

**MODELISATION DES PROCESSUS D'ANALYSE POUR LA
COMPREHENSION DES STRATEGIES DE SUPERVISION ET DE
RETROACTION DES ENSEIGNANTS PENDANT LEUR ACTIVITE EN FACE
DES ELEVES**

(ÖĞRENCİLERİN ÖNÜNDE ÖĞRETMENLERİN DAVRANIŞLARINI VE
TEPKİLERİNİ ANLAMAYA YÖNELİK ANALİZ İŞLEMİNİN MODELLENMESİ)

Réalisé par

Gülay YILDIRIM, Ingenieur en Génie Informatique

Thèse

Présenté en vue de

l'obtention du diplôme de

MASTER

en

GENIE INFORMATIQUE

de

L'INSTITUT DES SCIENCES

de

L'UNIVERSITE GALATASARAY

Septembre 2014

**MODELISATION DES PROCESSUS D'ANALYSE POUR LA
COMPREHENSION DES STRATEGIES DE SUPERVISION ET DE
RETROACTION DES ENSEIGNANTS PENDANT LEUR ACTIVITE EN FACE
DES ELEVES**

(ÖĞRENCİLERİN ÖNÜNDE ÖĞRETMENLERİN DAVRANIŞLARINI VE
TEPKİLERİNİ ANLAMAYA YÖNELİK ANALİZ İŞLEMİNİN MODELLENMESİ)

Réalisé par

Gülay YILDIRIM, Ingenieur en Génie Informatique

Thèse

Présenté en vue de

l'obtention du diplôme de

MASTER

Date de soumission : Septembre 23, 2014

Date de soutenance : Novembre 3, 2014

Directeur : M. de Conf Gönen EREN

Membres de Jury : M. de Conf Teoman NASKALI

M. de Conf Hakan AKYUZ

PREFACE

Tout d'abord, je tiens à exprimer mes sincères remerciements à mon conseiller adjoint. M.Conf Gönen EREN pour le soutien continu de mon étude de maître et de la recherche, pour sa patience, la motivation et des connaissances immenses.

Je tiens à exprimer mes remerciements les plus chaleureux et sincères pour adjoint. M. Conf Vanda Luengo, qui m'a présenté à Metah Lab à Grenoble, France. Son aimable soutien et des conseils ont été d'une grande valeur dans cette étude, ainsi que l'essentiel m'apprendre tout ce que je sais à propos du sujet, j'ai travaillé en France. J'ai réussi de finir la formation co-diplôme entre les universités UJF et Galatasaray grâce à elle.

Je tiens à remercier ma famille et mes amis pour leur encouragement et leur soutien.

Gülay YILDIRIM

Septembre, 2014

TABLE DES MATIERES

PREFACE	ii
TABLE DES MATIERES	iii
LISTE DES FIGURES	v
RESUME	vi
ABSTRACT	viii
ÖZET	x
1. INTRODUCTION	1
2. ETAT DE L'ART	3
3. CONTEXT	6
3.1. Hypothèses du projet SuperViseur	7
3.1.1. Hypothèse 1: Conscience de la classe	7
3.1.2. Hypothèse 2: Groupe de pilotage	8
3.1.3. Hypothèse 3: Facteurs influençant le type de rétroaction	9
3.2. Protocole expérimental	9
4. QUESTIONS DE RECHERCHE	11
5. METHODOLOGIE POUR REpondre AUX QUESTIONS DE RECHERCHE	13
5.1. Les méthodes d'exploration de données d'éducation	13
5.1.1. GSP	14
5.1.2. PrefixSpan	15
5.1.3. SPADE	15
5.1.4. SPAM	15
5.2. Outils et Plateformes	16
5.2.1. Oculométrie (eye-tracking)	16
5.2.2. ANVIL	17
5.2.3. Orange	17
5.2.4. CLASS	18
5.2.5. TDOP	18
5.2.5.1. Au niveau de l'enseignant	19
5.2.5.2. Au niveau de l'interaction entre l'enseignant et les élèves	19

5.2.5.3. Au niveau des élèves	20
5.2.6. Sequential Pattern Mining Framework	20
5.3. Concevoir de processus.....	20
5.4. Les widgets créés	23
5.5. Les analyses réalisées pour chaque hypothèse.....	26
5.5.1. Analyse pour l'hypothèse 1	26
5.5.2. Analyse pour l'hypothèse 2.....	30
6. CONCLUSION ET DISCUSSION	36
BIBLIOGRAPHIE	39
APPENDICE	44
BIOGRAPHIE	45

LISTE DES FIGURES

5-1. Le format pour l'algorithme Hirate-Yamana	16
5-2. Le processus entier	21
5-3. La sortie de l'outil ANVIL.....	22
5-4. La visualisation des données des élèves	24
5-5. Visualisation des données de regard.....	25
5-6. La génération des données synthétiques	26
5-7. Les séquences des identifiants des élèves	27
5-8. La sortie du widget "Convert2HY" et l'entree de l'algorithme Hirate-Yamana	28
5-9. L'interface de l'algorithme Hirate-Yamana	28
5-10. La sorties de l'algorithme Hirate-Yamana.....	28
5-11. Une autre version de sequence	29
5-12. Distribution des regard	30
5-13. Le resultat de clustering k-mean.....	31
5-14. La visualisation de k-mean.....	32
5-15. Les membres de clusters.....	33
5-16. Les traitements complets pour cette hypothèse	34

RESUME

La gestion d'une salle de classe est l'un des aspects le plus difficile pour de nombreux enseignants. Plusieurs recherches dans la littérature montrent qu'il existe une relation directe entre l'enseignant et les notes des étudiants. En ce point, le comportement de l'enseignant dans la classe et l'interaction avec les étudiants sont très importantes. Dans notre travail, nous nous intéressons aux comportements des enseignants et à la façon qu'ils analysent les étudiants. Pour le côté analyse des travaux, la branche de l'exploration de données que l'on appelle "educational data mining" est utilisée. Le principal objectif de l'exploration de données est d'extraire les informations pertinentes de la masse de données en utilisant plusieurs techniques différentes.

Dans notre recherche, il y a deux hypothèses attachées aux comportements des enseignants et le but du projet est de prouver ces hypothèses en utilisant des données expérimentales. A cet effet, plusieurs sources de données différentes sont rassemblées. D'un côté, avec eye-tracking qui est le processus de mesure du point de regard, il est possible de détecter où l'enseignant regarde. De l'autre côté, nous avons des documents sur les comportements des élèves remplis avec leurs enseignants. Ces deux types de données différentes aident afin d'analyser et de prouver les hypothèses.

Dans ce travail, selon les hypothèses, tout d'abord, nous avons travaillé sur l'existence d'un motif spécial; quand l'enseignant regarde les étudiants. Nous avons cherché s'il y a un ordre spécifique dans son regard c'est à dire avec quel fréquence l'enseignant regarde ses élèves, si l'enseignant regarde chaque étudiant avec la même durée ou non. En ce point, la méthode d'extraction de motifs séquentiels est impliquée dans notre travail. En outre, nous avons cherché s'il y a un groupe d'élèves que l'enseignant se concentre en circulant dans le cours. Nous avons appelé ce groupe "un groupe de référence".

