Cas d'étude : PACES

Date de rédaction : 9 novembre 2015
Scénarios : S2 et S3
Auteurs de l'analyse : Samy Foudil, Laura Dupuis, Muriel Ney, Nadine Mandran
Période de l'analyse : printemps 2014 puis été-automne 2015

La problématique posée pour l'analyse :

PRINTEMPS 2014 (données 2011-2012) :
- Typologie d'étudiants en fonction de leur profil et de leur classement dans les séances de tutorat et au concours (**S2**)
- Evolution temporelle de leurs résultats en fonction des 4 types d'étudiants déterminés auparavant (**S3**)

ETE-AUTOMNE 2015 (données 2012-2013) :
Evolution de l'activité des étudiants (en préparation au concours de médecine) au cours de l'année (**S3**)
L'activité sera mesurée par (1) la présence et les résultats aux tests QCM (séances de tutorat) et (2) la participation aux FLQ (formulation en ligne des questions et votes).
On regardera aussi l'impact du résultat au premier concours (fin du semestre 1) sur l'activité du semestre 2.
Objectif : Construire, enrichir et explorer le fichier de données dans le but de caractériser les étudiants par l'évolution de leurs notes, de leurs questions et de leur vote d'une part. D'autre part, prédire le comportement de l'étudiant en vue d'appliquer des rétroactions adaptées dans le cas où ce dernier présente des difficultés.
L'évolution d'un étudiant est représentée par :

- les courbes d'évolution de la note de chaque matière (4 matières au S1 et 4 au S2) et du rang = au total 16 séries chronologiques
- les courbes d'évolution du nombre de questions posées = 8 séries
- les courbes d'évolution du nombre de votes = 8 séries

Description du stockage des données:

En 2015 : Stockage sur Google Drive à accès limité pour partage des données avec les partenaires du LIUM

Plateformes/outils utilisés:

Points forts	Points faibles
sous Google drive : Synchronisation, simplicité, robustesse	sous Google drive : Ethique (données accessibles par Google…)

sous UT : sauvegarde des données sur le serveur d'UT	

Autre solution : UT ou OwnCloud (sur un serveur local à Grenoble ?)

Questions en suspend pour les trois plateformes citées : gestion des très gros fichiers ? trous de sécurité ?

Description des pré-traitements:

Plateformes/outils utilisés:

Première étape (Laura Dupuis) :
Programmation VBA sous excel pour fusionner des fichiers. Et SQL ?
Mise en place d'une procédure de contrôle de qualité des données comprenant entre autre :
- mise en évidence des doublons (un même étudiant/date avec deux résultats)
- traitement des données manquantes
- mise en évidence des pertes lors de la fusion des fichiers
- vérification cohérence entre fichiers brut et transformé
- enrichir les données (ajouter valeurs issues de divers fichiers)

Création des variables assiduité, notes QCM, rang QCM, nb+, nb-

Deuxième étape (Samy Foudil) :
Programmation Python

Création du fichier final avec les nouvelles variables : sequence, deltaNote, delatRang, penteNote, penteRang, min/cycle, max/cycle, moyenne/cycle, nbrQuest/vote, résultats concours (12 variables), activité

Variables de temps : semestre, cycle, matière, sequence

Variables construites	Description
	Identifiant de l'étudiant
	Identifiant de l'étudiant
	Sexe étudiant

	Code numérique du type de Bac
	Code alphanumérique
	Code numérique indiquant la situation professionnelle des parents.
	Il s'agit de l'orientation de l'étudiant. (ME : Medecine, MA : Maïeutique, OD : Odontologie, PA : Pharmacie)
Profil Orientation	Code numérique de l'orientation de l'étudiant
	Groupe de l'étudiant
	Date de naissance de l'étudiant
	Nombre d'inscritption de l'étudiant
Semestre	Semestre dans lequel s'inscrit la séquence (Voir colonne séquence)
Cycle	Cycle dans lequel s'inscrit la séquence
	Matière de la séquence en cours
	Date de la séquence en cours
Sequence	Variable importante contenant le libellé de chaque séquence S1_C1_BCH correspond à la séquence BCH du Cycle 1 du semestre 1.
	Note de l'étudiant au qcm de la séquence en cours
	Rang de l'étudiant sur la séquence en cours (sur nombre de présent dans cette séquence)
Nbr quest	Nombre de questions posées par l'étudiant sur la séquence en cours.
Assiduite	Variable 0-1 désignant le taux de présence de l'étudiant aux séquences. 0 toujours absent, 1 toujours présent.
	Resultat concours de l'étudiant code alpha (ADM,AJ,EX, etc) rang (sur nombre d'étudiant présent).
Note_Dessous / Note_Dessus	Nombre de notes en dessous de la moyenne et de notes au dessus de la moyenne de la promo entière
Changement_Pente_Note	Variable qui indique la variation des notes de l'étudiant dans une série de 3 notes (matière + cycle + semestre) On a deux valeurs "BOSSE" ou "CREUX"
Changement_Pente_Rang	Comme CPN, sur le rang.
Delta_Note	Calcul de la différence entre la note de la séquence en cours et de la note de la séquence précédente dans le but de mettre en évidence l'évolution de l'étudiant. La variable peut évoluer entre une valeur numérique négative ou positive selon l'évolution de l'étudiant.
Delta_Rang	Comme Delta_Note, on calcul la différence de rang entre la séquence en cours et la séquence précédente.

