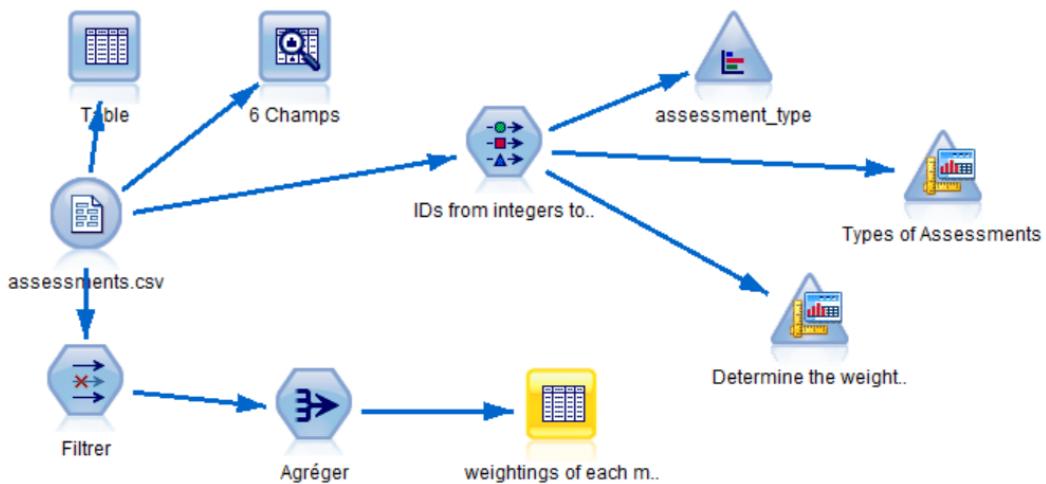


Figure 9: Schema Assessments

Maintenant, on voulait vérifier les pondérations des résultats d'évaluation, et déterminer les pondérations de chaque module. On remarque que:

- La pondération des examens est de 100%
- La pondération de la somme des évaluations est de 100%
- Les modules avec évaluations et examens auraient une pondération de 200%

	code_module	code_presentation	assessment_type	Weight_by_type
1	AAA	2013J	TMA	100
2	AAA	2013J	Exam	100
3	AAA	2014J	TMA	100
4	AAA	2014J	Exam	100
5	BBB	2013B	CMA	5
6	BBB	2013B	TMA	95
7	BBB	2013B	Exam	100
8	BBB	2013J	CMA	5
9	BBB	2013J	TMA	95
10	BBB	2013J	Exam	100
11	BBB	2014B	CMA	5
12	BBB	2014B	TMA	95
13	BBB	2014B	Exam	100
14	BBB	2014J	TMA	100
15	BBB	2014J	Exam	100
16	CCC	2014B	CMA	25
17	CCC	2014B	TMA	75
18	CCC	2014B	Exam	200
19	CCC	2014J	CMA	25
20	CCC	2014J	TMA	75

Figure 10: Pondération des modules

Cela indique que les modules ont à la fois des évaluations (100%) et des examens (100%), c'est pourquoi leur pondération est de 200. Les exceptions sont :

- Module CCC qui a un score de 200 pour les examens. Cela suggère 2 examens.
- Module GGG qui a un score de 0 pour les devoirs. Cela ne suggère aucune affectation.

Après, on a mis en évidence les différents types d'évaluations disponibles, comme suivant:

Valeur	Proportion	%	Comptage
CMA		36.89	76
Exam		11.65	24
TMA		51.46	106

Figure 11: Types of Assessments

On remarque donc que les **TMA** sont les types d'évaluation les plus courants.

Ensuite, on cherche à avoir la répartition de l'évaluation par module.

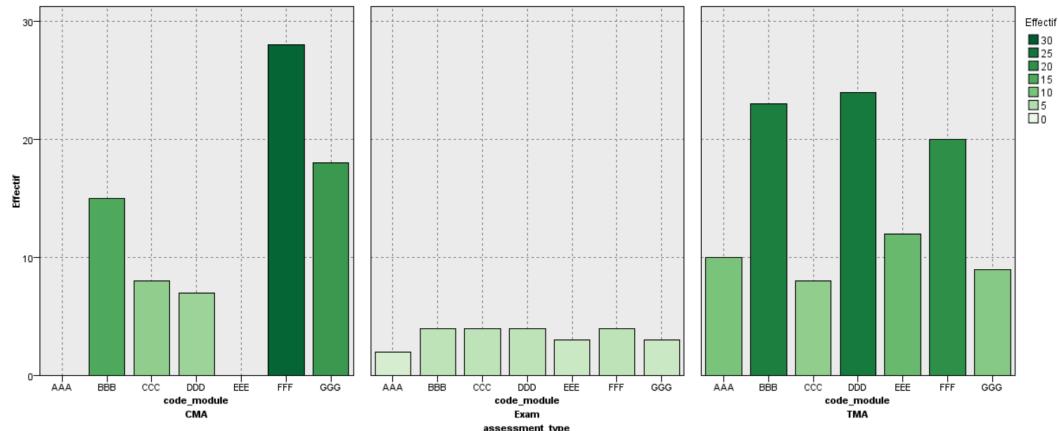


Figure 12: Assessment per module

On remarque que:

- Chaque module comporte un examen et des TMA.
- AAA et EEE n'ont pas d'évaluations CMA.

Pour avoir une vue globale sur les évaluations par modules et par présentation, on a visualisé le graphe suivant:

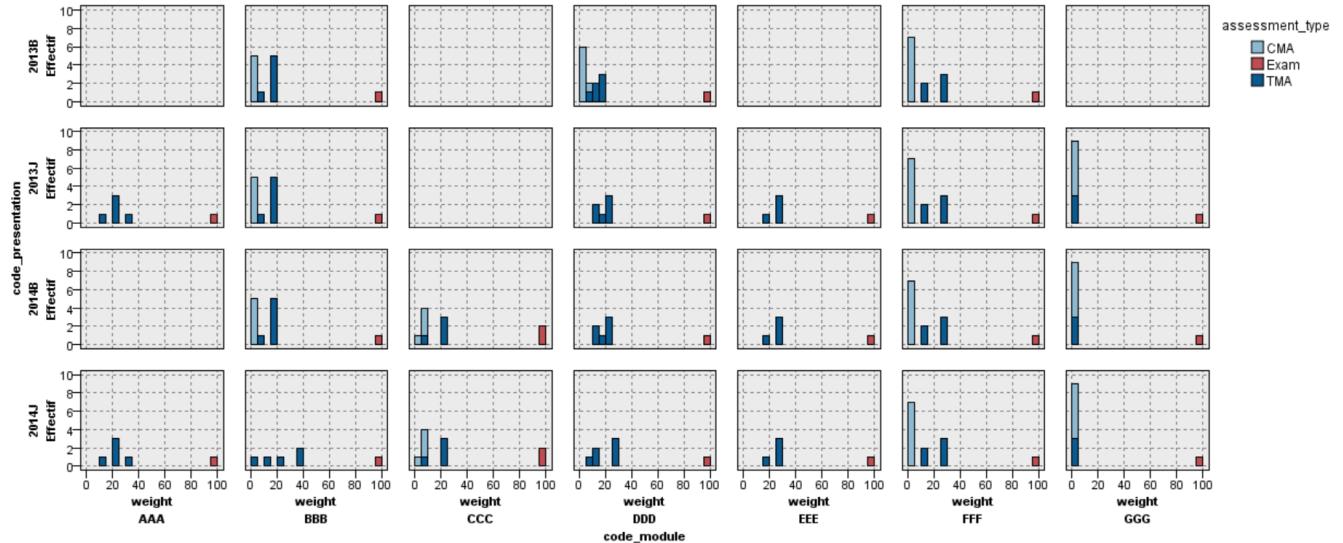


Figure 13: Assessment per module and presentation

2.3.3 Fichier StudentAssssments

Ce fichier contient les résultats des évaluations des élèves. Si l'étudiant ne soumet pas l'évaluation, aucun résultat n'est enregistré. Les soumissions d'examen final sont manquantes, si le résultat des évaluations n'est pas stocké dans le système.

	<i>id_assessment</i>	<i>id_student</i>	<i>date_submitted</i>	<i>is_banked</i>	<i>score</i>
1	1752	11391	18	0	78
2	1752	28400	22	0	70
3	1752	31604	17	0	72
4	1752	32885	26	0	69
5	1752	38053	19	0	79
6	1752	45462	20	0	70
7	1752	45642	18	0	72
8	1752	52130	19	0	72
9	1752	53025	9	0	71
10	1752	57506	18	0	68

Figure 14: Chargement de la table StudentAssssments

Tout d'abord, on a commencer par faire plusieurs manipulations sur les données du fichier **StudentAssssments** pour bien les comprendre. Comme suit:

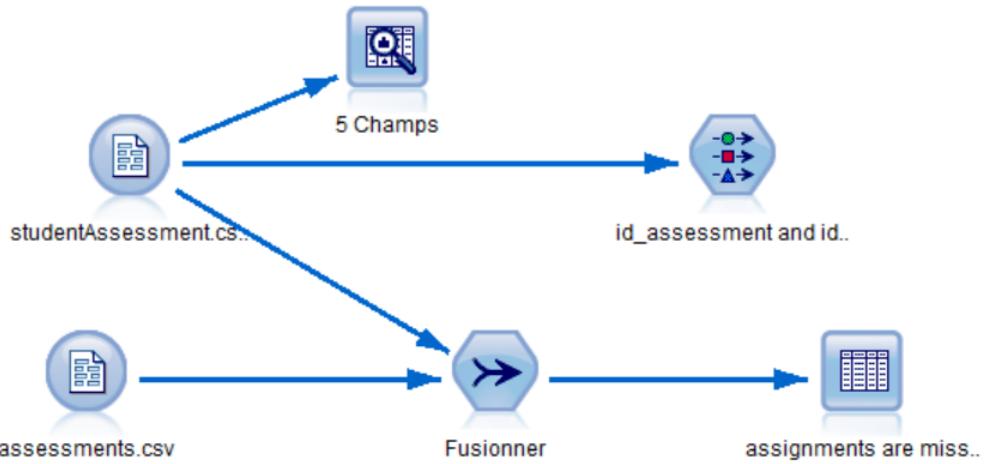


Figure 15: Schema StudentAssssments

Maintenant, on veut déterminer les valeurs manquantes de la table des résultats.

	code_module	code_presentation	id_assessment_result	assessment_type	date	weight
1	AAA	2013J	1757	Exam	\$n...	100
2	AAA	2014J	1763	Exam	\$n...	100
3	BBB	2013B	14990	Exam	\$n...	100
4	BBB	2013J	15002	Exam	\$n...	100
5	BBB	2014B	15014	Exam	\$n...	100
6	BBB	2014J	15025	Exam	\$n...	100
7	EEE	2013J	30713	Exam	235	100
8	EEE	2014B	30718	Exam	228	100
9	EEE	2014J	30723	Exam	235	100
10	FFF	2013B	34872	Exam	222	100
11	FFF	2013J	34885	Exam	236	100
12	FFF	2014B	34898	Exam	227	100
13	FFF	2014J	34911	Exam	241	100
14	GGG	2013J	37424	Exam	229	100
15	GGG	2014B	37434	Exam	222	100
16	GGG	2014J	37444	Exam	229	100
17	CCC	2014B	40087	Exam	\$n...	100
18	CCC	2014J	40088	Exam	\$n...	100

Figure 16: Schema StudentAssssments

On remarque que tous les devoirs manquants dans le tableau des résultats sont des examens avec un poids de module de 100%.

2.3.4 Fichier VLE

Le fichier contient des informations sur les matériaux disponibles dans le VLE. Il s'agit généralement de pages html, de fichiers pdf, etc. Les étudiants ont accès à ces documents en ligne et leurs interactions avec les documents sont enregistrées.

	<code>id_site</code>	<code>code_module</code>	<code>code_presentation</code>	<code>activity_type</code>	<code>week_from</code>	<code>week_to</code>
1	546943	AAA	2013J	resource		
2	546712	AAA	2013J	oucontent		
3	546998	AAA	2013J	resource		
4	546888	AAA	2013J	url		
5	547035	AAA	2013J	resource		
6	546614	AAA	2013J	homepage		
7	546897	AAA	2013J	url		
8	546678	AAA	2013J	oucontent		
9	546933	AAA	2013J	resource		
10	546708	AAA	2013J	oucontent		

Figure 17: Chargement de la table VLE

Tout d'abord, on a commencer par faire plusieurs manipulations sur les données du fichier **VLE** pour bien les comprendre. Comme suit:

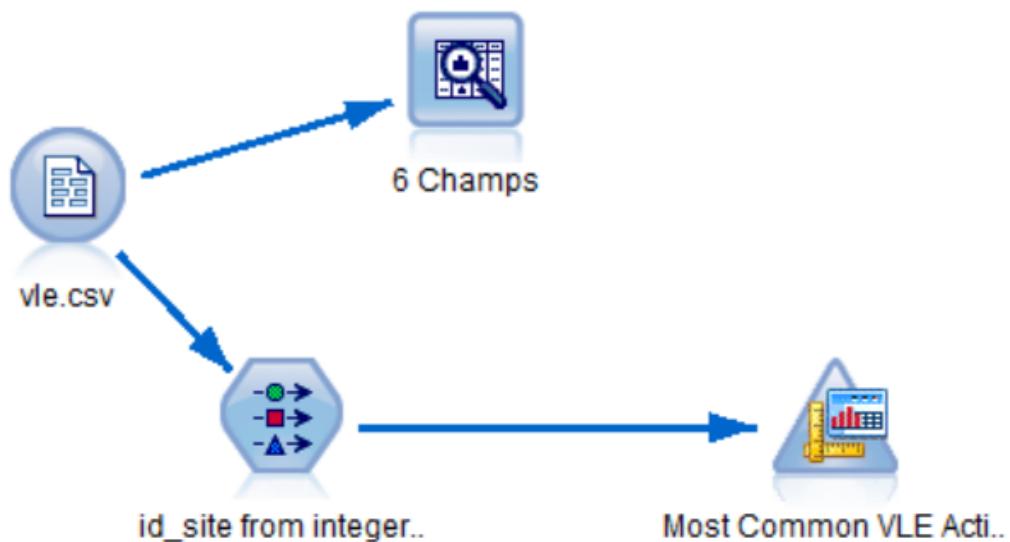


Figure 18: Schema VLE

Après, on cherche à visualiser les Activités VLE les plus courantes.