En conclusion, au cours de ce travail, les méthodes d'exploration de données sur l'éducation sont appliquées sur les données du système éducatifs en utilisant la plateforme de data mining. Par conséquent, ces données provenant de diverses sources sont analysées afin de prouver les hypothèses. Pendant le processus d'analyse, l'outil Orange Data Mining et ses widgets sont utilisés. Selon les besoins des nouveaux widgets sont mis en oeuvre. L'implémentation des widgets est réalisée de façon modulaire et réutilisable ainsi, en cas d'augmentation de nombre de participants à l'ensemble des expérimentations même composant peut traiter le nouvel ensemble de données.

Avec les données expérimentales, on a trouvé l'existence de patterns spéciales dans le comportement et les regards des enseignants. Par exemple, il existe un ordre précis dans les regards de l'enseignant, il existe un groupe des élèves que l'enseignant regarde le plus souvent et aussi l'enseignant regarde plutôt aux élèves handicapés.

ABSTRACT

Classroom management is one of the difficult aspects of teaching for many teachers. Many works in the literature show that there is a direct relation between the teacher and the student grades. In this point, analyze of teacher behavior in the class and interaction with the students become very important. In our work, we are interested in behaviors of the teachers and how they analyze the students. For the analyze side of the work, branch of data mining that is called educational data mining is used. Main goal of data mining is to extract relevant information from big data mass by using several different techniques. Educational Data Mining (EDM) is the application of data mining techniques on educational system data.

In our research, there are two hypotheses related with teacher behaviors and the aim of the project is to prove these hypotheses by using experimental data. For this purpose several different data sources are gathered. On one side, with eye-tracking which is the process of measuring either the point of gaze, it is possible to detect where the teacher is looking. On the other side, we have documents on students behaviors filled with their teachers. These two different types of data are gathers in order to analyze and prove the hypotheses.

In this work, according to the hypotheses, first of all we worked on existence of special pattern while teacher is looking to the students and we searched if there is a specific order of looking, how of the teacher looks students, if the teacher looks each student with same duration or not. In that point, sequential pattern mining method is involved in our work. Furthermore, we searched if there is a group of student that teacher focus on for the flow of the course. We called this group as a group of reference.

As a conclusion, during this work, educational data mining methods are applied on educative system data by using data mining platform. Hence these data coming from

different sources are analyzed in order to prove hypotheses. During the analyze process, Orange Data Mining tool and its widgets are used. Depending on the need new widgets are implemented. The implementation of widgets is realized in a modular way thus, in case of increase of number of participants to the experimental set, same widget can process the new dataset.

At the end of this study, with the experimental dataset, we figured out existence of special pattern in teacher's behavior and look. For instance, we found that there is an order of looking students and it exists a reference group that the teacher looks usually. Also, teacher insists on looking handicapped students.

ÖZET

Birçok öğretmen için sınıf yönetimi eğitimin ve öğretimin en zorlu kısmıdır. Literatür araştırması yapıldığında öğretmen ile öğrencilerin notları arasında doğrudan bir ilişki olduğu açıkça görülmektedir. Bu bağlamda düşündüğümüzde öğretmenin sınıf içindeki davranışları, öğrencileriyle olan etkileşimi büyük önem arz etmektedir. Bu çalışmamızda öğretmenlerin öğrenci önünde, ders anlatırkenki davranışları ve öğrencilerin ders sırasındaki hareketleri incelenmiştir. Analiz sırasında veri madenciliği tekniklerinden yararlanılmıştır. Özellikle veri madenciliğinin eğitim dalı olan eğitsel veri madenciliği bu çalışmada kullanılmıştır. Büyük veri yığınlarından ilgili bilgileri çeşitli teknikler ile elde etmek veri madenciliğinin ana prensibidir.

Bu çalışmamızda konu ile alakalı yürütülmekte olan proje bünyesindeki hipotezleri başlangıç noktası olarak ele aldık. Bu hipotezlerden iki tanesi öğretmenin davranışları üzerinedir. Amacımız ise elimizdeki deneysel verileri ve eğitsel veri madenciliği metotlarını kullanarak bu hipotezlerin doğru olduğunu kanıtlamaktır. Çalışma sırasında farklı kaynaklardan deneysel veriler toplandı. Eye-tracking adı verilen cihaz yardımıyla öğretmenin nereye baktığı bilgisi kolaylıkla elde edilebilmektedir. Diğer taraftan öğretmenler tarafından doldurulan öğrencilerin davranışlarını açıklayan dokümanlar da bu çalışmada kullandığımız diğer bir kaynaktır. Bu birbirinden farklı iki kaynak hipotezleri doğrulamak ve sınıf ortamını anlayabilmek için önemlidir.

Çalışmaya, projeden gelen hipotezleri analiz ederek başladık. Hipotezlerin ışığında öğretmenlerin öğrencilerine bakarken belirli, özel bir modele sahip olup olmadığını araştırdık. Bu bakışların bir sırası olup olmadığını sorguladık. Örneğin bir öğrenciye hangi sıklıkla bakıyor, tüm öğrencilere eşit sürelerde mi bakıyor yoksa daha fazla yoğunlaştığı öğrenciler var mı? Varsa bu öğrencilerin özellikleri nelerdir? Dahası öğretmenin bakışlarından yola çıkarak öğrenci öbeklerinin olup olmadığını araştırdık.

Örneğin öğretmenin bir konunun anlaşılıp anlaşılmadığını öğrenmek için baktığı bir öğrenci öbeği varsa bu gruba referans grubu denir.

Sonuç olarak, veri madenciliği platformu kullanılarak veri madenciliği metotları eğitim deney verileri üzerinde uygulanmıştır. Farklı kaynaklardan toplanan veriler ile projedeki hipotezlerden ikisinin doğruluğu kanıtlanmıştır. Verilerin analiz aşamasında veri madenciliği platformlarından biri olan Orange Platformu kullanıldı ve platform bünyesinde widget adı verilen araçlardan çalışmada kullanılmak üzere yenileri geliştirildi. Bu geliştirme işlemi ve programlaması, verilerin artması, deney setine ilerde eklenecek sınıflar düşünülerek modüler yapıda ve işlevsel şekilde gerçekleşti.

Çalışmamızın sonunda elimizdeki deneysel veriler ile öğretmenlerin sınıf ortamında öğrencilere ders anlatırken bakışlarında ve davranışlarında belirli bir modele sahip olduğunu saptadık. Örneğin sınıftaki öğrencilere aynı sıra ile baktığını, belirli öğrenci gruplarına daha çok baktığını, özel ilgiye muhtaç olan öğrencilere daha sık ve uzun süreli baktığını bulduk.

1. INTRODUCTION

La gestion de la classe n'a pas une seule définition mais en générale contient les actions des enseignants pour construire et pour maintenir l'ordre, l'engagement des élèves ou extraire les coopérations entre eux (Lewis, 1999). La classe dont on parle ici est en fait un environnement dynamique dans lequel il y a plusieurs variables qui interagissent au cours du temps.