Distance_Note	Calcul de la distance de la note par rapport à un referentiel de niveau (Moyenne, Moyenne générale, médiane ? etc)
	Calcul de la distance par rapport à la limite d'admissibilité

PROGRAMME

Programme : extraire les données, objet étudiant.
Le programme proposé ici est conçu en Python, langage multi-plateforme et adaptable. La plupart des fichiers de données étant des fichiers Excel, le langage C# peut s'avérer plus approprié. ier ? Cependant, comme expliquer plus haut, la plateforme OrangeUT fournit un environnement de développement intéressant pour le traitement de données via des scripts python. Ce programme n'est pas une application, il a été conçu pour répondre au besoin de construire un fichier exploitable rapidement.

Num_UJF : la première partie du code consiste en l'extraction des numéros identifiants uniques des étudiants (Num UJF). A partir desquelles on génère un tableau d'objet *étudiant* possédant plusieurs attributs (décrit plus bas) permettant le stockage, par étudiant, des informations nécessaires. Le concours de médecine de Grenoble dispose d'un numéro identifiant indépendant de l'université : numéro PACES. Il est important d'associer le numéro UJF avec le numéro PACES. Nous avons au final un tableau d'objet étudiant contenant Num_UJF et Num_Pace ainsi qu'un "idt" indiquant l'instance de l'objet dans le tableau.

Importance du fichier Concordance : la suite du programme repose entièrement sur ce fichier. Les numéro UJF qui ne sont pas instanciés seront ignorés dans le traitement des fichiers qui suivent.

Parcours des fichiers et extraction ciblée.
On parcours chaque fichier en localisant l'identifiant UJF ou PACES. On localise les variables pertinentes : date de naissance, sexe, note, etc… Pour chaque identifiant, on vérifie qu'il soit instancié dans le tableau d'objet étudiant créé plus tôt. Cette méthode d'extraction a le mérite de s'assurer que les données enregistrées correspondent bien aux identifiant de l'étudiant et permet d'ignorer ou de renvoyer une erreur si le programme rencontre un identifiant (UJF, PACES) inconnu.

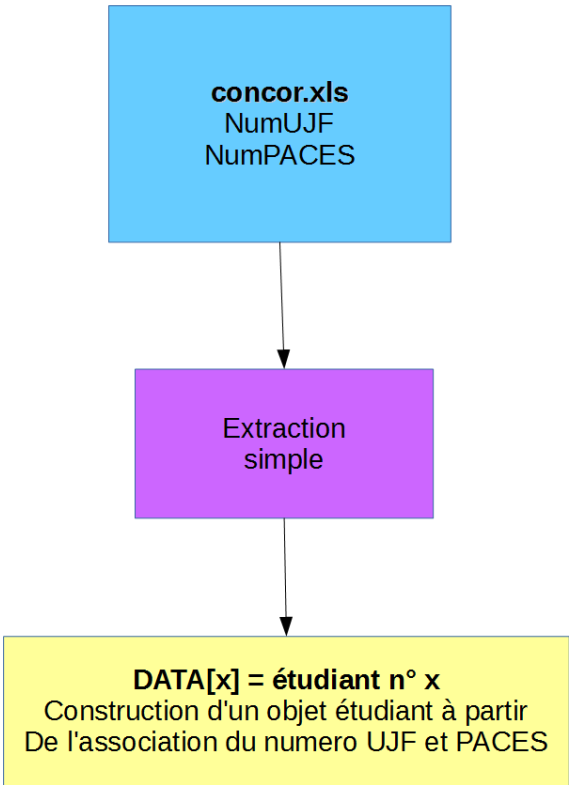
Calculer les variables d'enrichissements

Programme : les variables construites sont calculées à partir d'autres variables. Certaines sont calculées à l'instant de l'extraction de ces dernières, d'autres sont calculées après construction du fichier.