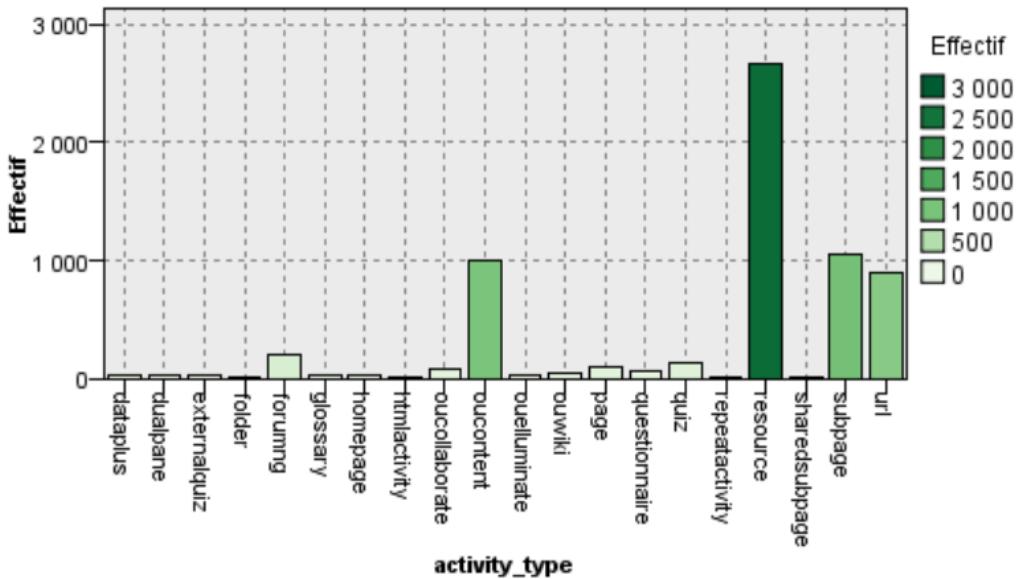


Figure 19: Most Common VLE Activities

On remarque que *Resource*, *oucontent*, *subpage* and *url* sont les activités les plus populaires sur le VLE.

2.3.5 Fichier StudentInfo

Ce fichier contient des informations démographiques sur les étudiants ainsi que leurs résultats.

	code_module	code_presentation	id_student	gender	region	highest_education	imd_band	age_band
1	AAA	2013J	11391 M	East Anglian Region	HE Qualification	90-100%	55<=	
2	AAA	2013J	28400 F	Scotland	HE Qualification	20-30%	35-55	
3	AAA	2013J	30268 F	North Western Region	A Level or Equivalent	30-40%	35-55	
4	AAA	2013J	31604 F	South East Region	A Level or Equivalent	50-60%	35-55	
5	AAA	2013J	32885 F	West Midlands Region	Lower Than A Level	50-60%	0-35	
6	AAA	2013J	38053 M	Wales	A Level or Equivalent	80-90%	35-55	
7	AAA	2013J	45462 M	Scotland	HE Qualification	30-40%	0-35	
8	AAA	2013J	45642 F	North Western Region	A Level or Equivalent	90-100%	0-35	
9	AAA	2013J	52130 F	East Anglian Region	A Level or Equivalent	70-80%	0-35	
10	AAA	2013J	53025 M	North Region	Post Graduate Qualification		55<=	

Figure 20: Chargement de la table StudentInfo

Tout d'abord, on a commencer par faire plusieurs manipulations sur les données du fichier **StudentInfo** pour bien les comprendre. Comme suit:

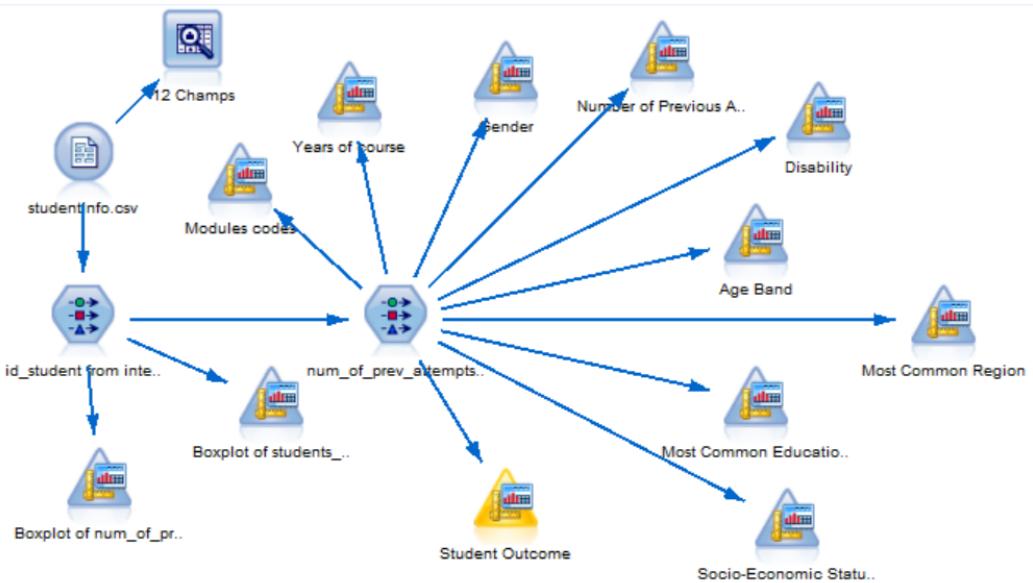


Figure 21: Schema StudentInfo

On a vu d'inspecter les boîtes à moustaches pour les variables numériques comme illustre les schémas suivants:

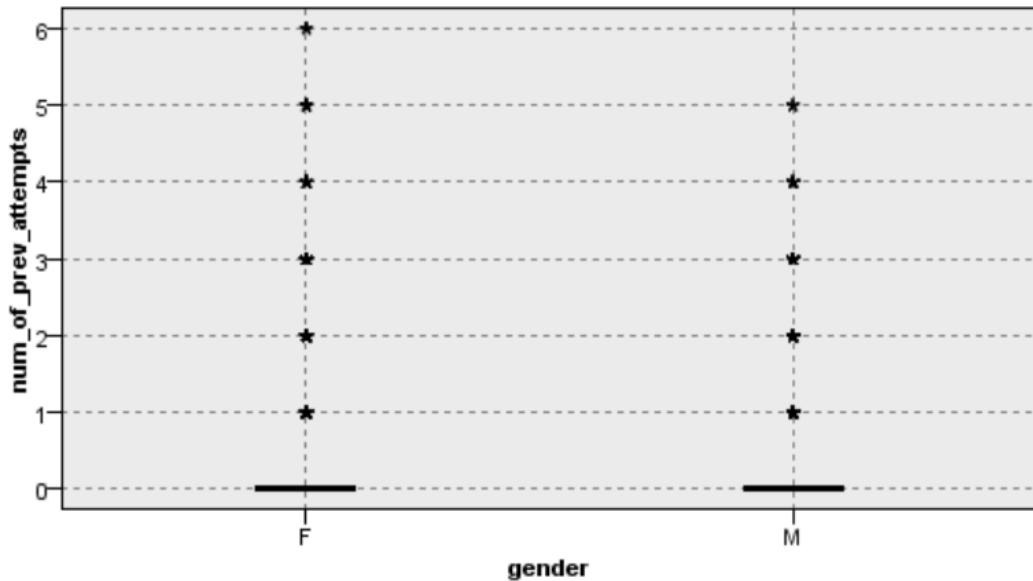


Figure 22: Boxplot of num_of_prev_attempts

On remarque bien que la majorité des étudiants ont terminé le module lors de leur première tentative. Les catégories doivent être réduites davantage lors du nettoyage.

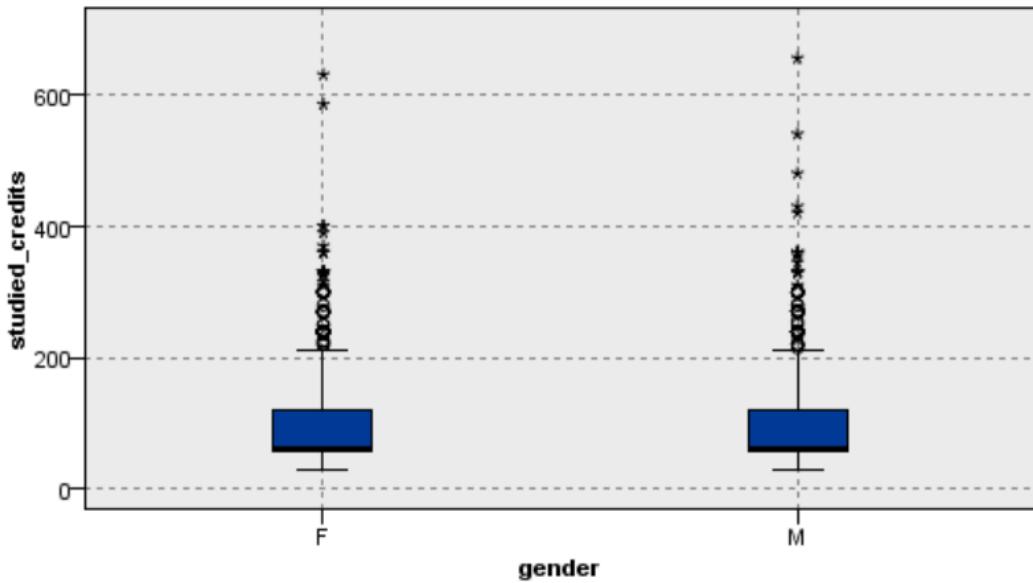


Figure 23: Boxplot of student_credits

Cela met en évidence l'existence des points atypiques dans la colonne student_credits. Cela devra être nettoyé plus tard. Il est également clair que le num_of_prev_attempts est une variable ordinaire, pas une variable continue

On visualise les variables catégorielles à savoir: *code_module*, *Code_Presentation*, *Gender*

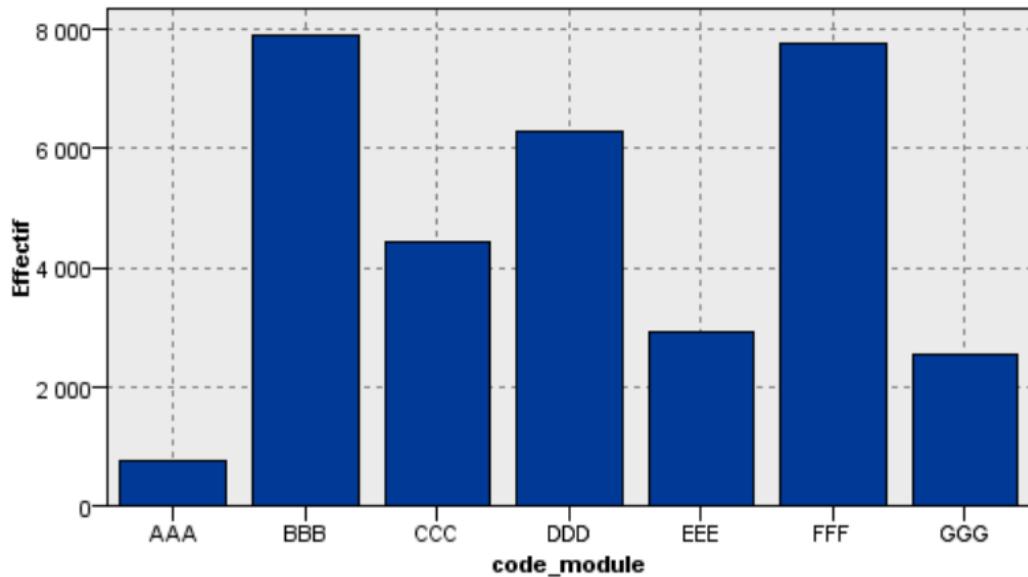


Figure 24: Modules codes

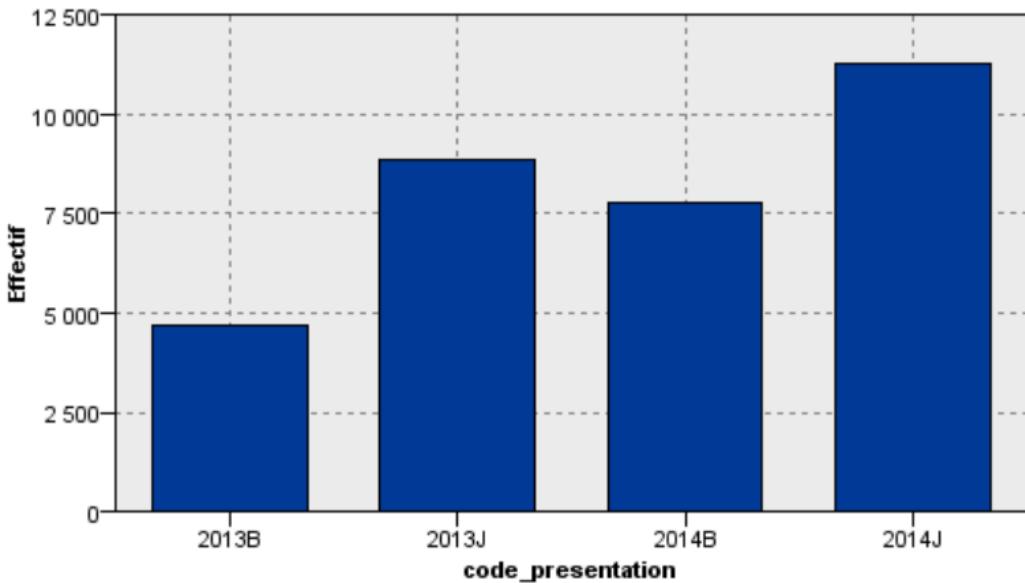


Figure 25: Years of course

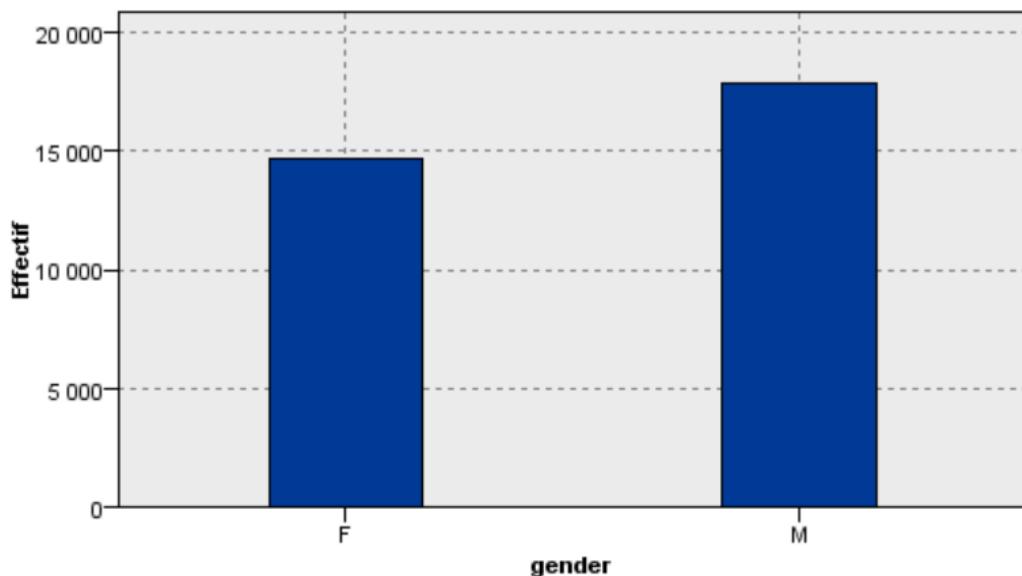


Figure 26: Gender

On constate que:

- Il existe 7 codes de module. Ces catégories doivent être condensées davantage lors du nettoyage. Le *code_presentation* pourrait être condensé en deux groupes d'années
- Plus de mâle étaient inscrits que de femelle.