De notre point de vue, dans un système d'éducation, l'élève n'est pas le seul acteur. L'enseignant et les élèves construisent un système. L'enseignant a trois propriétés caractéristiques: la conscience de la situation, le groupe de référence et la prise des informations lors d'épisodes de rétroaction. La conscience de la situation nous explique que l'enseignant est au courant de ce qui se passe dans la classe. Il suit les événements qui ont lieu, les comportements des élèves. Au niveau de groupe de référence, un groupe d'élève peut être choisi par l'enseignant pour le déroulement du cours. L'enseignant peut regarder ces élèves pour vérifier s'ils ont compris le cours ou non. Le dernier point nous explique l'importance de rétroaction vient des élèves, c.à.d. les comportements des élèves favorisent la compréhension des élèves. L'enseignant peut superviser la bonne progression des événements.

Dans ce travail, le but principal est de modéliser les processus d'analyse pour la compréhension des stratégies de supervision et de rétroaction des enseignants pendant leur activité en face des élèves. On a des hypothèses déjà soulevées pour ce domaine de recherche. Pour prouver les hypothèses, on a besoin d'analyser des données expérimentales qui sont récupérées dans un environnement écologique, i.e. en situation réel sans modification expérimentale. On obtient des données quantitatives et qualitatives par différents outils comme la suivi de regard (l'oculométrie), CLASS (ClassroomAssessmentScoring System). La liste complète des outils utilisés dans le

projet sera expliquée dans la partie Outil. Les données venant de ces outils doivent être traitées par des méthodes de l'exploration de données d'éducation.

Educational Data Mining (EDM) est une discipline dans laquelle les méthodes de l'exploitation des données sont appliquées sur des données de l'éducation. En fait EDM contient les mêmes étapes que l'exploration de données comme prétraitement, application des méthodes de DM et post-traitement. Pendant l'étape d'application de DM, dans le domaine EDM, il existe des techniques comme régression, corrélation et visualisation qu'il n'existe pas dans DM (Romero, 2010). La différence majeure d'EDM vient de la nécessité de tenir compte explicitement de la hiérarchie multi-niveaux et non-indépendance des données éducatives (Baker, 2009).

Comme indiqué avant, dans ce travail on a deux types de données à traiter: des données quantitatives c.à.d. des mesures numériques et des données qualitatives c.à.d. des mesures non-numériques. Une des difficultés de ce travail est de proposer des processus de traitement pour d'analyser ces données et de traiter ensemble ces deux types de données sur la même plateforme.

2. ETAT DE L'ART

Même si l'exploration des données d'éducation est un sujet de recherche très récent, il y a plusieurs travaux réalisés au tour d'EDM. Baker, dans son papier (Baker, 2010), cite différentes méthodes d'EDM. Il les catégorise comme les suivantes: prediction, clustering, relationship mining, discovery with models, distillation of data for human judgement. Il a aussi présenté quatre différentes sous-méthodes de relationship mining: association rule mining, correlation rule mining, sequential pattern mining, causal data mining. Parmi ces méthodes, dans notre travail, pour prouver nos hypothèses, clustering, relationship mining et ses sous-méthodes sequential pattern mining et causal data mining peuvent être utiles. Pour regrouper nos données, pour mettre ensemble les données similaires, c.à.d. diviser les données en K groupes, on va utiliser la méthode clustering. Comme nos données sont séquentielles, on peut utiliser l'exploration du pattern séquentiel. On va appliquer l'exploration des données causales pour extraire les liaisons entre nos données.

Dans la littérature, la plupart des travaux réalisés sur la gestion de classe ont considéré les élèves comme l'acteur principal et ils analysent les activités des élèves. Par exemple, Romero et Ventura (2007) indique que l'exploration de données peut être utilisé pour comprendre comment les élèves exploitent le cours, pour analyser l'effet d'une stratégie pédagogique sur des différents types d'élèves, pour voir comment les élèves s'organisent parmi leurs sujets, pour mesurer combien et comment les élèves travaillent sur une page/un chapitre d'un sujet. Dans notre travail, au lieu des apprenants, on s'intéresse aux comportements des enseignants, leurs interactions vis-à-vis les élèves.

Bien sûr qu'il y a des travaux qui s'intéressent aux enseignants mais peu de travaux se sont intéressés à l'analyse automatique de l'activité de l'enseignant. Ainsi, comme (Marzano & Pickering, 2003) indique dans son livre, les enseignants jouent divers rôles

dans une salle de classe, mais sûrement l'un des plus importants est celui de gestionnaire de salle de classe. L'enseignement et l'apprentissage efficace ne peut avoir lieu dans une salle de classe mal gérée. Par exemple, l'étude sur la stratégie de classe menée par JereBrophy et McCaslin (1992) a été la gestion de la salle de classe d'étude. Il s'agissait d'entrevues en profondeur avec des observations et avec 98 enseignants, dont certains ont été identifiés comme des gestionnaires efficaces et certains d'entre eux ne l'étaient pas. L'étude présentait à des enseignants des vignettes sur les types spécifiques d'étudiants (ex: les étudiants hostiles-agressifs, les étudiants passifs-agressifs, hyperactifs étudiants) dans des situations spécifiques. Parmi les nombreuses conclusions de l'étude, il en était que les gestionnaires de classe efficaces avaient tendance à utiliser différents types de stratégies avec différents types d'étudiants, tandis que les gestionnaires inefficaces avaient tendance à utiliser les mêmes stratégies, indépendamment du type de l'étudiant ou de la situation. L'une des recommandations fortes de l'étude était que les enseignants doivent développer un ensemble de "compétences aidant" à utiliser avec différents types d'étudiants. Dans notre travail, surtout dans les salles d'expérimentales on a aussi différents types d'élèves. On cherche quel type d'élève (le groupe de références) que l'enseignant regarde souvent.

La compréhension commune dans la communauté d'EDM (Romero, 2010) est qu'il y a plusieurs types de formation: l'éducation traditionnelle, l'éducation basée sur le web(e-éducation), avec des systèmes informatiques de gestion de l'apprentissage (comme Moodle, Chamilo,..), les systèmes de tutorat intelligent etc. Dans notre cas les expérimentations viennent des salles de classe et d'un système éducation traditionnelle. On rappelle bien que (Heiner et al. 2007) définit l'exploitation des données d'éducation comme un processus de conversion des données brutes pris dans des systèmes éducatifs. Le même travail pense qu'il est possible d'utiliser ces données pour éclairer les décisions de conception. De même point de vue, dans notre travail, on essaie de chercher les comportements spécifiques dédiés aux enseignants. L'utilisation des outils complémentaires comme eye-tracking, vidéo etc. donne la chance d'utiliser ces technologies dans le domaine d'éducation avec le but d'analyser les données de l'éducation surtout les comportements des enseignants et des élèves.