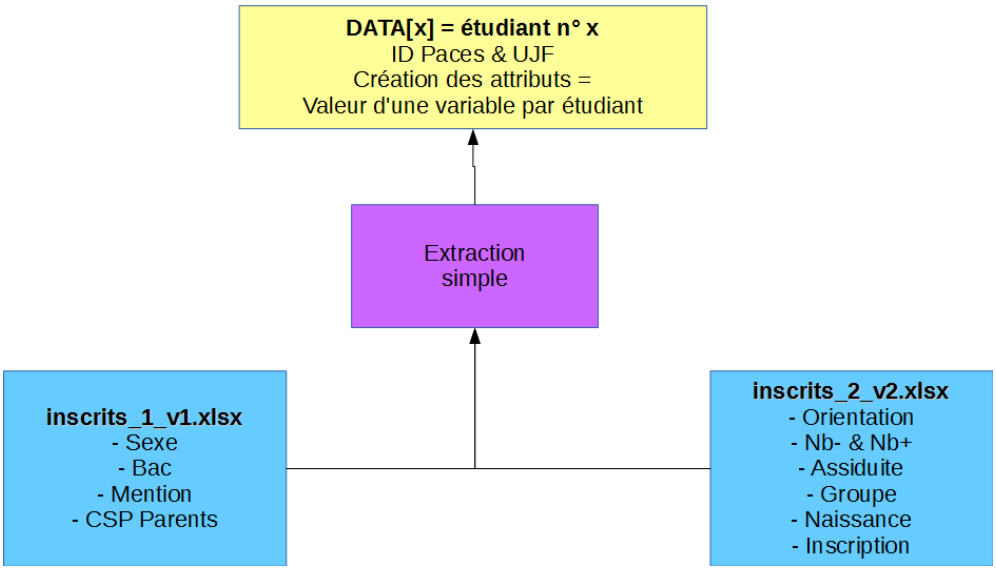
Post traitement : certains variables sont construites après traitement du programme sur le fichier de données. Il s'agit d'un ensemble de formule Excel. La variable séquence est construite de cette manière là.

Programme : Variables DELTA et Changement Pente, ces variables seront calculées lors de la procédure de récupération des notes par matière par étudiant. Ces notes sont stockées dans un attribut de l'objet étudiant (en cours de traitement) de type dictionnaire.

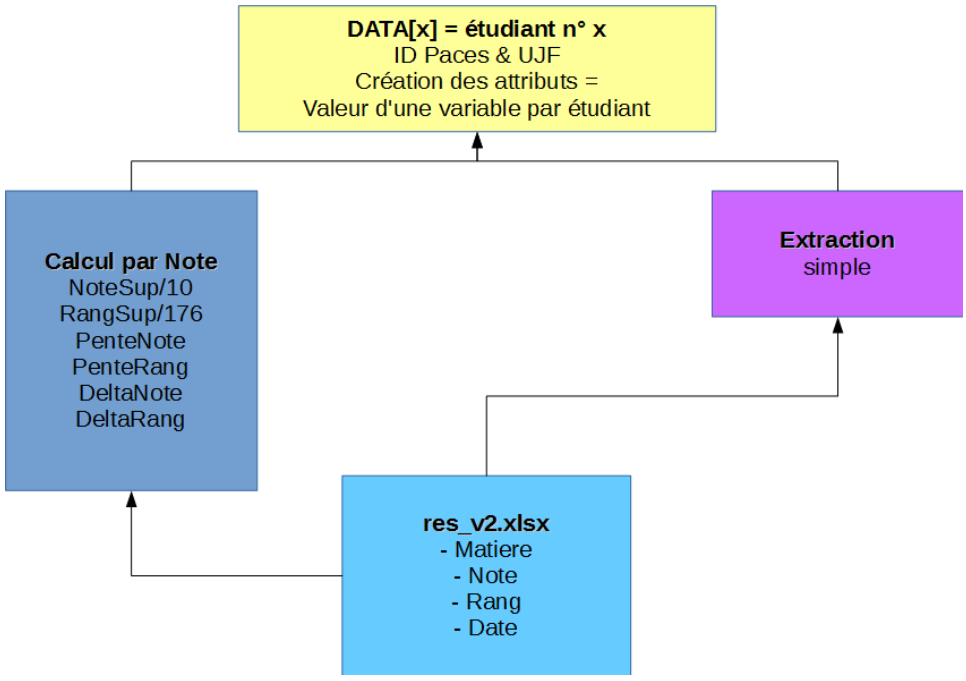
Mise en forme du fichier final
La mise en forme du fichier final consiste à nettoyer le fichier des caractères spéciaux non supportés par UnderTracks. Intégration dans UnderTracks
Détail des étapes de traitement :
Chaque schéma représente une étape du traitement du programme. Ce peut être des modules indépendants transformable en opérateur dans OrangeUT.