Après, on a visualiser d'autres variables catégorielles à savoir: *num_of_prev_attempts*, *Disability*, *Age_band*

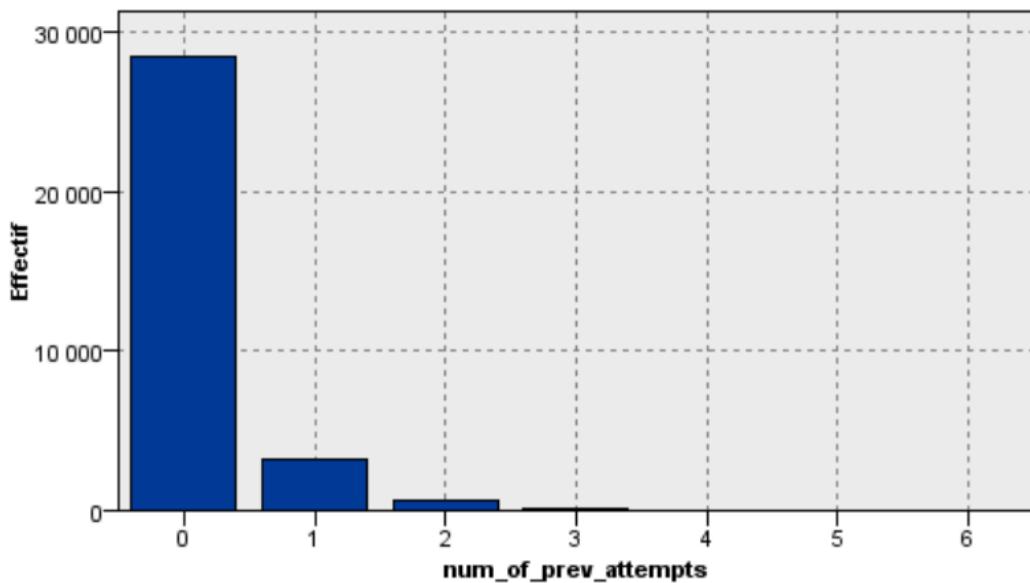


Figure 27: Number of Previous Attempts

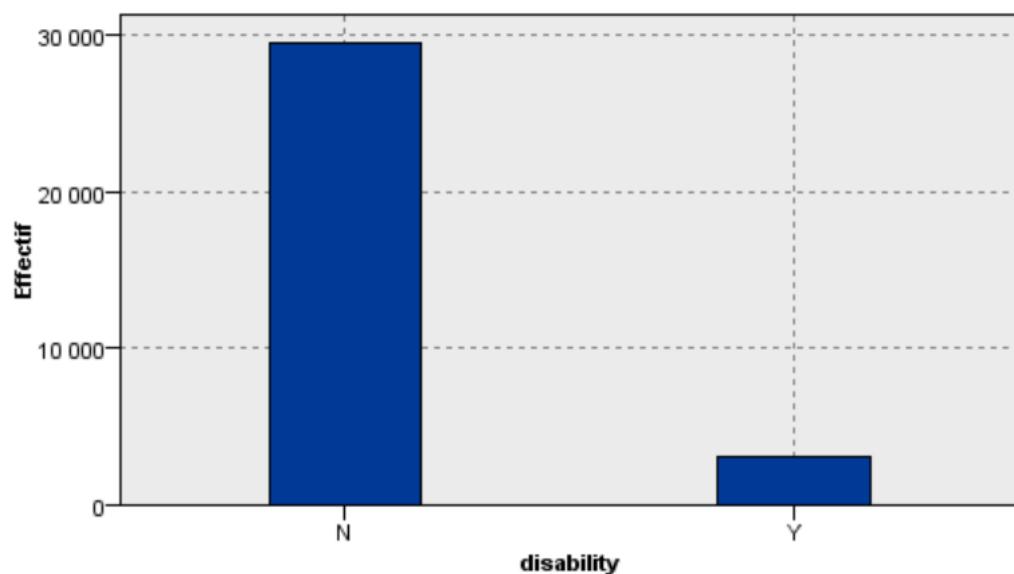


Figure 28: disability

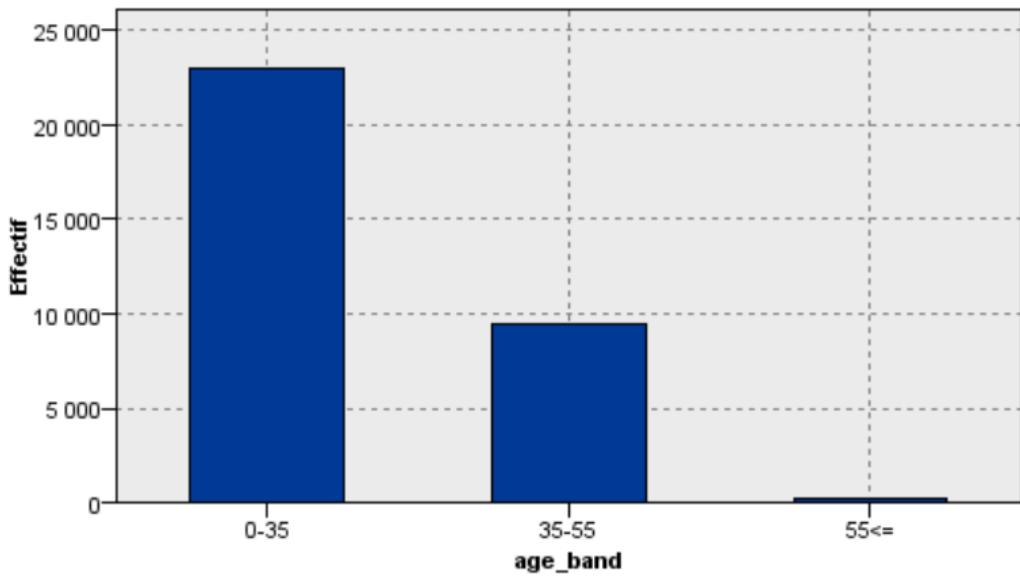


Figure 29: Age band

On remarque que:

- La grande majorité des étudiants ont terminé le cours à leur première tentative.
- La plupart étaient âgés de 35 ans et moins.
- La plupart des étudiants ont réussi le cours mais les abandons sont très élevés.
- Très peu est *disabled*

Ensuite, on a visualiser la région la plus commune d'où vient les étudiants.

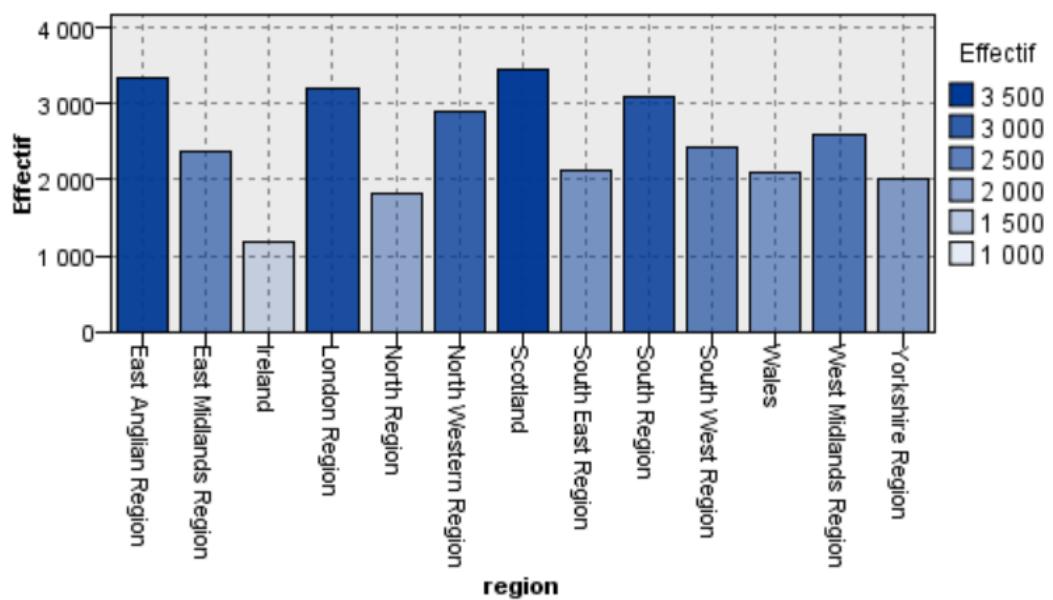


Figure 30: Most Common Region

Donc, L'Écosse avait un plus grand nombre d'étudiants, mais dans l'ensemble, l'Angleterre avait le plus grand nombre d'étudiants et l'Irlande le moins.

On a aussi chercher à savoir le niveau d'éducation le plus courant. Voici le graphe qui illustre cela:

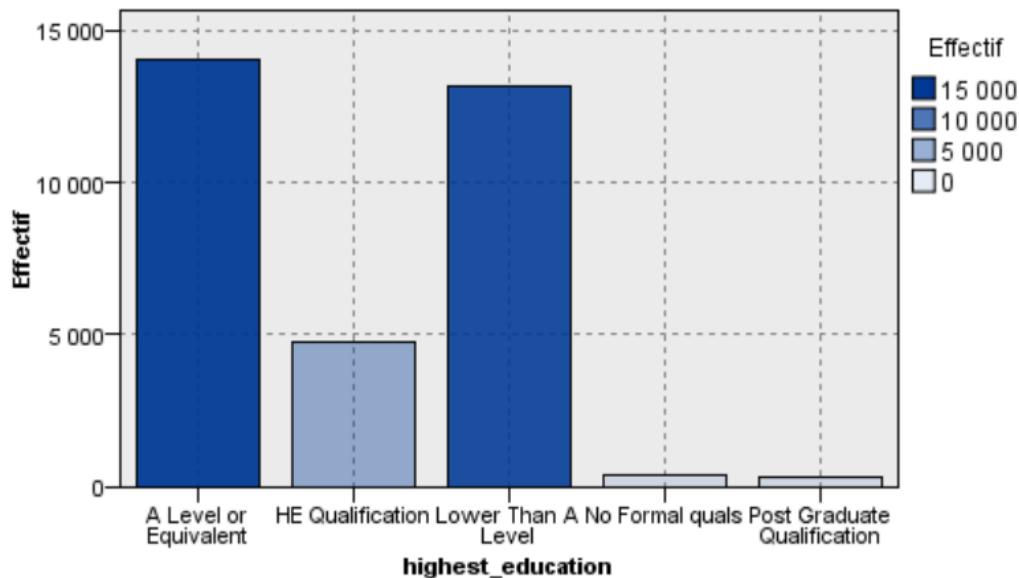


Figure 31: Most Common Education Level

Et par la suite, on a visualiser le status socio-économique des étudiants.

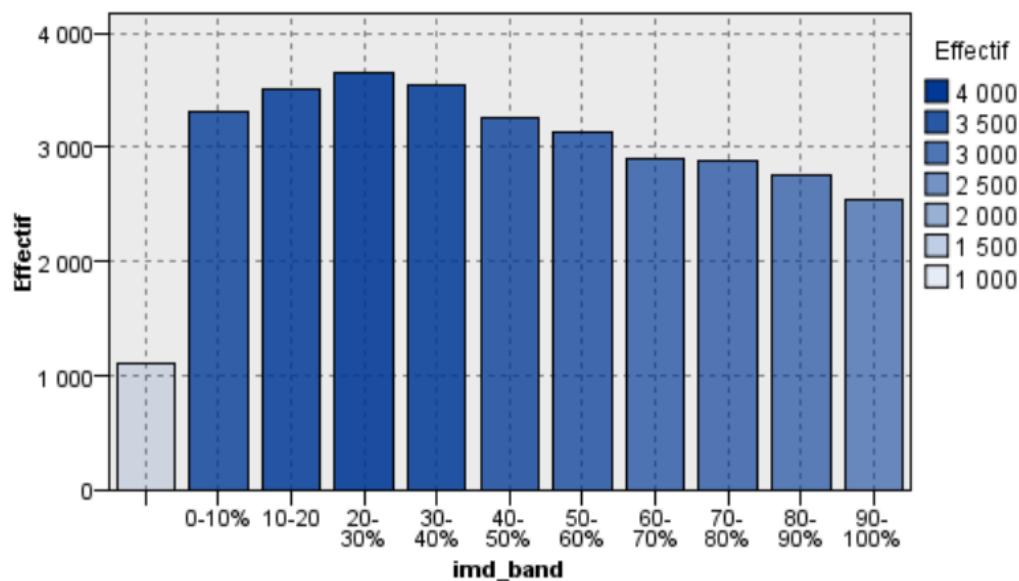


Figure 32: Socio-Economic Status

On remarque donc:

- Plusieurs étudiants avec un faible revenu étaient inscrits.
- Il semble y avoir quelques catégories redondantes dans la colonne *highest_education*. Cela devra être réglé lors du nettoyage.
- Pas beaucoup de variation dans les *imb_band* mais il y a trop de bandes donc cela devrait être condensé.