Dans la littérature, il existe aussi des chercheurs qui travaillent sur différents types d'enseignants surtout les enseignants experts et les nouveaux enseignants. Ils cherchent la différence entre ces deux niveaux et son effet sur la gestion de classe (van den Bogert et al., 2014). Ils ont montré que les enseignants experts peuvent traiter les données visuelles plus rapides que les nouveaux enseignants. De plus, les enseignants experts sont capables de propager leurs intérêts plus fréquemment et uniformément que les nouveaux enseignants. Dans notre travail, on observe aussi la différence entre expert-novice. Par contre pour nous, la différence entre expert et novice varie selon la score de CLASS. Par exemple dans une de nos hypothèses, on essaye de prouver que l'enseignant expert peut détecter l'élève qui est hors-tache plus vite que le novice.

3. CONTEXT

L'objet de travail est de proposer des processus d'analyse à partir des données de l'éducation qui viennent de différents outils. Une des données principales de notre travail est le pistage du regard. Le pistage de regard peut être capté par l'outil eye-tracking qui sera expliqué dans la partie Outil. Afin de mieux comprendre leur activité en classe, les stratégies de prise d'informations des enseignants dans un contexte écologique, ces données doivent être analysées. L'utilisation de ce type d'outil est bien adaptée à l'étude de superviseurs d'environnements dynamiques, où la pression temporelle et la diversité des sources potentielles d'évolution sont difficiles à capter avec des outils d'observation indirects. La problématique du projet s'articule sur les trois questions générales suivantes:

- Conscience de la classe: quelles sont les relations entre les prises d'indices visuels dans l'environnement scolaire et le niveau de la qualité d'interactions enseignants-élèves?
- Groupe de référence: y existe-t-il un groupe de référence que l'enseignant supervise régulièrement durant son activité et quelles sont ses caractéristiques (niveau et nombre des élèves, groupe unique ou multiple, quelles variations selon le type des décisions)?
- Prises d'informations lors de rétroactions: quelles sont les prises d'indices que l'enseignant réalise pendant une séquence de rétroactions, avec quels patterns d'interaction?

Notre travail n'est pas de répondre à ces questions mais de trouver des processus d'analyse permettant de répondre à ces questions à partir des données qui seront décrites. Pour cela, il faut construire une liaison entre les données oculaires et d'autres données situationnelles.

D'où, pour le travail, il existe trois hypothèses qu'on essaie de prouver avec l'analyse des données. Ces trois hypothèses sont construites sur la relation entre l'enseignant et les élèves, le comportement des enseignants dans la salle de cours.

3.1. Hypothèses du projet SuperViseur

3.1.1. Hypothèse 1: Conscience de la classe

La première hypothèse du projet est reliée avec les événements pédagogiques qui se sont réalisés dans la salle de cours. Un événement pédagogique est une activité qui concerne tous les méthodes et pratiques d'enseignement et d'éducation ainsi que toutes les qualités requises pour transmettre une connaissance, un savoir ou un savoir-faire*. Comme on a déjà mentionné, avec CLASS, il est possible de prendre des données sur plusieurs critères de suivi de l'enseignant. D'où notre hypothèse principale est constituée par les données de CLASS:

Les enseignants ayant des niveaux élevés de "sensibilité" et de "gestion des comportements" (mesurés par CLASS) auront un pattern de comportements oculaires spécifiques. Ici on peut relever trois patterns différents:

- Pour un événement pédagogique comparable, un plus grand nombre d'élèves seront scrutés, avec une fréquence plus importante chez les enseignants présentant des scores élevés sur les dimensions "sensibilité" et "gestion des comportements" de CLASS.
- Pour un événement pédagogique comparable, les enseignants présentant des scores élevés sur les dimensions "sensibilité" et "gestion des comportements" de CLASS regarderont moins longtemps et avec une fréquence moindre le matériel.
- A partir du moment où un élève est observé comme hors tâche, les enseignants présentant des scores élevés sur les dimensions "sensibilité" et "gestion des

*<http://fr.wikipedia.org/wiki/pedagogie>

comportements" de CLASS reviendront visuellement sur cet élève, et en moyen plus rapidement que les autres:

- les enseignants présentant des scores élevés sur les dimensions "sensibilité" et "gestion des comportements" de CLASS interagiront (verbalement ou comportementalement) plus fréquemment et plus rapidement, en moyen, que les autres.
- les enseignants présentant des scores élevés sur les dimensions "sensibilité" et "gestion des comportements" de CLASS auront des élèves hors-tâche qui passeront plus rapidement dans-la-tâche que les autres enseignants.

3.1.2. Hypothèse 2: Groupe de pilotage

Dans l'hypothèse 2, on parle d'un groupe de pilotage. Les enseignants regarderont plus fréquemment un groupe composé de quelques élèves, formant le groupe de pilotage (groupe de référence). La suite de l'hypothèse dit que ce groupe sera décomposée en 2 sous-groupes. Certains élèves sont scrutés pour des raisons comportementales (élèves à superviser pour assurer un bon déroulement de la classe) et auront une activité hors-tâche (EC) significativement plus élevée que ceux de l'autre groupe; d'autres pour des raisons cognitives (groupe " compréhension ", élèves à superviser pour s'assurer que la classe comprend le déroulement de l'enseignement).

Les élèves du groupe " compréhension " auront une activité dans-la-tâche significativement plus élevée que l'ensemble de la classe et aussi aura un niveau moyen de performances compris entre le 10e et le 25e percentile (Lundgren, 1993).

Tandis que les élèves du groupe " comportemental " auront une activité hors-la-tâche significativement plus élevée que l'ensemble de la classe et aura un niveau moyen de performances au BRIEF inférieur à l'ensemble de la classe. La BRIEF est un inventaire

permettant d'évaluer les comportements exécutifs dans les milieux scolaire et/ou familial. Elle est utile lors d'évaluations cliniques ou neuropsychologiques concernant les différents aspects du dysfonctionnement exécutif et ses répercussions dans la vie quotidienne. Elle permet de disposer d'un profil précis des troubles pour orienter la prise en charge de l'enfant ou de l'adolescent et l'aide aux familles (Gioia, 2000).

3.1.3. Hypothèse 3: Facteurs influençant le type de rétroaction

La troisième hypothèse est reliée avec la réaction des enseignants. On cherche les rétroactions générées par les enseignants et leurs raisons. On croit que les types de rétroactions formatives (Shute, 2008) produites par les enseignants auront des facteurs de déclenchement différents. Ça veut dire que le jugement de l'enseignant et les informations prises par l'enseignant préalables à la production de la rétroaction causeront des rétroactions formatives différentes, et reliées à un score CLASS "qualité de la rétroaction" différent. D'où il est possible de trouver une causalité entre les prises d'information de l'enseignant, le type de rétroaction, et le score classe "qualité de la rétroaction". De plus les types rétroactions formatives correspondent à des patrons sur la forme de la rétroaction. La forme de la rétroaction sera caractérisée par le récepteur de la rétroaction (élève, groupe, classe) et le média (oral, écrit).