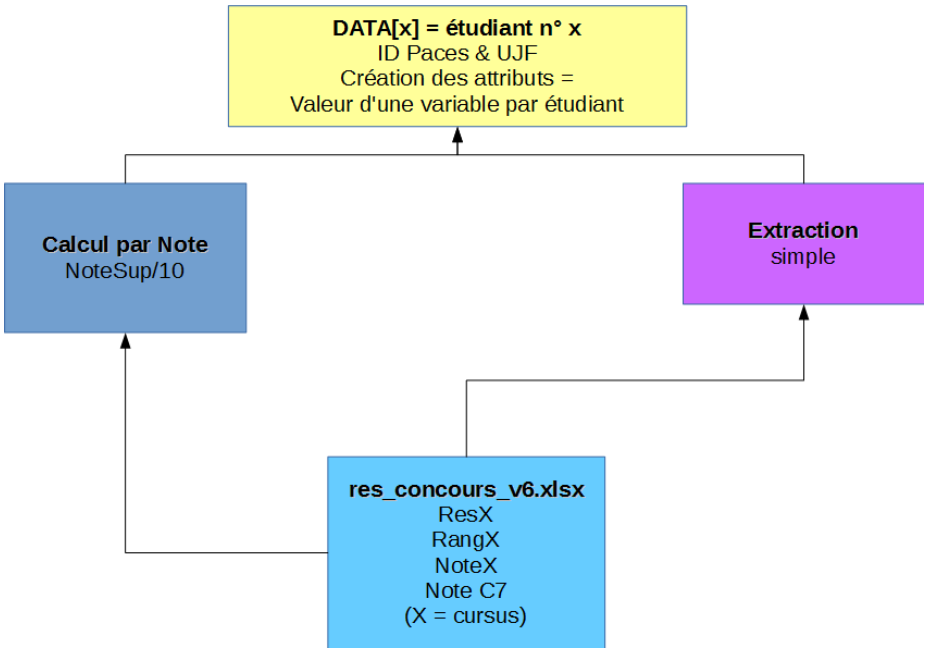
Construction d'un objet étudiant à partir de l'association du numéro UJF et PACES. On génère les attributs qui constitueront les variables du fichier finales en leur affectant des valeurs par défauts (0) en début de traitement.



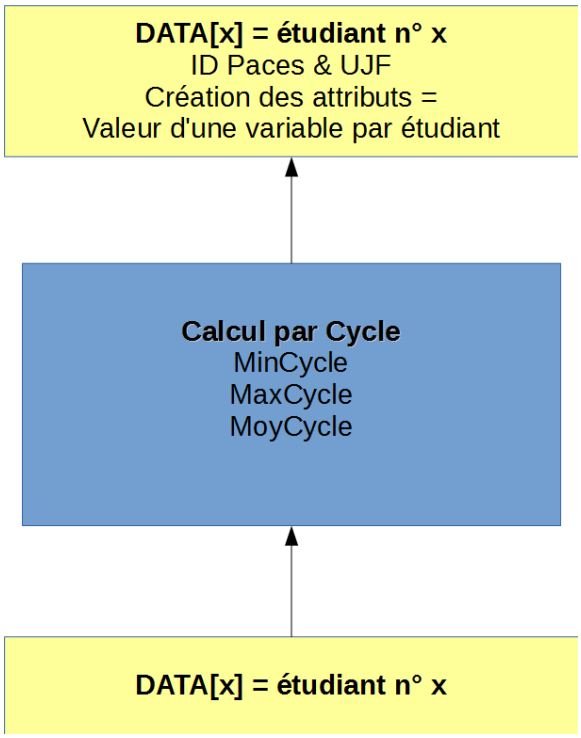
En localisant dans les fichiers les identifiant, le numéro UJF ou PACES, on affecte des valeurs aux attributs de l'objet étudiant ayant l'identifiant en question. De cette manière, on agglutine les connaissances autour du numéro UJF et PACES. Cette technique permet d'éviter les confusions et les doublons.