Enfin, on a représenté les résultats des étudiants pour avoir une vision clair sur leurs rendements.

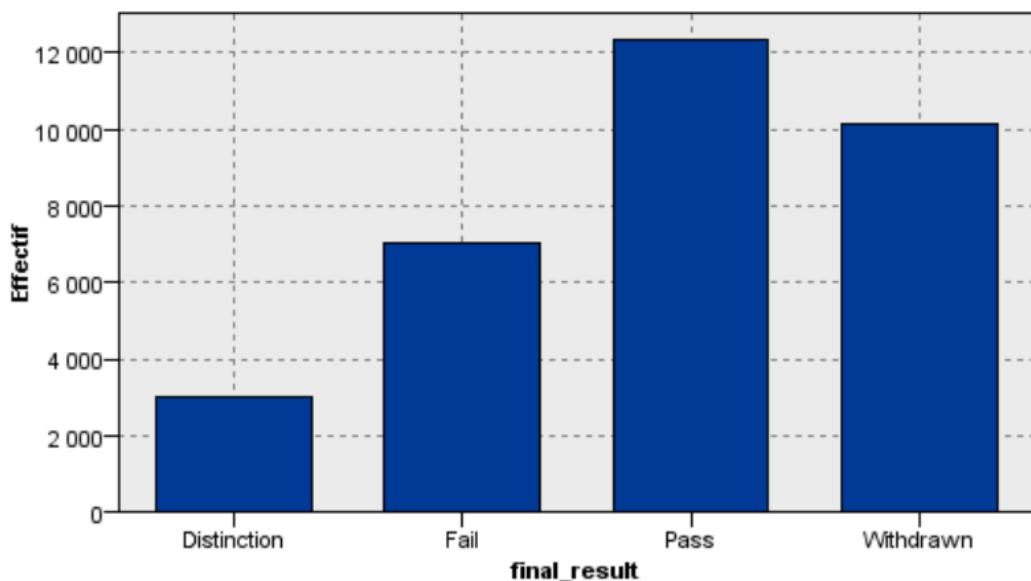


Figure 33: Student Outcome

on a bien constater que la plupart des étudiants ont réussi, mais les taux d'abandon et d'échec sont élevés.

Valeur	Proportion	%	Comptage
Distinction		9.28	3024
Fail		21.64	7052
Pass		37.93	12361
Withdrawn		31.16	10156

Figure 34: Statistic about Student Outcome

- Près de 52 % des étudiants ont abandonné ou échoué(*withdrew or failed*)
- Près de 21% des étudiants ont échoué(*failed*).
- 31% des étudiants ont abandonné leur courses.

2.3.6 Fichier StudentRegistration

Ce fichier contient des informations sur l'heure à laquelle l'étudiant s'est inscrit à la présentation du module. Pour les étudiants qui se sont désinscrits, la date de désinscription est également enregistrée.

	field1	code_module	code_presentation	id_student	date_registration	date_unregistration	reg_month	unreg_month
1	0 AAA	2013J		11391	-159.000	\$null\$ Sep	Jun	
2	1 AAA	2013J		28400	-53.000	\$null\$ Sep	Jun	
3	2 AAA	2013J		30268	-92.000	12.000 Sep	Oct	
4	3 AAA	2013J		31604	-52.000	\$null\$ Sep	Jun	
5	4 AAA	2013J		32885	-176.000	\$null\$ Sep	Jun	
6	5 AAA	2013J		38053	-110.000	\$null\$ Sep	Jun	
7	6 AAA	2013J		45462	-67.000	\$null\$ Sep	Jun	
8	7 AAA	2013J		45642	-29.000	\$null\$ Sep	Jun	
9	8 AAA	2013J		52130	-33.000	\$null\$ Sep	Jun	
10	9 AAA	2013J		53025	-179.000	\$null\$ Sep	Jun	

Figure 35: Chargement de la table StudentRegistration

Tout d'abord, on a commencé par faire plusieurs manipulations sur les données du fichier **StudentRegistration** pour bien les comprendre. Comme suit:

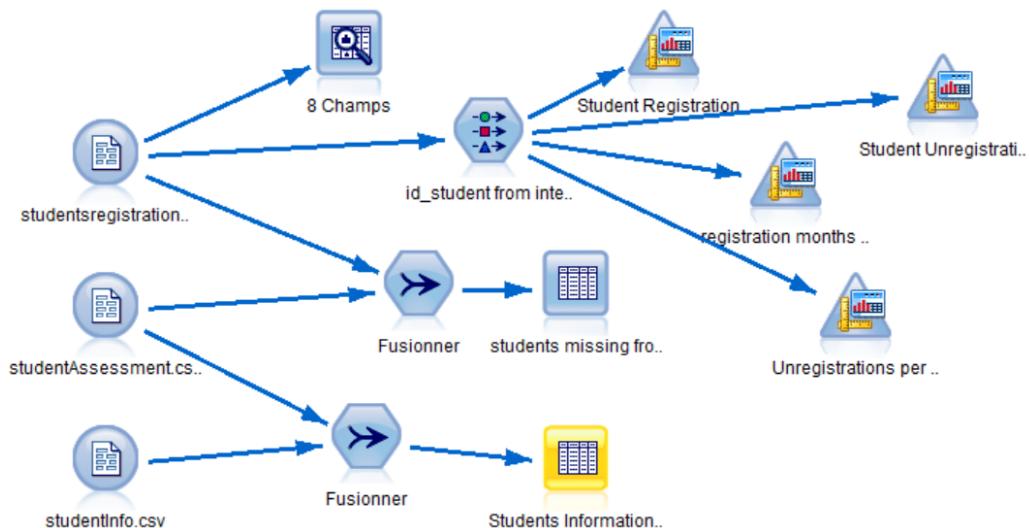


Figure 36: Schema StudentRegistration

On vérifie si tous les identifiants d'étudiants enregistrés dans les tables d'inscription sont enregistrés dans la table de résultats. On a utiliser des fusions.

students missing from the Results table (3 champs, 5 847 enregist...)

	id_student	reg_month	unreg_month
1	3733	Sep	Sep
2	23632	Sep	Sep
3	25629	Sep	Jan
4	25629	Sep	Sep
5	26269	Sep	Sep
6	26677	Sep	Sep
7	26915	Sep	Jun
8	27457	Sep	Sep
9	28149	Sep	Sep
10	28770	Sep	Jun
11	30268	Sep	Oct
12	32553	Sep	Sep
13	34694	Sep	Sep
14	35859	Sep	Jun
15	36842	Sep	Sep
16	38187	Sep	Oct
17	40333	Sep	Oct
18	40508	Sep	Oct
19	41489	Sep	Jun
20	43011	Sep	Sep

Figure 37: Students missing from the Results table

Il manque 5847 étudiants dans le tableau des résultats.

On vérifie s'il manque des étudiants du tableau d'informations sur les étudiants dans le tableau des résultats.

	id_student
5828	2685154
5829	2685825
5830	2687012
5831	2687644
5832	2687739
5833	2690077
5834	2690136
5835	2690719
5836	2692381
5837	2693772
5838	2693974
5839	2697608
5840	2697773
5841	2698109
5842	2698591
5843	2702660
5844	2707979
5845	2710343
5846	2710343
5847	2716795

Figure 38: Students Information table missing from the Assessment Results table

Il y a 5847 étudiants enregistrés dans le tableau des informations sur les étudiants manquants dans le tableau des résultats d'évaluation.

On vérifie qu'ils sont les même étudiants manquants dans les deux tables.

Maintenant, On cherche à inspecter la date d'enregistrement dans le cours.

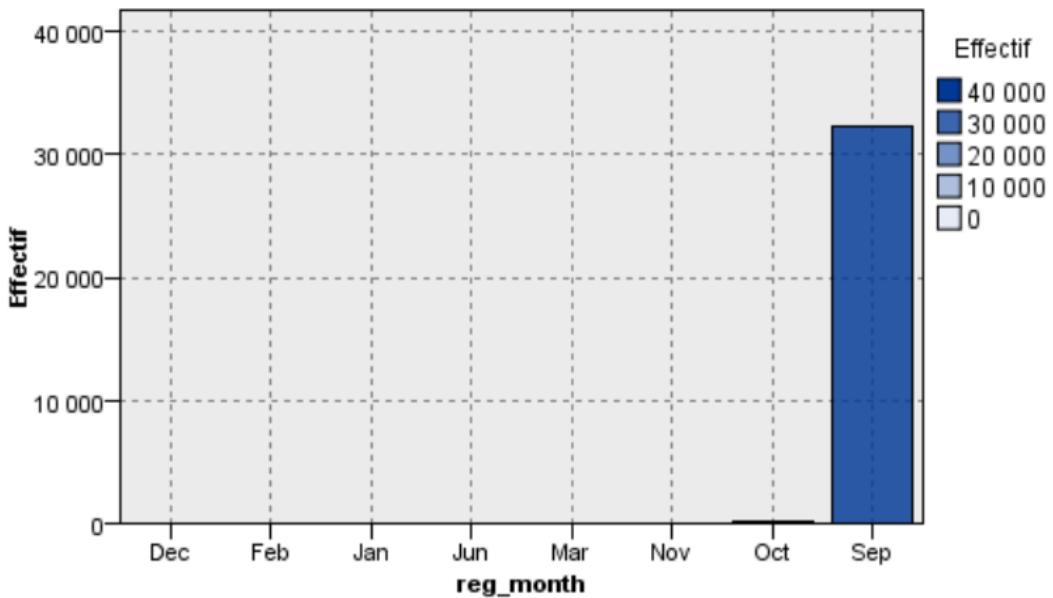


Figure 39: Student Registration Date (Month)

On remarque que 99% des inscriptions ont eu lieu en septembre.

Ensuite, on vérifier la date de désinscription. Et on remarque que 69% des étudiants se sont désinscrits en juin ce qui est logique car 31% des étudiants ont abandonné leur cours. La plupart des abandons se produisent au premier trimestre avec un nombre constant tous les deux mois de l'année. Le taux d'abondant le plus bas a été enregistré en mai.

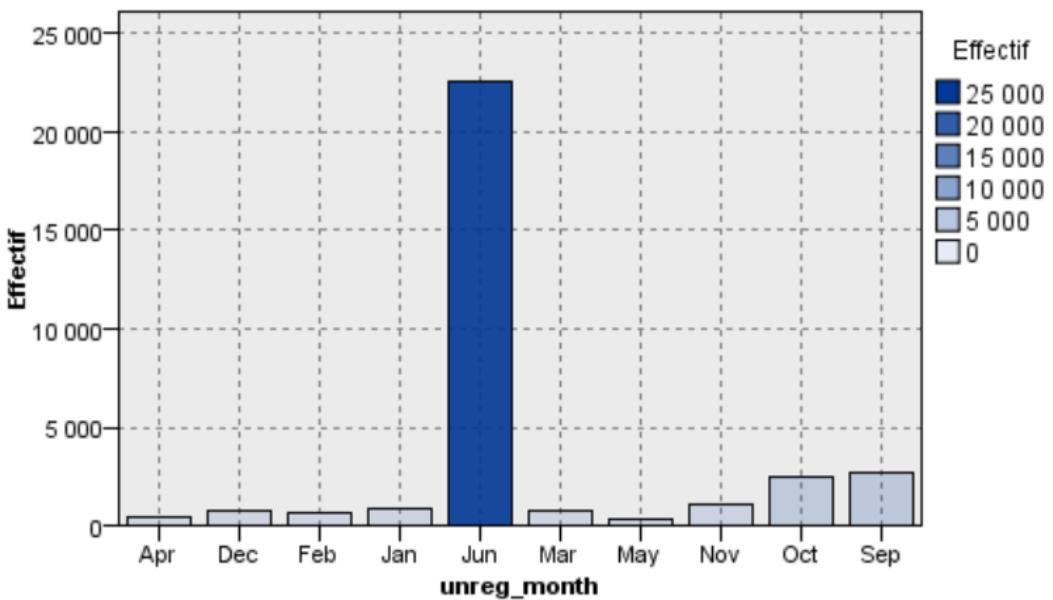


Figure 40: Student Unregistration Date (Month)

On veut vérifiez si les mois d'inscription varient selon les types de modules. Pour cela, on a

réaliser le graphe suivant:

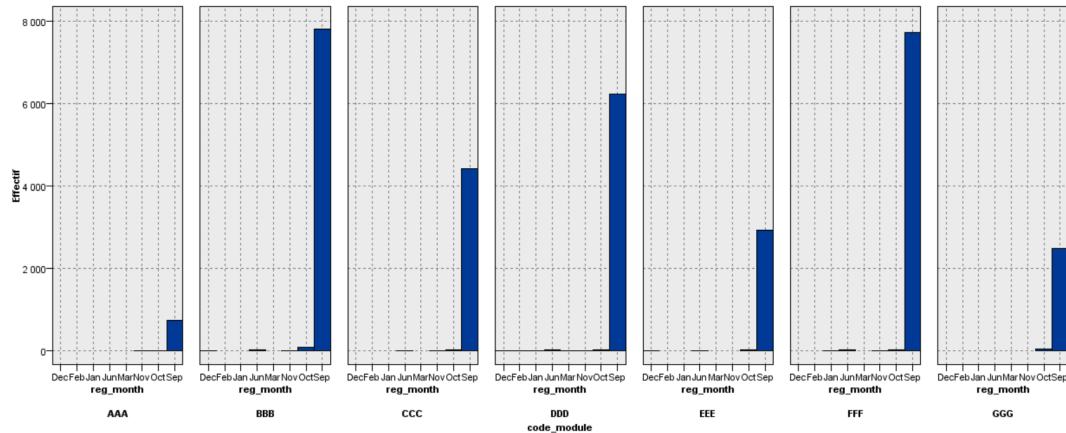


Figure 41: Registration months varied across module types

Seuls les modules BBB, DDD, FFF et GGG avaient des étudiants inscrits en octobre, mais c'était une minorité d'étudiants.