3.2. Protocole expérimental

Dans nos données expérimentales, on a au total quatre enseignants d'école primaire avec deux enseignants novices et deux enseignants expérimentés. Pour les enseignants expérimentés, ils doivent avoir au moins 15 années d'expérience et si possible ayant été formateur/formatrice d'enseignants. Pour les enseignants novices, ils ne doivent pas être en difficulté pendant leur stage, ayant un intérêt pour cette recherche. Au niveau des élèves, on a environ 22-26 élèves dans chaque classe. Comme on utilise l'outil eye-tracking pour capturer le regard, il y a quelques restrictions pour les enseignants : ils ne doivent pas porter des lunettes, ils ne doivent pas avoir de strabisme. On a choisi les cours similaires pour qu'on puisse comparer en utilisant les scores de CLASS. La durée

des cours est 45 minutes environs. En dehors des outils portés par les enseignants, on a un camera fixé pour capturer la classe en entier.

De plus, on a quatre types de tableaux que les enseignants doivent remplir avec des données des élèves :

1. Tableau démographiques : numéro d'identifiant d'élève, l'âge et le sexe
2. Tableau de niveau en français et en mathématiques (T, B, P, F)
3. Elèves qui ont des besoins particuliers : ici ils indiquent les élèves handicapés ou besoin d'aide pendant le cours.
4. Questionnaires sur des activités et les comportements scolaires des élèves : Il y a 13 questions à répondre. Pour chaque élève, les enseignants répondent aux questions en utilisant le barème 0-4. (voir l'annexe)

De l'autre côté, deux observateurs au fond de la classe feront 4 codages CLASS de 20 min d'interactions. Pour favoriser l'accoutumassions au matériel, les lunettes et la caméra seront installées pendant la demi-journée.

4. QUESTIONS DE RECHERCHE

Dans la partie précédente, on a présentée les hypothèses qu'on essaie de prouver pendant ce travail. Comme indiqué, par rapport à plusieurs travaux dans la littérature, on s'intéresse à analyser les traces de l'enseignant et non celles des apprenants. Dans cette partie du document, on va présenter les méthodes possibles et les analyses qu'on peut appliquer pour chaque hypothèse.

Dans la première hypothèse (hypothèse reliée avec les événements pédagogiques qui se sont réalisés dans la salle de cours), on parle de scores des enseignants mesurés par CLASS. L'hypothèse 1 cherche s'il existe un pattern de comportements oculaires. Pour le montrer, on a trois sources de données: les données de CLASS, le suivi de regard avec les données séquentielles, les événements pédagogiques.

Dans l'analyse de cette hypothèse pour le premier pas, on divise les vidéos en séquences selon des événements pédagogiques(EVP). Puis ces EVP seront groupés selon leurs types. Pour chaque EVP, on a les données de regard pendant cet EVP. On va chercher l'existence d'un pattern spécifique entre le même type d'EVP. Comme l'ordre des données est très important, l'algorithme de l'exploration des patterns séquentiels sera utilisé pour extraire des patterns. Cet algorithme sera appliqué sur des données de regard. A la fin de ce processus, la sortie de l'algorithme sera comparée avec les données de CLASS pour voir s'il y a une relation entre eux ou pas. Comme Relational Data Mining peut construire une liaison entre les variables, ça peut être utile pour notre analyse.

Pour la deuxième hypothèse, on essaie de trouver des groupes de pilotage. Selon l'hypothèse, ce groupe de pilotage est constitué par des élèves qui sont regardés plus fréquemment par l'enseignant. Pour le montrer, les données de regard et les notes de comportements données par enseignant pour chaque élève seront utilisées. Pour

chaque élève dans une classe, on calcule " indicateur d'élève " en regardant leurs notes des cours et les notes de comportements. On a besoin d'avoir un indicateur d'élève pour classier les élèves parce qu'on a plusieurs classes et on doit les classier d'une façon ou d'une autre. De l'autre côté, on a aussi les données du regard pour ces élèves. Avec ces deux groupes de données de différents types, un des algorithmes de cluster sera appliqués (comme k-mean). La sortie de cet algorithme nous montrera s'il y a un groupe de pilotage ou non.

Pour la troisième hypothèse, on parle d'une causalité entre les informations prises par l'enseignant et leur rétroaction. On cherche les types d'information que l'enseignant peut prendre et en même temps les types de rétroactions générées par les enseignants. Ensuite, il sera possible de trouver une relation entre ces deux. Dans cette hypothèse, les rétroactions seront classifiées par la classification de Shute(2008). La liaison entre les informations et les rétroactions sera cherchée.

5. METHODOLOGIE POUR REpondre AUX QUESTIONS DE RECHERCHE

Dans cette partie du document, avant d'expliquer les analyses, les méthodes de l'exploration de données seront présentées. Puis parmi les méthodes, les algorithmes qui sont choisis pour utiliser dans notre travail seront donnés. Ensuite pour pouvoir appliquer ces méthodes, les outils et les plateformes utilisées seront présentés. La dernière partie de section sera l'explication des analyses pour chaque hypothèse et les implémentations sur des plateformes.

5.1. Les méthodes d'exploration de données d'éducation

Comme souligné au début, la plupart des méthodes de l'exploration de données peuvent être appliquées sur des données éducatives. Parmi ces divers algorithmes, la première méthode qu'on utilise est le clustering. Surtout en analysant de données statistiques, la méthode clustering décrit des méthodes de classification de données. En effet, cette méthode sert à diviser un ensemble de données en différents groupes homogènes, d'où les données regroupées dans chaque sous-groupe partagent des caractéristiques communes (similaires). Dans le domaine EDM, le clustering a été développé pour modéliser le comportement des élèves (Ritter et al., 2009).

Parmi les algorithmes de clustering, dans notre travail on utilise l'algorithme de k-mean clustering. L'algorithme de k-mean est un algorithme de regroupement de données correspondantes des statistiques et de l'apprentissage automatique. Dans ce cas, k-mean contient l'apprentissage non supervisé. Il permet de diviser les données en K groupes (clusters) dans lesquelles les éléments ont des propriétés similaires. L'algorithme k-mean trouve un groupe de telle sorte que l'erreur quadratique entre la moyenne empirique d'un groupe et les points dans le groupe est réduite au minimum (Jain, 2010). La raison pourquoi on choisit k-mean est que l'algorithme est facile à utiliser, de plus la sortie de k-mean est compréhensible.

D'autre part, dans notre travail, l'extraction des liaisons entre les données nécessite l'application de la méthode d'exploration relationnelle. Selon Baker (2010), le but de l'exploration relationnelle est de trouver des relations entre les variables. Dans son papier, il décrit quatre types de l'exploration relationnelle: association rule mining, correlational mining, sequential pattern mining, causal data mining.

Parmi ces types de l'exploration relationnelle, causal data mining sera utilisé pour l'hypothèse trois. Selon (Baker, 2010) le but de causal data mining est de chercher s'il existe une liaison entre ces deux événements, c.à.d. si un événement est déclenché par un autre événement. Dans notre cas, pour l'hypothèse 3, on cherche la relation entre les infos pris par l'enseignant (comportements des élèves etc) et la rétroaction générée par l'enseignant, surtout les raisons de ces rétroactions. D'où ça sera utile d'appliquer causal data mining méthode dans notre cas.