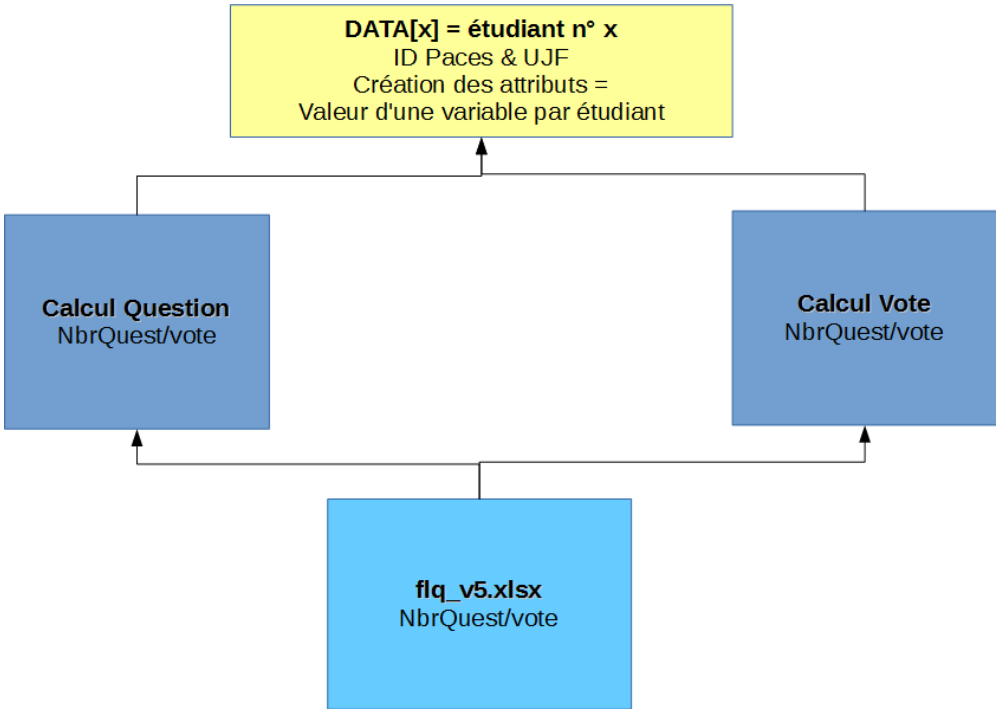
Les différentes itérations de l'analyse ont nécessité la construction de nouvelles variables qu'il faut calculer. Pour le fichier res_v2.xls, contenant les informations sur les résultats aux concours, nous les avons calculé lors de l'extraction même. Il n'est cependant pas nécessaire de procéder ainsi.



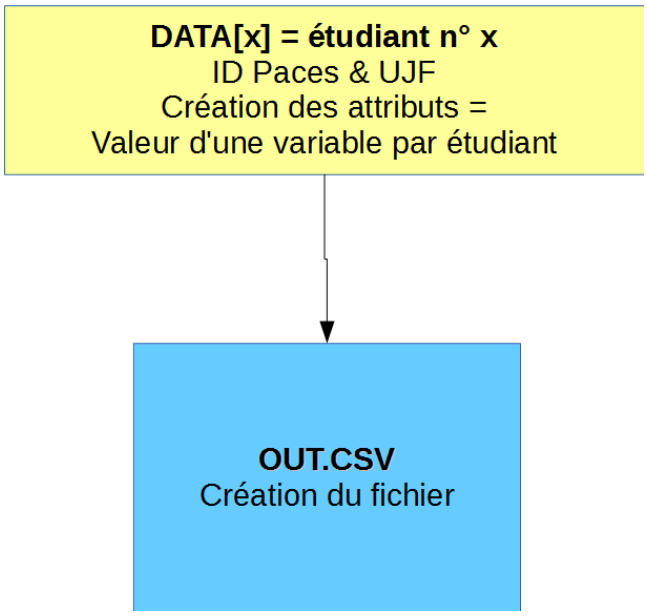
On procède ici de la même manière que pour l'étape précédente. Le fichier res_concours_v6.xlsx a été modifié en amont. Suppression des colonnes non pertinentes, changement de l'entête des colonnes. Exemple : S2_C7_PHS pour les résultats PHS du semestre 2 en physiologie.