Puis, On vérifie si les mois de désinscription varient selon les types de module.

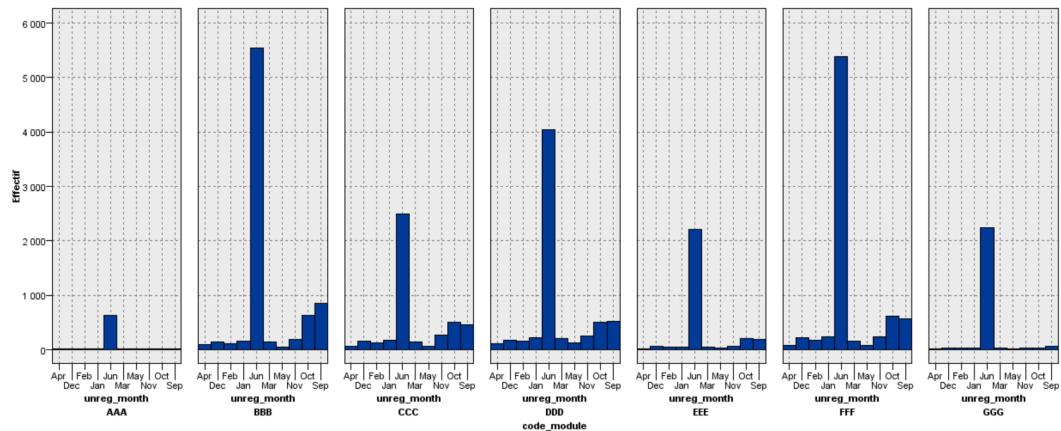


Figure 42: Unregistrations per Module per Month

Il semble que AAA n'ait pas de décrochages et que le module GGG en ait très peu.

2.3.7 Fichier studentVle

Le fichier studentVle.csv contient des informations sur les interactions de chaque étudiant avec le matériel du VLE.

	code_module	code_presentation	id_student	id_site	date	sum_click
1	AAA	2013J	28400	546652	-10	4
2	AAA	2013J	28400	546652	-10	1
3	AAA	2013J	28400	546652	-10	1
4	AAA	2013J	28400	546614	-10	11
5	AAA	2013J	28400	546714	-10	1
6	AAA	2013J	28400	546652	-10	8
7	AAA	2013J	28400	546876	-10	2
8	AAA	2013J	28400	546688	-10	15
9	AAA	2013J	28400	546662	-10	17
10	AAA	2013J	28400	546890	-10	1

Figure 43: Chargement de la table studentVle

Tout d'abord, on a commencer par faire plusieurs manipulations sur les données du fichier **studentVle** pour bien les comprendre. Comme suit:

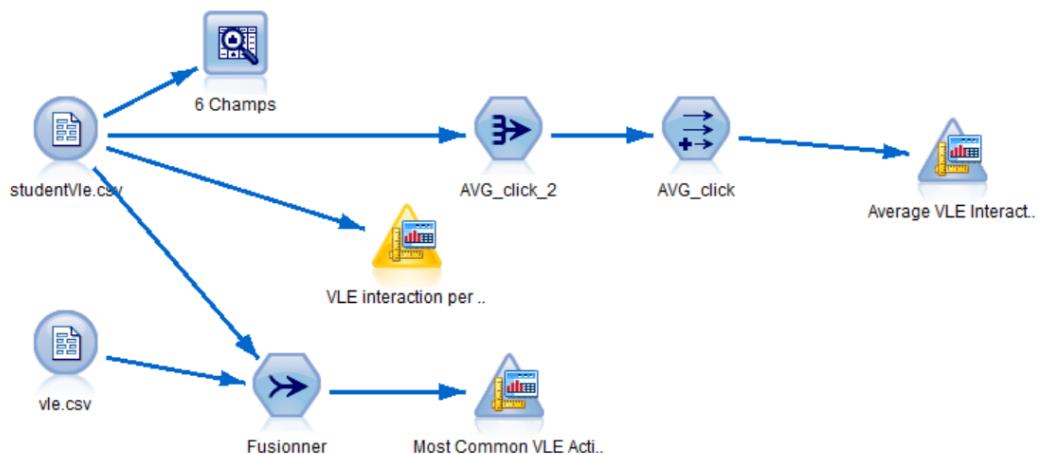


Figure 44: Schema studentVle

On visualise combien de matériel vle dans chaque module à l'aide du graphe suivant:

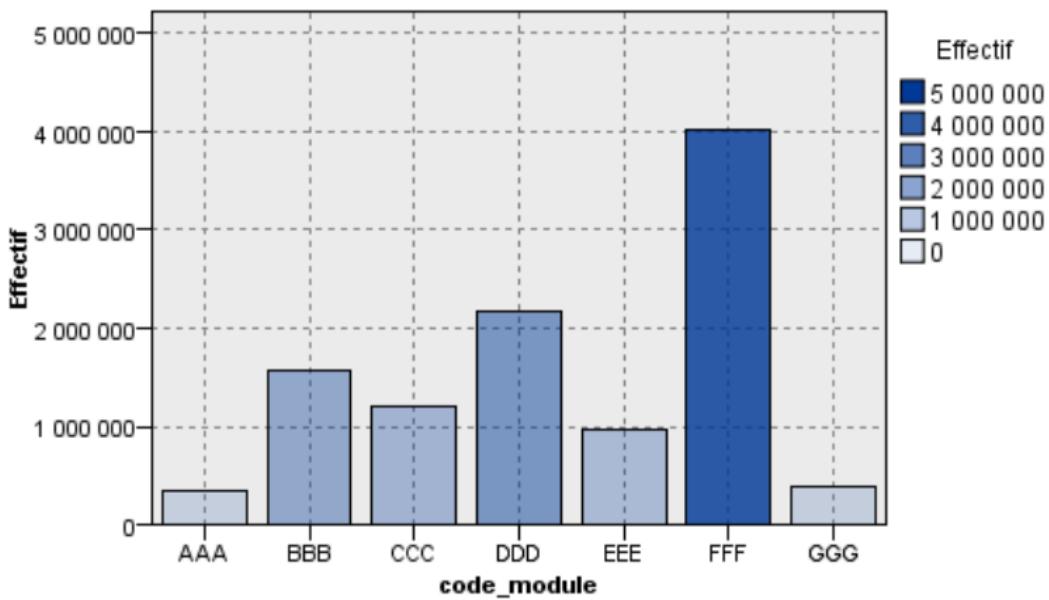


Figure 45: VLE interaction per Module

On remarque que les modules BBB, DDD, FFF peuvent avoir une charge de travail plus lourde car il y a plus d’interaction VLE que les autres modules.

On va utiliser la colonne pour indiquer la moyenne de clics par étudiant, et pouvoir la visualiser.

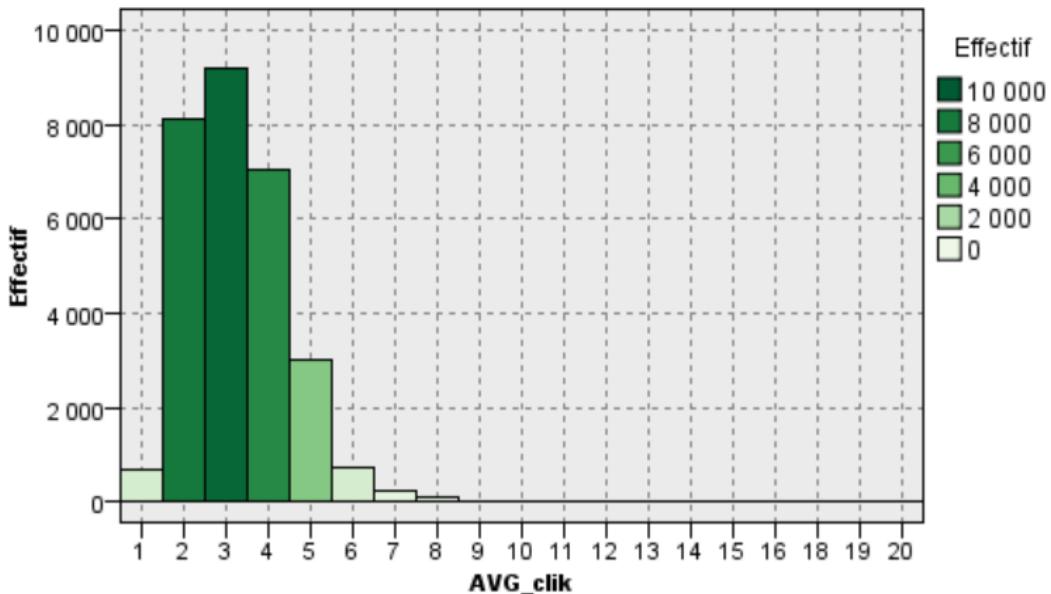


Figure 46: Average VLE Interaction

La plupart des étudiants ont cliqué sur le matériel trois fois par jour.

Et enfin, on présente les VLE Activités les plus courantes.

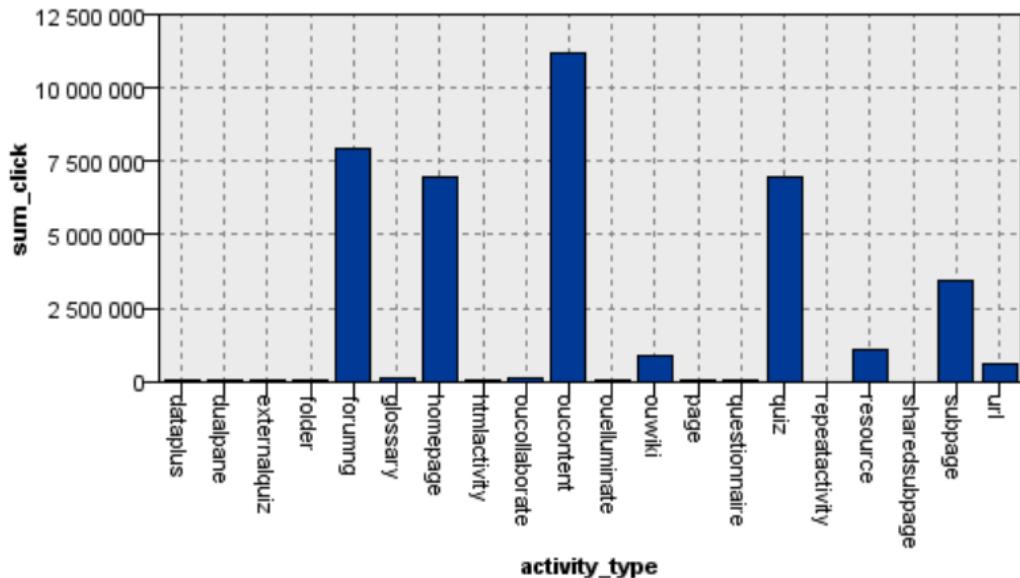


Figure 47: Most Common VLE Activity

2.4 Vérification de la qualité des données

On prend l'exemple des tables suivantes:

Qualité							
Audit		Champs complets (%) : 80 %		Enregistrements complets (%) : 99,9 %			
Champ	Mesure	Valeurs extrêmes	Extrêmes	Action	Attribuer une ent...	Méthode	% terminé
id_assessm...	Continue	0	0 Aucun	Jamais	Fixe		
id_student	Continue	8290	0 Aucun	Jamais	Fixe		
date_submitt...	Continue	24	38 Aucun	Jamais	Fixe		
is_banked	Continue	0	1909 Aucun	Jamais	Fixe		
score	Continue	2212	0 Aucun	Jamais	Fixe		

Figure 48: Qualité assessment

Qualité							
Audit		Champs complets (%) : 66,67 %		Enregistrements complets (%) : 17,61 %			
Champ	Mesure	Valeurs extrêmes	Extrêmes	Action	Attribuer une ent...	Méthode	% terminé
id_site	Continue	0	0 Aucun	Jamais	Fixe		1
code_module	Catégorielle	--	--	Jamais	Fixe		1
code_presen...	Catégorielle	--	--	Jamais	Fixe		1
activity_type	Catégorielle	--	--	Jamais	Fixe		1
week_from	Catégorielle	--	--	Jamais	Fixe		17,6
week_to	Catégorielle	--	--	Jamais	Fixe		17,6

Figure 49: Qualité VLE

Audit	Qualité	Annotations					
Champs complets (%) : 100 %		Enregistrements complets (%) : 100 %					
Champ	Mesure	Valeurs extrêmes	Extrêmes	Action	Attribuer une ent...	Méthode	% terminé
code_module	Catégorielle	--	--	Jamais	Fixe		1
code_presen...	Catégorielle	--	--	Jamais	Fixe		1
id_student	Continue	402518	0 Aucun	Jamais	Fixe		1
id_site	Continue	0	0 Aucun	Jamais	Fixe		1
date	Continue	0	0 Aucun	Jamais	Fixe		1
sum_click	Continue	68336	58289 Aucun	Jamais	Fixe		1

Figure 50: Qualité StudentVLE

Chapitre 3

3 Préparation des données

Ce chapitre sera consacré à la préparation des données qui est l'un des aspects les plus importants et les plus coûteux en temps en Data Mining. Dans cette section, on va voir la préparation et la fusion des données.