D'autre part, pour la première hypothèse, on a besoin d'utiliser l'exploration des patterns séquentiels pour extraire des patterns. Surtout pour pouvoir analyser le regard des enseignants, cette méthode peut arriver à extraire des patterns dans les données consécutives.

Les algorithmes les plus connus sont GSP (Generalized Sequential Pattern), PrefixSpan (Prefix projected Sequential pattern mining), SPADE (Sequential Pattern Discovery using Equivalence classes) et SPAM (Sequential Pattern Mining).

5.1.1. GSP

Pendant le premier tour, il trouve les objets le plus fréquents qui composent un seul objet le plus fréquent. Chaque sous-séquence commence avec une graine (seed en anglais) reliée avec le pattern trouvé. Cette graine sera utilisée pour les prochains patterns possibles. Chaque candidat de pattern contient au moins un objet de la graine. L'algorithme se termine quand il n'y a plus de candidat pour faire partie du pattern séquentiel (Pei et al., 2004).

5.1.2. PrefixSpan

L'algorithme de Prefix est un dérivé de FreeSpan. Dans FreeSpan l'idée est d'utiliser des éléments fréquents et de projeter de manière récursive les bases de données de séquences dans une petite base de données projetée et grandir les fragments de sous-séquences dans chaque base de données projetée. Par contre dans l'implémentation de Prefix, il y a moins de projections et des séquences rétrécissent rapidement (Pei, 2001).

5.1.3. SPADE

Les caractéristiques principales de SPADE est que l'algorithme décompose l'endroit où le pattern est cherché. Chaque sous-groupe peut être exécuté indépendamment. De plus, cet algorithme utilise la liste d'identification verticale pour le format de base de données. SPADE non seulement réduit les coûts d'E/S en réduisant les analyses de bases de données, mais aussi réduit les coûts informatiques en utilisant des systèmes de recherche efficaces (Zaki, 2001).

5.1.4. SPAM

SPAM est un nouvel algorithme pour trouver toutes les séquences fréquentes au sein d'une base de données transactionnelle. L'algorithme est particulièrement efficace lorsque les patterns séquentiels dans la base de données sont très longs. Une stratégie de recherche en profondeur est d'abord utilisée pour générer des séquences de candidats, et divers mécanismes d'élagage sont mis en œuvre pour réduire l'espace de recherche (Ayres et al., 2002).

Dans notre travail on utilise une version de GSP nommée l'algorithme Hirate-Yamana. Hirate-Yamana est un algorithme pour découvrir des patterns séquentiels fréquents sur certaines contraintes de temps pour filtrer les modèles inintéressants. L'idée d'utiliser des contraintes de temps est intéressante car elle permet de réduire considérablement le nombre de patterns trouvés et il est également plus rapide et utilise moins de mémoire que si tous les modèles étaient découverts (Hirate & Yamana, 2006) et algorithme (peut varier selon les versions) en général prend 5 paramètres. Le premier est le pourcentage d'existence de pattern dans toutes les séquences. Le deuxième et le troisième paramètre

sont min et max l'intervalle de temps entre deux objets consécutifs dans le même pattern. Le quatrième et le dernier paramètre sont encore la min et max l'intervalle de temps entre le début et la fin de pattern à chercher.

Pour l'algorithme Hirate-Yamana, il faut donner un fichier d'entrée. Ce fichier a un format spécial comme la Figure 5.1:

```
<0> 1 -1 <1> 1 2 3 -1 <2> 1 3 -1 -2
<0> 1 -1 <1> 1 2 -1 <2> 1 2 3 -1 <3> 2 -1 -2
<0> 1 2 -1 <1> 1 2 -1 -2
<0> 3 -1 <1> 1 2 3 -1 -2
```

5-1. Le format pour l'algorithme Hirate-Yamana

Les timestamps sont écrits entre '<' et '>'. Puis les données consécutives sont présentées. Chaque séquence doit finir avec "-1" et la ligne doit finir avec "-2".

5.2. Outils et Plateformes

5.2.1. Oculométrie (eye-tracking)

Par définition, L'oculométrie tout simplement est la mesure du mouvement des yeux. Cette technique regroupe un ensemble de méthodes permettant d'enregistrer les mouvements oculaires. Les oculomètres les plus courants analysent des images de l'œil humain capturées par une caméra souvent en lumière infrarouge, pour calculer la direction du regard du sujet. En fonction de la précision souhaitée, différentes caractéristiques de l'œil sont analysées. D'autres techniques sont basées sur les variations de potentiels électriques à la surface de la peau du visage ou encore sur les perturbations induites par une lentille spéciale sur un champ magnétique.

Avant son utilisation dans l'exploitation des données de l'éducation, il est utilisé en psychologie (Rayner, 1995 ; Rayner, 1998). Les chercheurs ont également commencé à examiner les données du regard comme une source d'information pour modéliser les processus cognitifs pertinents des utilisateurs (Kardan & Conati, 2012). Les données du regard ont été étudiées:

- Pour capturer les processus de prise lors de tâches de recherche d'information (Simola, 2008 ; Day, 2010).
- Pour la reconnaissance de l'activité au cours du travail (Courtemanche et al., 2011),
- Pour prédire la pertinence d'un mot dans une tâche de lecture (Loboda et al., 2011)
- Pour prédire comment les utilisateurs traitent une visualisation de l'information donnée (Loboda & Brusilovsky, 2010 ; Conati et al., 2011)
- Pour estimer la charge de travail mental (Iqbal et al., 2005)

Dans le papier (Muldner et al., 2009), a porté sur la dilatation des pupilles pour détecter les états affectifs de l'utilisateur. (Knoepfle et al., 2009) a utilisé des données eye-tracking pour comparer les théories existantes pour des jeux de stratégies.

5.2.2. ANVIL

Anvil (Annotation of Video and SpokenLanguage) est un outil de recherche pour l'annotation de vidéo. C'est un outil qui est utilisé dans plusieurs domaines comme Interface homme-machine, Linguistique, la recherche des gestes. ANVIL permet une interface graphique pour créer les annotations temporelle. Ces annotations peuvent être sous forme de différents types comme string, booléen etc. Les éléments non-temporels peuvent être aussi annotés (Kipp, 2005).

5.2.3. Orange

Orange est un framework global, basé sur des composants pour l'extraction de données d'expérience. L'utilisation de l'Orange Canvas et des widgets permet d'explorer visuellement de données de conception par scripts Python courts.

Orange est une bibliothèque d'objets C++ de base et des routines qui comprend une grande variété d'algorithmes d'apprentissage machine et exploration de données standard et non-ainsi-standard, plus les routines pour l'entrée et la manipulation des données. Orange est également un environnement de script-mesure pour le prototypage

rapide de nouveaux algorithmes et systèmes d'essais. Il s'agit d'une collection de modules en Python qui sont assis sur la bibliothèque de base et mettent en œuvre certaines fonctionnalités pour lesquelles le temps d'exécution n'est pas cruciale et qui est plus facile à faire en Python qu'en C++. Orange est également un ensemble de widgets graphiques qui utilisent des méthodes de la bibliothèque de base et des modules d'Orange et fournissent une interface d'utilisateur agréable. Les Widgets sont conçus sur la métaphore de communication fondée sur le signal et peuvent être assemblés dans une application par un outil de programmation visuel appelé orange Canvas (Demsar et al., 2004).