Ici on calcul les variables MinCycle MaxCycle MoyCycle en reprenant les informations contenues dans le tableau DATA (Tableau d'étudiant).



On récupère le nombre de question et le nombre de vote par étudiant.



On restitue toutes les connaissances sur un étudiant dans un fichier csv.

LES FICHIERS

Processus

concor.xls	Contient les associations entre numéro UJF et numéro PACES. Ce fichier est le plus important dans la mesure ou il permet d'identifier et récolter les informations sur un étudiant dans les différents fichiers exploités.
inscrits_1_v1.xlsx	Le fichier original contenant les informations sur l'inscription des étudiants PACES. Les variables Sexe, Bac, Mention, Csp, sont extraites ici.
inscrits_2_v2.xlsx	Le fichier de Laura Dupuis contenant les variables construites : Orientation (code binaire), assiduité, Nb- et Nb+.
res_v2.xlsx	Ce fichier contient l'ensemble des notes et du rang par matière des étudiants. La matière est représentée par une variable concaténé (Séquence, dans le fichier final). Elle contient le semestre, le cycle et la matière dans laquelle l'étudiant a eu la note. Exemple S1_C1_BCH = Semestre 1 du cycle 1 dans la matière BCH.
res_concours_v6.xlsx	Ce fichier est une fusion du fichier résultat aux concours Médecine, Maïeutique, Odontologie et Pharmacie. Il contient ainsi tous les résultats des étudiants traités. On récupère ici les

	résultats du concours final avec les résultats du cycle 7. (Rappel : C7 est un cycle ajouté représentant les concours passés en première et seconde année par l'étudiant)
flq_v5.xlsx	Ce fichier contient l'ensemble des questions posées par les 8 groupes durant l'année 2011/2012. Ici nous gardons le nombre de question par étudiant dans chaque matière et le nombre de votes par étudiant dans chaque matière. On construit une variable concaténée comme dans res_v2.xlsx : Q_S1_C1_BCH = Nombre de question dans la matière BCH du cycle 1 au semestre 1. Q_S1_C1_BCH, la même pour le vote.

Points forts	Points faibles
Python pour : éviter l'importation des erreurs du traitement précédents ou des fichiers bruts, portabilité du langage d'une machine à l'autre, scripts utilisables sous Orange Langage C# aurait été plus adapté ?	Fichiers bruts dispersés, à fusionner avec une table de concordance (numéro étudiants différents) Obligation de nettoyer les fichiers bruts (format des données, ; et accents, colonnes en trop) avant de passer dans le traitement Progammes adaptés aux fichiers PACES donnés, pas utilisable sur d'autres fichiers Nécessité compétence programmation sous excel ou SQL ou Python

Points forts UT/Orange	Points faibles UT/orange
on aurait pu faire la chaine de traitements sous UT/orange qui offre un environnement de développement dédié aux traitements et un moyen de conserver et rejouer toute la chaine possibilité de créer des opérateurs (python, R)	sous UT, il faut créer un nouvelle étude à chaque nouveau fichier et la liste des études est peu praticable (longue) sous Orange : documentation faible, difficultés à faire fonctionner certains opérateurs (Bayes, régression)

Description des analyses

PROCESSUS (réalisé jusque l'étape 4 par S Foudil) :

- Sélectionner un échantillon d'individu pour faciliter les analyses.
- Permet de fournir une visualisation des variables selon des profils variés
- Créer de nouvelles variables dans le but de caractériser au mieux les comportements.