3.1 Intégration des données

Notre dataset est constituée de données réparties sur plusieurs fichiers, d'où la nécessité de les intégrer dans une seule Dataset. Ceci est réalisé par le biais de la fusion. Nous avons fait en premier lieu la fusion des tables StudentInfo, Courses, et studentRegistration par une jointure interne. La table obtenu est surnommé regCoursesInfo

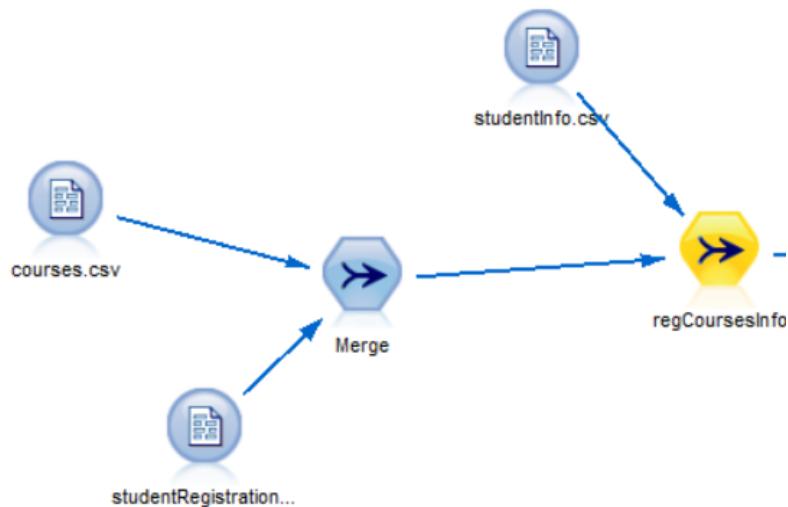


Figure 51: Première étape du fusionnement

Puis nous avons effectué une autre jointure interne des tables Assessemments et studentAssessements. La table obtenu est surnommé ass Results.

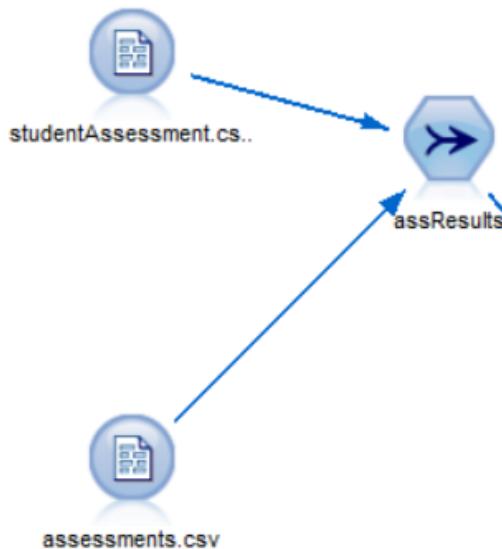


Figure 52: Deuxième étape du fusionnement

Après ajout des colonnes dérivés, nous avons effectué une jointure externe (gauche) de la table

regCourseInfo avec la table vle_interaction et assessments.

Le processus d'intégration est le suivant:

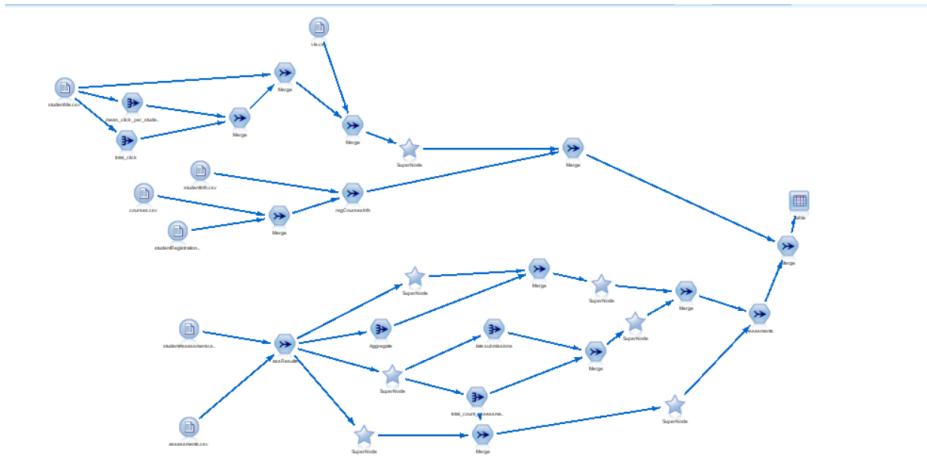


Figure 53: Processus du fusionnement

3.2 Creation de nouvelles colonnes

On a estimé important de créer plusieurs nouvelles colonnes offrant une meilleure compréhension et visualisation des données, notamment :

- **Weighted_Score**: afin que le poids total de tous les modules puisse être créé.
 - **late_rate**: pour indiquer le pourcentage de devoirs remis en retard.
 - **fail_rate** pour indiquer le pourcentage de devoirs non remis. On a posé le score de passage de 40%, et on a ainsi calculé le nombre de fails ($score < 40\%$). Après on divise le résultat sur le compte total des assessments.



Figure 54: Calcul de fail_rate

- AVG click pour indiquer le nombre moyen de clics par étudiant.

3.3 Sélection des données

Après avoir fusionné les données en une seule table, il est judicieux maintenant de minimiser les colonnes à étudier en ne sélectionnant que celles qu'on estime être intéressantes.

Au début, après avoir visualisé l'audit des données, on observe que les données des attributs 'num_of_prev_attempts', 'total_assessments' et 'total_late_submission' sont plus catégorielles que continues (figure de l'Histogramme de num_of_prev_attempts). En effet, les variables continues sont des variables ayant un nombre infini de valeurs entre deux valeurs, alors que les variables catégorielles n'ont qu'un nombre fini de groupes distincts. C'est pour cela qu'on a pensé à les abandonner.

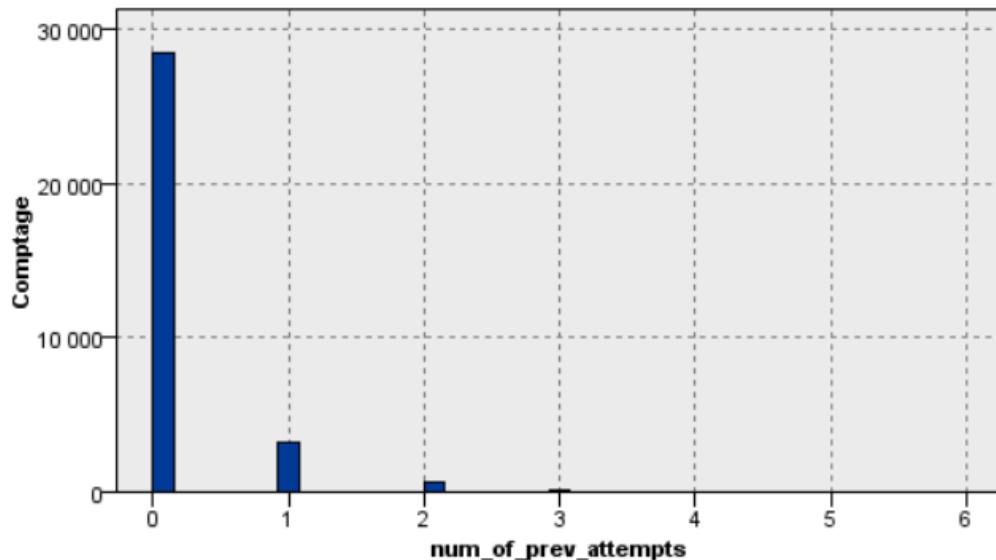


Figure 55: Histogramme de num_of_prev_attempts

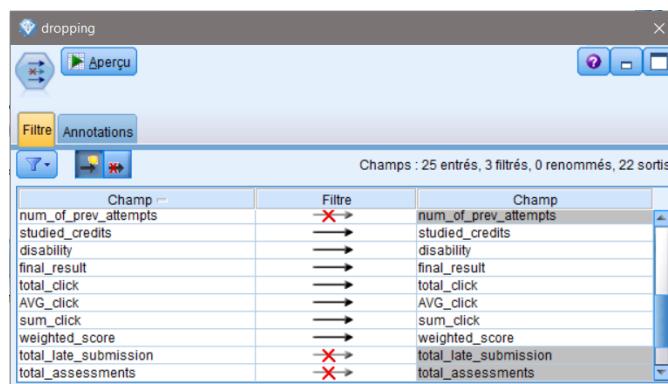


Figure 56: colonnes abondonnées 1

On a abouti au statistiques de weighted_score pour étudier ses corrélations avec les autres variables. C'est ainsi qu'il n'y a pas de corrélation entre module_presentation_length ou sum_click avec weighted_score, elles doivent donc être supprimées.

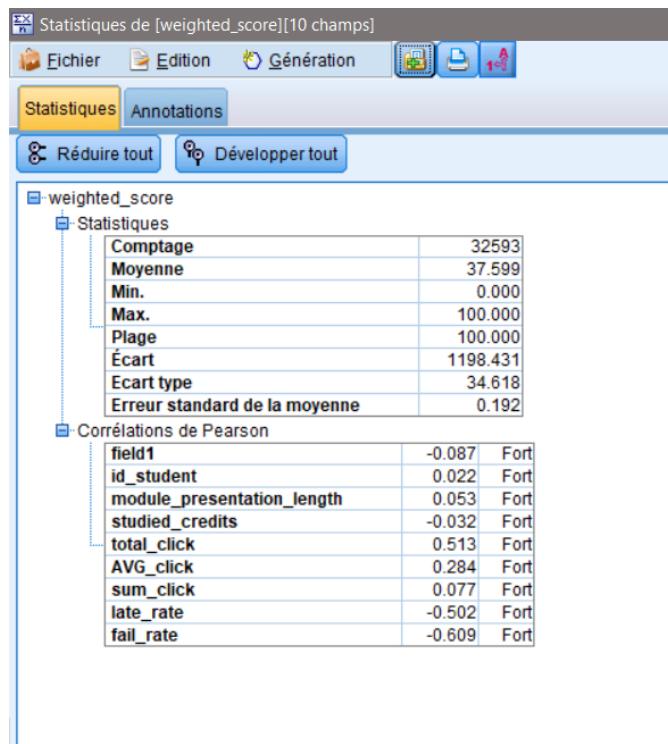


Figure 57: Corrélation de weighted_score

3.4 Nettoyage des données

Afin de renforcer la pertinence et l'intégrité des données, on a eu recours au nettoyage des données, afin d'éviter les erreurs et les incohérences pour obtenir des résultats plus précises.

3.4.1 Données manquantes

D'après ce qui est visualisé dans le chapitre de la compréhension des données, l'audit montre qu'il y a de nombreuses données manquantes dans quelques colonnes notamment : total_click, AVG_click, weighted_score, total_late_submission, late_rate, fail_rate, sum_clic. Le nettoyage de ces données manquantes se font par le biais du noeud remplacer :

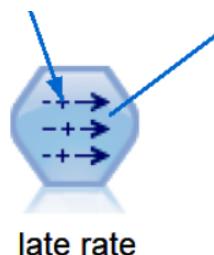


Figure 58: Noeud Remplacer

	total_click	Continue	6519
#	AVG_click	Continue	6519
#	sum_click	Continue	6519
#	weighted_sc...	Continue	8862
#	total_late_su...	Continue	6750
#	total_assess...	Continue	6750
#	late_rate	Continue	6750
#	fail_rate	Continue	6750

Figure 59: Données manquantes

- **weighted_score** : Remplaçant les valeurs nulles par 0 car ces étudiants n'ont fait aucune soumission.

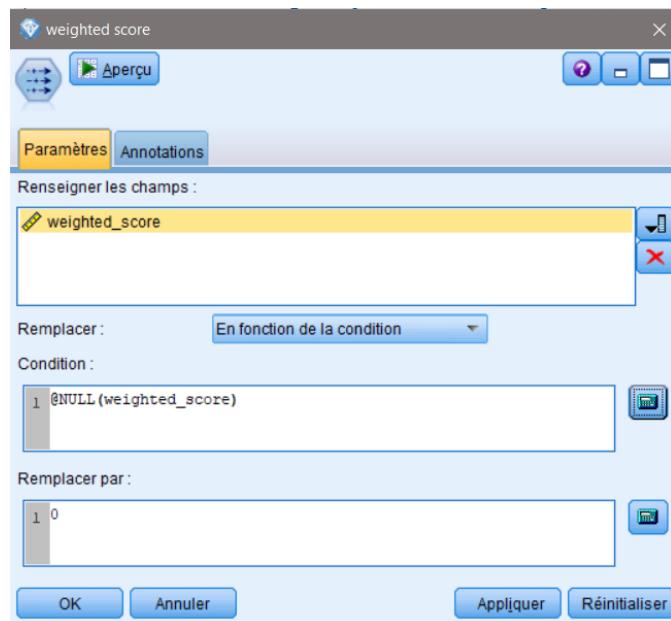


Figure 60: remplacer weighted_score

- **late_rate** : Remplaçant les valeurs nulles par 100% en retard parce qu'ils n'ont fait aucune soumission.

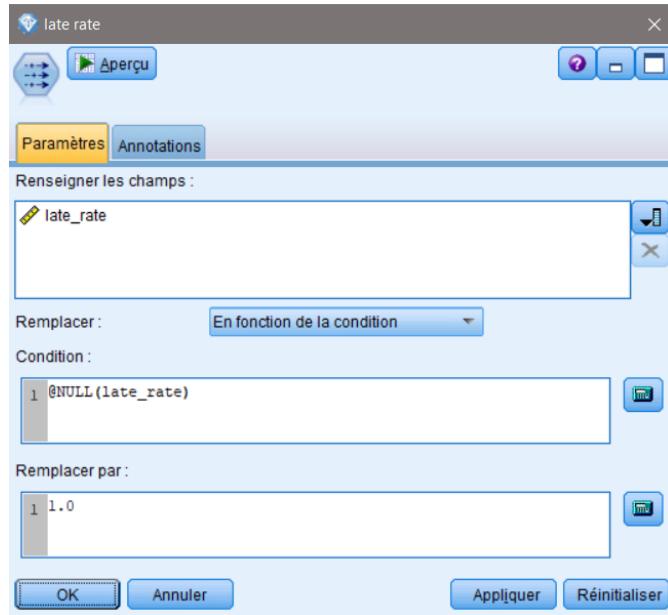


Figure 61: remplacer late_rate

- **total_late_submission** : Remplaçant les valeurs nulles par 100% en retard parce qu'ils n'ont fait aucune soumission.
- **fail_rate** : Remplaçant les valeurs manquantes par 100% d'échec parce qu'ils n'ont fait aucune soumission.
- **AVG_click** : Remplaçant les valeurs manquantes par 0, car ces étudiants n'ont pas interagi avec le VLE.
- **total_click** : Remplaçant les valeurs manquantes par 0, car ces étudiants n'ont pas interagi avec le VLE.
- **sum_click** : Remplaçant les valeurs nulles par 0, car ces étudiants n'ont pas interagi avec le VLE.
- **total_asssesments** : Remplaçant les valeurs manquantes par la médiane observé depuis l'audit.