La flexibilité et la facilité d'usage sont des raisons principales pour choisir la plateforme Orange. Dans plusieurs cas, il ne nécessite pas la connaissance sur la programmation. Il offre aux utilisateurs la programmation visuelle. Aussi il permet de visualiser les données sur les tableaux. C'est très pratique d'appliquer les méthodes de l'exploitation des données. De plus c'est gratuit et " open source ", prêt à développer.

5.2.4. CLASS

CLASS, Classroom Assessment Scoring System est une méthodologie d'observation et d'évaluation de l'enseignant. Les dimensions de classe sont basées sur la théorie du développement et de la recherche suggèrent que les interactions entre les élèves et les adultes constituent le principal mécanisme de développement et de l'apprentissage des élèves. Dans ce travail, on n'a pas forcément utilisé CLASS mais une partie des données qu'on a utilisé sont venue par CLASS (Pianta et al., 2008).

5.2.5. TDOP

TDOP (Teaching Dimensions Observation Protocol) est un protocole structuré conçu pour capturer les comportements des enseignants. TDOP a été développé dans le cadre d'une étude empirique des déterminants des pratiques d'enseignement postsecondaire. Dans notre travail, TDOP est utilisé pour déterminer les événements pédagogiques dans les cours et les classifier. Voici les types des EVP et leurs caractéristiques (Hora et al., 2013).

5.2.5.1. Au niveau de l'enseignant

C : Cours. L'enseignant parle aux élèves sans utiliser aucune aide visuelle ni équipement

C-AV : Cours, mais avec des aides préparées (transparents, posters, notes pré-écrites au tableau, etc.). L'enseignant peut référer au sujet affiché.

C-EV : Cours avec visuels écrits en direct. L'enseignant fait le cours et écrit au fur et à mesure, présente des notes.

C-Démo : Cours étayé par une démonstration d'un sujet ou phénomène. Un équipement est utilisé (simulation info, objets physiques autres que notes écrites) pour faire comprendre un contenu.

TRANS : L'enseignant annonce le contenu ou le plan de ce qui va se passer ensuite, indique clairement la transition entre 2 activités.

ORG : L'enseignant fait une annonce, discute ou est engagé dans une activité à propos de points hors de la tâche en cours (p. ex. cantine, logistique, problèmes élèves extrascolaires, etc.)

5.2.5.2. Au niveau de l'interaction entre l'enseignant et les élèves

COURS-INT : Cours interactif, l'enseignant parle aux élèves (ou à un élève) en posant des questions multiples et successives auxquelles les élèves (ou l'élève) répond (ent), et ces réponses sont guidées ou intégrées à la discussion. Il va pouvoir servir à récupérer les rétroactions fines.

EXE-INT : Exercice interactif; comme **COURS -INT**, mais des exercices sont réalisés collectivement (souvent au tableau mais pouvant être sur ardoise ou cahier), et les élèves donnent des réponses. Il va pouvoir servir à récupérer les rétroactions fines.

D-GR : Discussion en petits groupes. Les élèves forment des groupes pour travailler ou discuter. L'enseignant peut passer dans les groupes et discuter avec les élèves. Elle va pouvoir servir à récupérer les rétroactions fines.

Disc-Classe : Discussion en classe entière où l'enseignant mène une discussion ouverte où les élèves répondent/ se posent des questions entre eux (p. ex., créent, résolvent un problème). L'enseignant dirige moins les échanges que dans **C-INT**. Assez proche de

D-GR, mais en classe entière et pour un contenu plus ouvert. Elle va pouvoir servir à récupérer les rétroactions fines.

5.2.5.3. Au niveau des élèves

Trav-IND : Les élèves travaillent individuellement à leur bureau (les consignes liées au travail sont intégrées et non codées TRANS).

Vidéo : L'enseignant diffuse une vidéo sans parler pendant que les élèves la regardent.

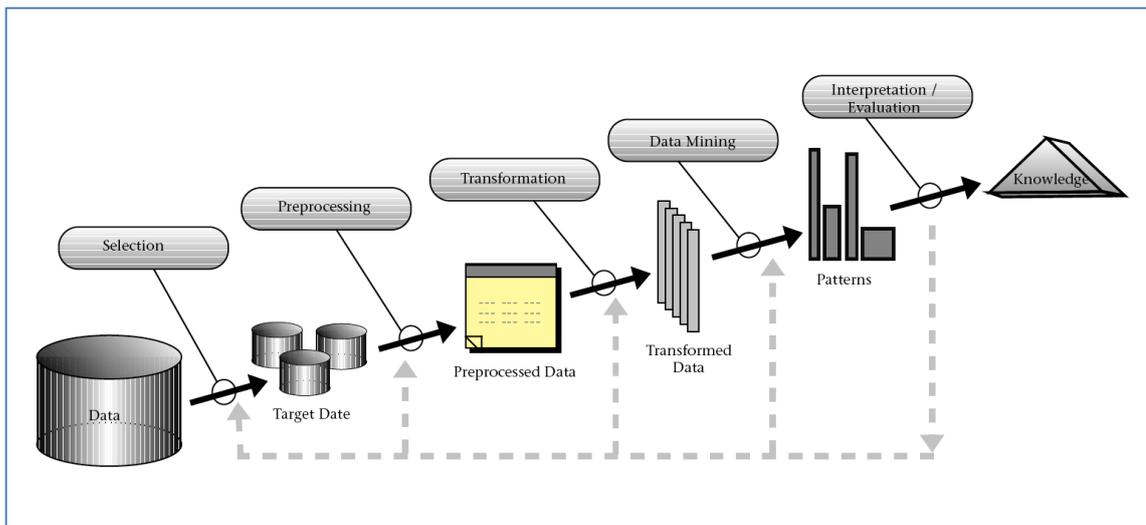
Prés-E : Les élèves font une présentation orale à la classe et dirigent le cours (l'enseignant étant en retrait).

5.2.6. Sequential Pattern Mining Framework

Sequential Pattern Mining Framework (SPMF) est une bibliothèque de l'exploration de données développé en Java. La plateforme permet aux utilisateurs d'utiliser plus de 52 algorithmes d'exploration de données pour découvrir des motifs séquentiels, les règles séquentielles, règles d'association, les articles le plus fréquemment, clusters etc (Fournier, 2011).

5.3. Concevoir de processus

Il existe des étapes qu'il faut suivre avant d'appliquer les méthodes d'exploration de données. Selon Fayyad (Fayyad et al., 1996) il existe 3 étapes qu'il est nécessaire de faire après avoir pris les données expérimentales. Le processus entier est donné dans la Figure 5.2.