1. Intégrer les données à la plate-forme UnderTracks
2. Visualiser un groupe de 5 étudiants, dans la catégorie ADM, AJ et EX. L'objectif ici est de visualiser les 8 courbes notes, les 8 courbes rang, les 8 courbes questions et votes. Trouver à partir des observations de ces courbes des invariants. Typologie d'étudiants.
3. Définir de nouvelles variables si nécessaire. On peut penser à la pente moyenne par matière, toute matière confondue, etc…
4. Stats descriptives : dans un premier temps pour observer la "surface" des données. (Distribution des Deltas et autres variables créées).
5. Clustering : sur les nouvelles variables définies plus tôt. Mettre en évidence les comportements des étudiants.
6. HMM : Dans l'idéal, appliquer un modèle de markov caché sur les données de manière à prédire un comportement "défaillant".
7. NB : Pour obtenir des prédictions sur les données, on peut utiliser un classifieur naïf de bayes.
8. on peut faire un modèle de régression logistique, courbe de ROC, Intervalle de confiance, pour prédire la réussite en fonction des résultats du S1

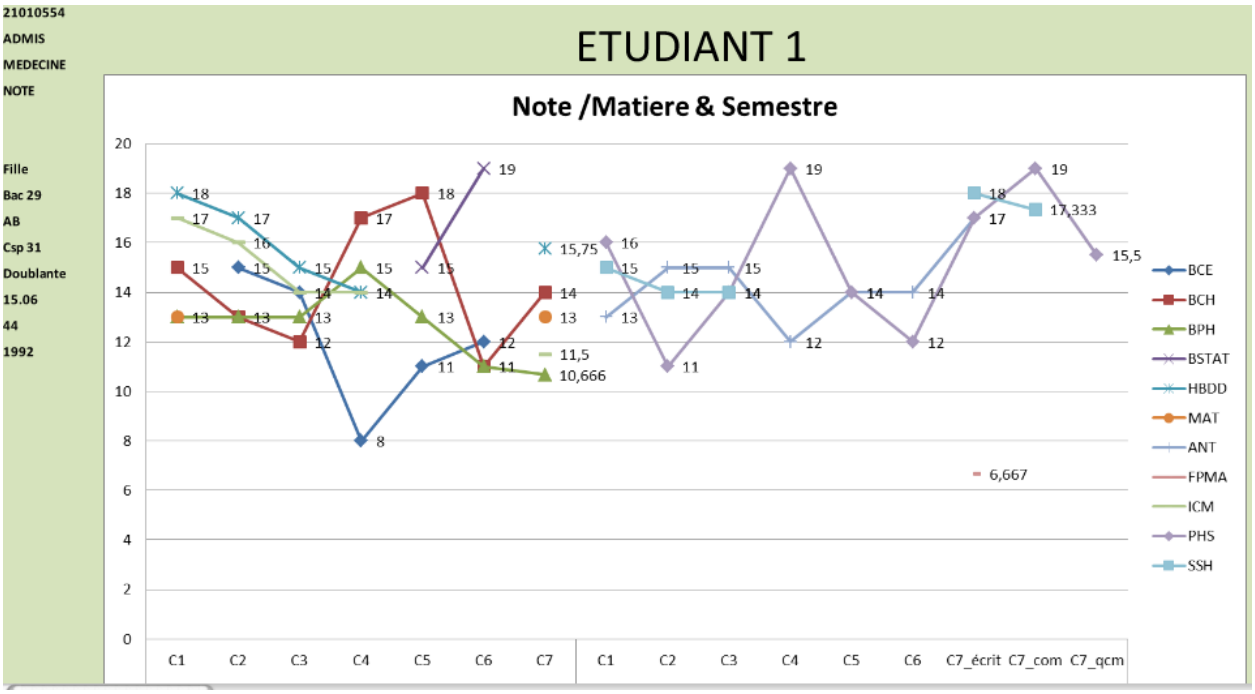
Mode opératoire méthodologique
On se donne 4 types d'étudiants (suggéré par les enseignants)
On visualise un échantillon d'étudiants dans chaque type, sous forme de graphique en fonction du temps

Mode opératoire technique, logiciels utilisés
Analyses sous excel et sous Orange/UT

Scripts produits pour l'analyse des données
feuille excel

Résultats obtenus:

Exemple de visualisation



Points forts des analyses	Points faibles des analyses
les visualisations ont permis de choisir de nouvelles variables pertinentes et des analyses à faire sur ces variables	typologie d'étudiant choisie a priori (4 types) difficile de voir des invariants sur un échantillon de 5 étudiants par type, grande variabilité, mais quelques pistes

Description des itérations:

Décrire la production des nouvelles données

suite aux visualisations, création de nouvelles variables

Décrire le changement dans les méthodes d'analyse

Points forts des itérations	Points faibles des itérations