Figure 62: Médiane de total_asssesments

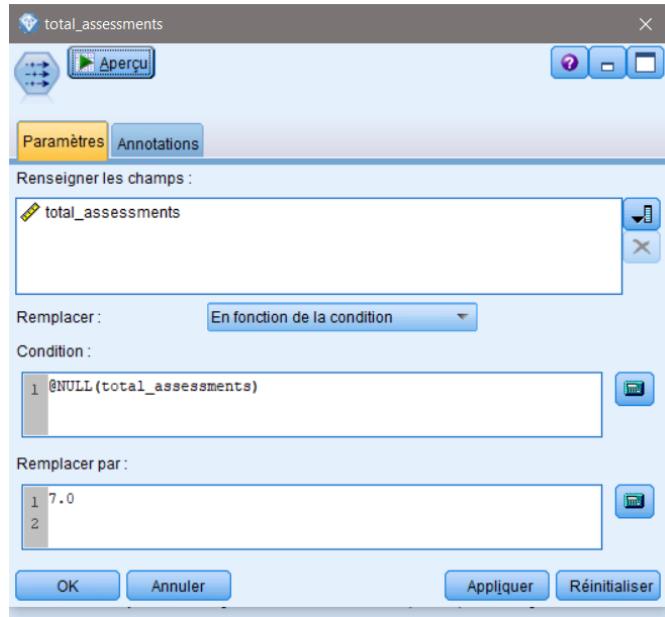


Figure 63: remplacer total_asssesments

3.4.2 Vérification de la distribution des données

- age_band :

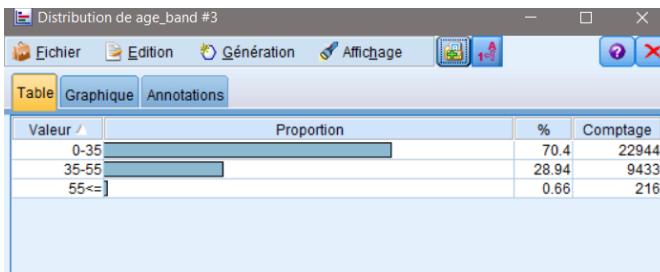


Figure 64: distribution de age_band

Les proportions des trois groupes ne sont pas du tout équitables. La plus dominante est celle de '0-35'. C'est ainsi qu'on a regroupé les groupes restants en '+35'.

On se retrouve avec la nouvelle distribution suivante:

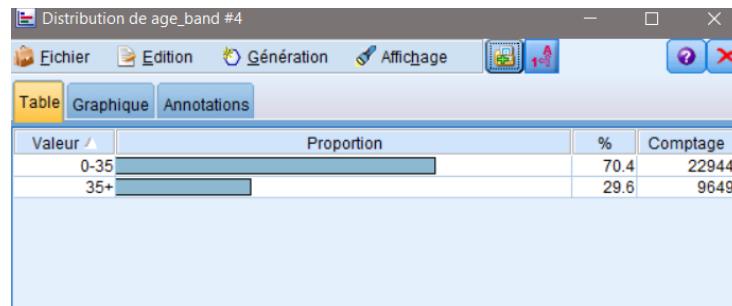


Figure 65: nouvelle distribution de age_band

- **region :**

Valeur	Proportion	%	Comptage
East Anglian Region		10.25	3340
East Midlands Region		7.26	2365
Ireland		3.63	1184
London Region		9.87	3216
North Region		5.59	1823
North Western Region		8.92	2906
Scotland		10.57	3446
South East Region		6.48	2111
South Region		9.49	3092
South West Region		7.47	2436
Wales		6.4	2086
West Midlands Region		7.92	2582
Yorkshire Region		6.15	2006

Figure 66: distribution de region

De même, les proportions des treize groupes ne sont pas du tout équitables et de petites proportions. Une simple recherche a permis de regrouper plusieurs régions en un seul. Par exemple regrouper 'East Anglian Region' et 'East Midlands Region' en 'East UK'.

On se retrouve avec la distribution suivante:

Valeur	Proportion	%	Comptage
East UK		23.98	7816
North UK		25.95	8459
South UK		19.35	6308
West UK		30.71	10010

Figure 67: nouvelle distribution de region

3.4.3 Variable cible

On a créé une variable cible qu'on a nommé 'Dropout'. Celle-ci traite les étudiants qui se sont retirés comme des dropouts, c'est-à-dire qui ont pour final_result la valeur 'Withdrawl'. La nouvelle colonne aura pour valeur 1 s'il a abandonné et 0 sinon.

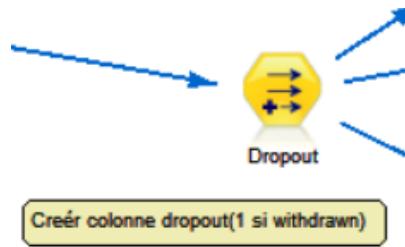


Figure 68: Noeud Calculer

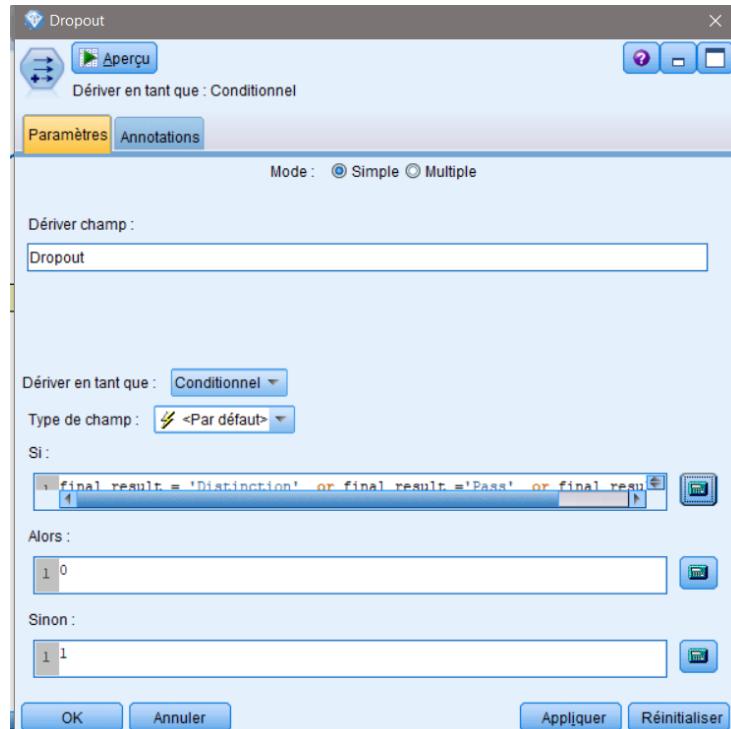


Figure 69: Dropout -1

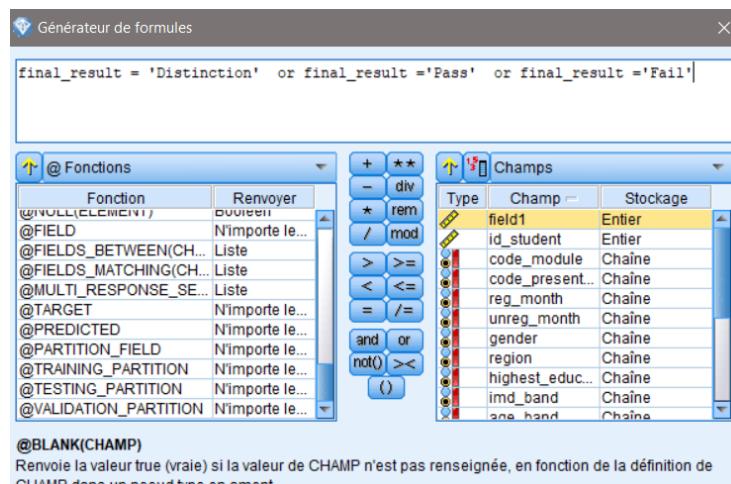


Figure 70: Dropout -2

La réalisation dans SPSS Modeler s'affiche comme ci-dessus. On a regroupé les noeuds de même objectifs (Données manquantes, distributions) pour faciliter la visualisation.

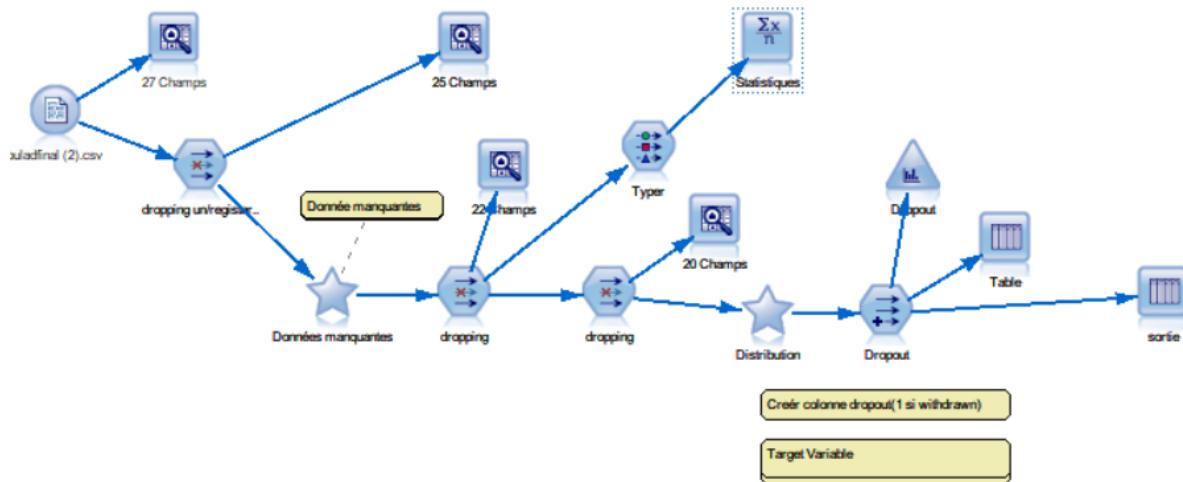


Figure 71: nettoyage dans SPSS

On donne l'exemple du super noeud regroupant les noeuds de remplacement des champs manquants des colonnes.

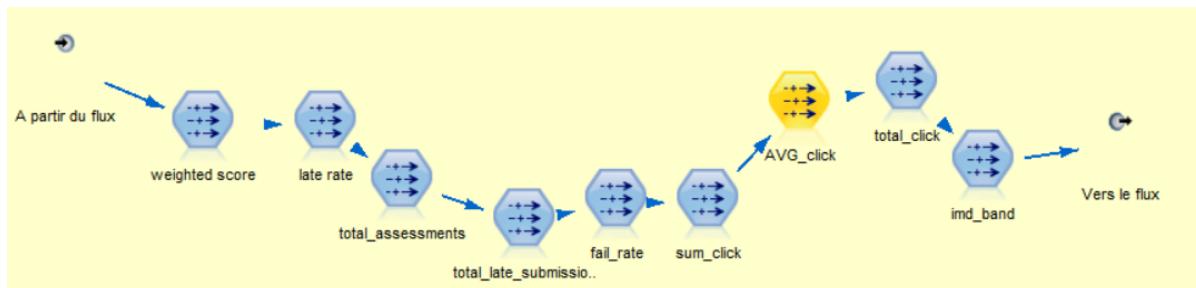


Figure 72: Super Noeud Données manquantes

Chapitre 4

4 Modélisation

Dans ce présent chapitre, on détaillera l'étape de la modélisation. Elle sera consacrée aux sélections de ses techniques et la construction du modèle.

La quatrième étape du CRISP-DM est la modélisation où on utilise l'une des techniques de Machine Learning. Elle comprend le choix, le paramétrage et le test des différents algorithmes. Les résultats commencent à éclaircir la problématique posée lors de la compréhension du métier.

4.1 Sélection des techniques de modélisation

Pour choisir le modèle qui pourra répondre le mieux aux objectifs du projet, il faut prendre en compte les critères suivants :

- Les types de données préparés.
- Les objectifs d'exploration de données
- Les exigences de modélisation particulières.

Comme l'objectif du projet est de prédire l'abandon des cours, on a affaire à un problème de classification supervisé. Le but de la classification supervisée est d'affecter un nouvel objet à une classe d'un ensemble donné de classes en fonction des valeurs d'attributs de cet objet et d'un ensemble d'apprentissage. Il existe plusieurs méthodes de classification supervisée notamment : KNN, les arbres de décision, Support Vector Machine (SVM), Réseaux de neurones ...

Nous avons opté de travailler avec trois de ces méthodes et les comparer pour voir laquelle permet les résultats les plus pertinentes :

- Les réseaux neuronaux
- Réseau bayésien
- Arbre de décision à l'aide de la détection automatique d'interaction Chi_Square

4.1.1 Les réseaux neuronaux

Les réseaux de neurones sont constitués de différentes couches de nœud, contenant une couche en entrée, une ou plusieurs couches cachées et une couche en sortie.

Les réseaux neuronaux utilisent des données de formation pour apprendre et améliorer leur précision au fil du temps, ils constituent de puissants outils permettant de classer et de regrouper rapidement les données. [1]

4.1.2 Réseau bayésien

Un réseau bayésien est un modèle graphique probabiliste de représentation des connaissances sur un domaine incertain où chaque nœud correspond à une variable aléatoire et chaque arête représente la probabilité conditionnelle pour les variables aléatoires correspondantes. [2]

4.1.3 Arbre de décision

Les arbres sont un ensemble de stratégies de résolution de problèmes consistant à diviser pour mieux régner qui utilisent des structures arborescentes pour prédire la valeur d'une variable de résultat. Les algorithmes basés sur des arbres se distinguent. Ce sont des modèles prédictifs avec une plus grande précision, une compréhension simple.

Chi-Square est une mesure statistique pour trouver la différence entre les nœuds enfants et parents. Pour calculer cela, nous trouvons la différence entre les nombres observés et attendus de la variable cible pour chaque nœud et la somme au carré de ces différences standardisées nous donnera la valeur du chi carré. [3]

4.2 Génération d'une conception de test

Afin de valider notre modèle et comparer la performance entre les trois modèles choisis, nous avons utilisé la méthode de validation croisé grâce au noeud "Partitionner" de SPSS. En effet, la validation croisée est une technique d'évaluation des modèles d'apprentissage-machine via la formation de plusieurs modèles d'apprentissage-machine sur des sous-ensembles des données d'entrée disponibles et via leur évaluation sur le sous-ensemble complémentaire des données. [4]

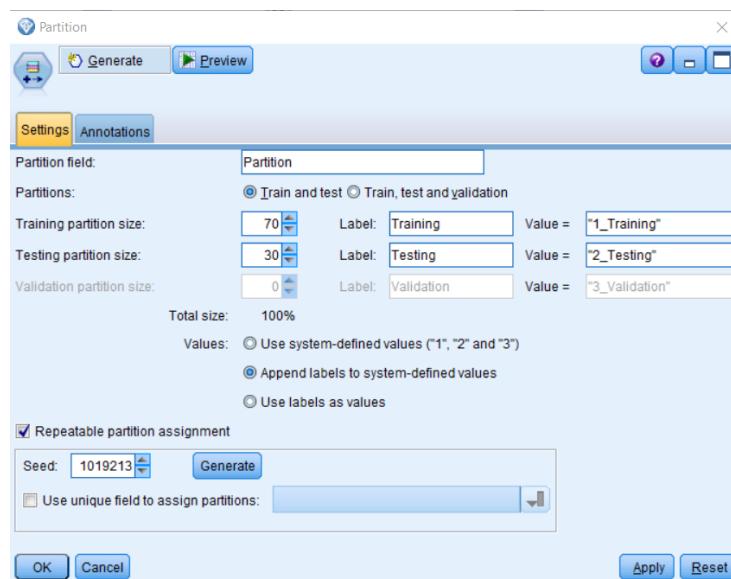


Figure 73: Implémentation de la validation croisée avec SPSS

4.3 Construire le modèle

Nous avons commencé par ajouter le noeud "typer" pour définir les données entrantes du modèle.

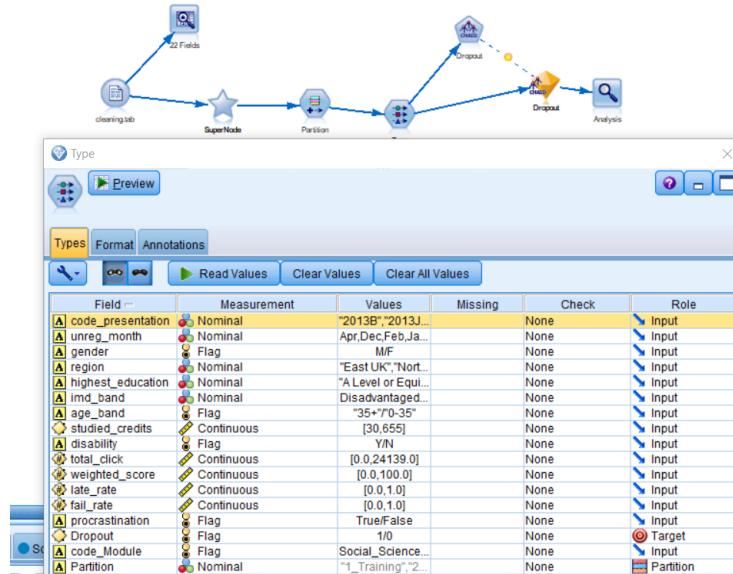


Figure 74: Choix de la colonne Dropout comme cible et les autres colonnes comme entrée

4.3.1 Implémentation de l'arbre de décision

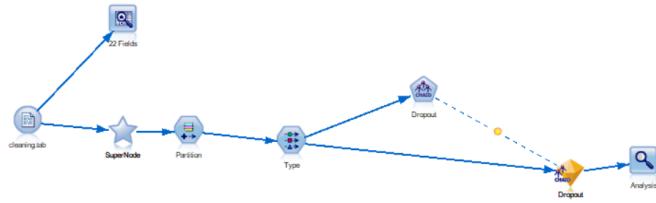


Figure 75: Arbre de décision par SPSS

On obtient alors la représentation graphique de la procédure de la classification sous forme de l'arbre de décision suivant:

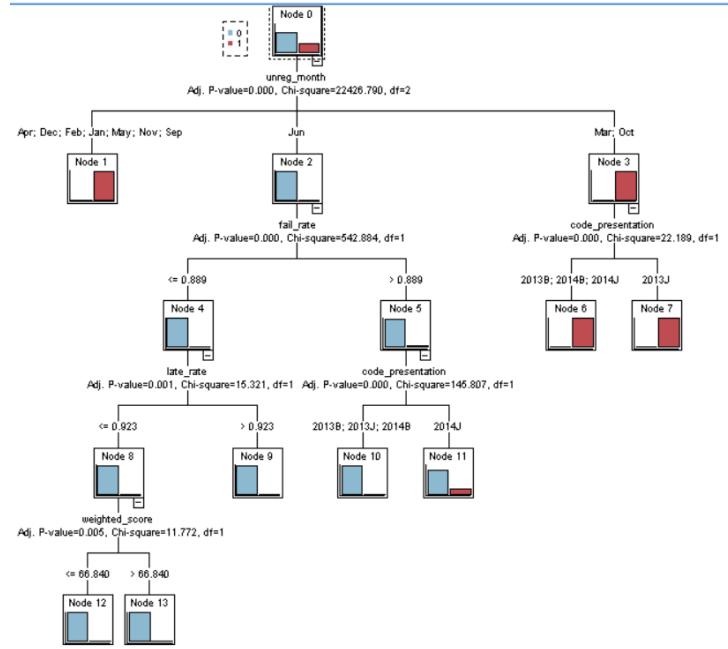


Figure 76: Arbre de décision

4.3.2 Implémentation du réseau bayésien

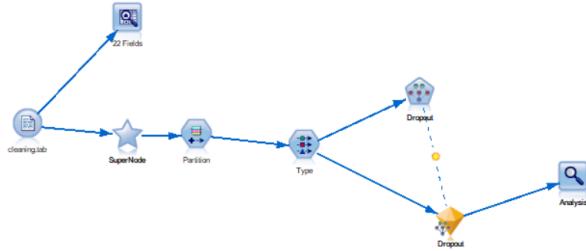


Figure 77: Réseau bayesien par SPSS

Le résultat est un modèle graphique probabiliste représentant l'ensemble des attributs sous la forme d'un graphe tel que chaque noeud comprend une table de probabilité conditionnelle.

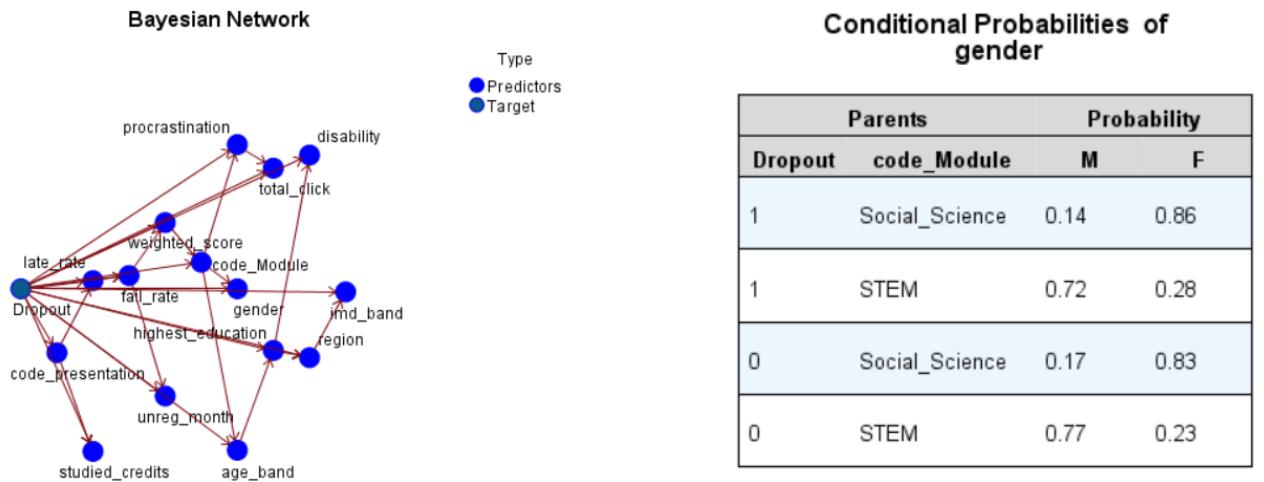


Figure 78: Réseau bayésien

4.4 Évaluation de la performance des modèles

Nous calculons la précision en divisant le nombre de prédictions correctes par le nombre total d'échantillons. Le résultat nous indique que notre modèle a atteint une précision de 99.67% pour les arbres de décisions, de 82.88% pour les réseaux de neurones et 99.64% pour les réseaux bayésiens.

Comparing \$R-Dropout with Dropout				
'Partition'	1_Training	2_Testing		
Correct	22,713	99.66%	9,770	99.67%
Wrong	78	0.34%	32	0.33%
Total	22,791		9,802	

Results for output field Dropout				
Comparing \$N-Dropout with Dropout				
'Partition'	1_Training	2_Testing		
Correct	18,885	82.86%	8,124	82.88%
Wrong	3,906	17.14%	1,678	17.12%
Total	22,791		9,802	

Figure 79: Performances respectives de l'arbre de décision et réseaux de neurones

Results for output field Dropout				
Comparing \$B-Dropout with Dropout				
'Partition'	1_Training	2_Testing		
Correct	22,715	99.67%	9,767	99.64%
Wrong	76	0.33%	35	0.36%
Total	22,791		9,802	

Figure 80: Performance du réseau bayésien

Chapitre 5

5 Evaluation et déploiement

Ce chapitre vise à évaluer les résultats obtenus qui est la cinquième étape du CRISP-DM pour ensuite expliquer le déploiement.

5.1 Evaluation

Durant les quatre premières phases du modèle de processus CRISP-DM, nous avons établi les objectifs du projet suite à la compréhension du métier pour ensuite explorer les données et construire des modèles prédictifs de classification supervisée. La prochaine étape est naturellement l'évaluation des résultats. Nous évaluerons ainsi non seulement les modèles créés, mais également le processus que nous avons utilisé pour les créer et leur potentiel d'utilisation pratique.

L'évaluation vise à vérifier les connaissances obtenues afin de s'assurer qu'ils répondent aux objectifs formulés au début. Elle contribue aussi à la décision de déploiement du modèle et son amélioration. A ce stade, on teste notamment la robustesse et la précision des modèles obtenus.

Il s'agit de l'étape finale du processus. Elle consiste en une mise en production pour les utilisateurs finaux des modèles obtenus. Son objectif : mettre la connaissance obtenue par la modélisation, dans une forme adaptée, et l'intégrer au processus de prise de décision.

La phase d'évaluation comprend trois tâches, notamment :

- Évaluation des résultats
- Déterminer les prochaines étapes

5.1.1 Évaluation des résultats

Afin d'évaluer les résultats pour les objectifs du projet, il faut résumer les résultats par rapport aux critères de réussite que nous avons établi lors de la phase de compréhension. Vu le taux élevé des abandons des cours en lignes pendant les dernières années, l'étude de ce phénomène s'est avérée être très intéressant pour prédire la probabilité de ces abandons qui saura être utile pour maintenir et encourager les activités d'apprentissage des élèves. En effet, en se basant sur la construction des modèles choisis et leur finalité, on peut observer que l'objectif a été bien atteint avec plus de pertinence de résultats lors de l'utilisation des arbres de décision. C'est ainsi qu'on peut estimer que le défi établi au début de la conception et la compréhension du projet dont la prédiction de l'échec scolaire des étudiants.

5.1.2 Déterminer les prochaines étapes

On peut extraire d'après les résultats des modèles des informations très intéressantes qu'on pourra exploiter pour la prise de décision ainsi pour encourager les cours en ligne. On pourra donc étudier les régions où le taux des abandons est le plus élevé et motiver les élèves de cette région le plus.

5.2 Déploiement

Le déploiement peut aller, selon les objectifs, de la simple génération d'un rapport décrivant

les connaissances obtenues jusqu'à la mise en place d'une application, permettant l'utilisation du modèle obtenu, pour la prédiction de valeurs inconnues d'un élément d'intérêt.

De façon générale, la phase de déploiement de CRISP-DM comprend deux types d'activité :

- Planification et surveillance du déploiement des résultats
- Exécution des tâches de synthèse, telles que la production d'un rapport final et la révision du projet

6 Conclusion

Ce projet avait pour objet la prédiction des abandons des élèves inscrits en cours en ligne (MOOC). La mise en point de ce projet a été faite par le biais de la méthode CRISP-DM. Ce rapport trace les différentes étapes parcourues lors de l’élaboration du projet. Il s’agissait de bien comprendre le marché et déterminer les objectifs du projet tout en évaluant la situation actuelle. Pour ensuite passer à la compréhension et la préparation des données qui englobe le nettoyage de ceci et l’équilibrage des distributions. Les modèles ont été construit en se basant sur trois méthodes de la classification supervisée notamment les arbres décisions, les réseaux de neurones, et les réseaux bayésiens. En final, nous avons évalué les modèles et les objectifs atteints.

Ce projet a été pour nous l’occasion d’approfondir nos capacités de conception et nos acquis dans les concepts de datamining dont l’application du CRISP-DM qu’on estime essentiels pour le domaine de la Business Intelligence et la Data Science.

References

- [1] <https://www.ibm.com/fr-fr/cloud/learn/neural-networks>.
- [2] [https://www.sciencedirect.com/topics/mathematics/bayesian-network#:~:text=A%20Bayesian%20network%20\(BN\)%20is,corresponding%20random%20variables%20%5B9%5D](https://www.sciencedirect.com/topics/mathematics/bayesian-network#:~:text=A%20Bayesian%20network%20(BN)%20is,corresponding%20random%20variables%20%5B9%5D).
- [3] <https://www.analyticsvidhya.com/blog/2021/05/implement-of-decision-tree-using-chaid/>.