5-2. Le processus entier

Pendant l'étape Sélection (aussi nommé Réduction de données), parmi les données, on choisit les attributs reliés avec notre analyse. Au lieu de s'intéresser à toutes les données, on élimine les attributs qui ne sont pas reliés avec l'analyse. Deuxième étape à suivre est le prétraitement. Dans cette étape, le but est d'éliminer les données inutiles, des données incomplètes. Et la dernière étape avant le traitement c'est la transformation. Pour l'analyse des données, on utilise plusieurs outils. Chaque outil a besoin d'une entrée avec un format spécial. Pendant cette étape, on fait la transformation des données dans un format acceptable par les outils qu'on utilise.

Dans le premier temps, on a appliquée l'annotation des vidéos en utilisant l'outil ANVIL. A l'aide de cette l'interface, on a créé le schéma d'annotation. Dans un schéma d'annotation on définit toutes les pistes c.à.d. on décide ce qu'on va annoter dans le vidéo. Par exemple, dans notre travail on s'intéresse au regard de l'enseignant et pour le déterminer on a créé une piste. Les attributs de cette piste sont élèves, ardoises, cahiers, tableau de classe et autres. Après avoir déterminé les attributs, le processus d'annotation commence. On lit la vidéo. Dès que l'enseignant regarde un des nos attributs, en utilisant le timestamp (début et la fin du regard) on fait l'annotation. Donc sur la piste, on crée un segment en utilisant le début et la fin du regard. Puis on met la valeur sur ce segment comme a5 pour regarder l'ardoise de l'élève 5, c7 pour regarder le cahier de l'élève 7, e9 pour regarder l'élève 9 etc.

A la fin de processus de l'annotation, on prend la sortie d'outil. La sortie d'ANVIL est donnée dans la Figure 5.3. Dans la figure, on voit seulement la sortie de piste d'élève. La première colonne est le nom de piste, la deuxième colonne est le début de regard, la troisième est la fin de regard, la quatrième colonne est tout simplement la durée de regard et la dernière colonne montre l'identifiant de l'élève. Si on considère la première ligne, l'enseignant a regardé l'élève no 9 dans l'intervalle $t = [0, 1.76]$.

track	start	end	duration	no_eleve
eleves	0	1.76	1.76	9
eleves	1.76	2.24	0.48	7
eleves	3.44	3.56	0.12	3
eleves	6.24	6.52	0.28	1
eleves	6.68	7.04	0.36	7
eleves	7.08	7.4	0.32	9
eleves	11	11.95	0.96	9
eleves	11.99	12.43	0.44	7
eleves	12.43	12.83	0.4	14
eleves	12.83	13.75	0.92	16
eleves	13.79	13.99	0.2	4
eleves	13.99	14.91	0.92	21
eleves	14.91	15.15	0.24	6
eleves	20.35	20.55	0.2	2
eleves	24.87	25.07	0.2	20
eleves	26.23	26.43	0.2	5
eleves	30.59	30.79	0.2	7
eleves	32.95	33.07	0.12	21

5-3. La sortie de l'outil ANVIL

De cette façon, on applique l'annotation de tous les comportements ou les événements déroulant pendant la séance qu'on a déterminée au début (tours de parole de l'enseignant, événement pédagogique, regard enseignant etc.). Les données venant d'ANVIL sont quantitatives mais aussi on a des données qualitatives venant des autres outils comme CLASS et les fichiers remplis par les enseignants.

Après avoir reçu toutes les données, on passe à l'étape du prétraitement. Cette étape est très importante pour l'exploration de données. On applique les processus qu'on vient de présenter sur nos données. Premièrement, on décide les données qu'on va utiliser, c.à.d. faire on va faire la sélection. Pour chaque hypothèse, on utilise soit une source de données différente soit un attribut différent d'une même donnée. Par exemple pour l'hypothèse 2, on utilise seulement le regard d'enseignant parmi toutes les données d'ANVIL. Puis on élimine les données qu'on ne va pas utiliser. Dans l'hypothèse 2, on

n'utilise pas les données de début et la fin de regard. Après avoir choisi les données, on fait une transformation de ces données dans un format acceptable pour des outils.

Comme dans toutes les plateformes de l'exploration de données, dans Orange, il existe des widgets pour faire le prétraitement. Pour analyser les données et prouver les hypothèses, on a utilisé plusieurs widgets existants et dans le cas où il n'existe pas un widget pour notre besoin, on a implémenté notre propre widget. Les widgets qu'on a créés seront présentés dans la section suivante.

5.4. Les widgets créés

Pour l'analyse des données, dans ce travail on a utilisé la plateforme Orange. Cette plateforme nous permet d'analyser, de traiter des données en utilisant des widgets existants. Dans le cas où il n'y a pas de widget convenable pour notre travail, on a créé nos propres widgets. La plateforme Orange permet aux développeurs de créer des widgets nouveaux et réutiliser ces widgets sur des autres processus.

Tout au début, on a besoin d'accéder aux dossiers, de récupérer des données et de les visualiser sur un tableau. Pour cela, il existe le widget "File" pour récupérer des données par un fichier et il est possible de visualiser des données avec l'aide d'un autre widget "DataTable". Par contre, on a plusieurs classes et plusieurs fichiers. Orange ne supporte pas d'ouvrir plusieurs fichiers avec un seul widget. D'où on a dû créer et implémenter notre propre widget. On a un fichier ayant les notes de comportements de chaque élève. Le widget créé est nommé "VisualizeStudentInfo". Chaque widget a une entrée et une sortie. L'entrée de ce widget est le fichier qui contient les données. Le fichier doit se situer dans un dossier nommé "classe_NO". Les noms des dossiers doivent être sous forme de classe_1, classe_2 etc. Les données de comportements des élèves se trouvent dans le fichier nommé "classe_NO_comportement.csv". Dans ce fichier on a le numéro d'élève, leur nom et prénom, leur date de naissance, leur sexe, leurs notes de français, de mathématiques et les notes de comportements. Le widget doit ouvrir le fichier de comportement (pour chaque classe) et régler les colonnes. À la fin, la sortie de widget sera un tableau sous forme de "DataTable" dans lequel on met le numéro de classe, le numéro d'élève et leurs notes. Plus encore dans les classes on peut

avoir des élèves qui ont besoin d'aide pendant les cours. Ces élèves sont aussi indiqués dans le même fichier, donc dans le tableau de sortie pour ces élèves on met un signe spécial pour rappeler ces élèves. La Figure 5.4 montre le tableau qui contient les données des élèves. Si on prend la première ligne, on voit le premier élève de classe no 1, sa note de français est 2 sur 4, sa note de mathématiques est 2 sur 4. Puis on voit 13 notes de comportement et à la fin on voit s'il a besoin d'aide ou non.

No Classe	No Eleve	FrenchGrade	MathGrade	Comportement1	rter1	rter2	rter3	rter4	rter5	rter6	rter7	rter8	rter9	rter10	rter11	rter12	rter13	Comportement13	Special Case
1	1	2	2	3	1	3	2	3	3	2	2	4	4	4	4	4	4		Non
1	2	4	4	4	1	4	3	4	4	4	2	4	4	4	4	4	4		Non
1	3	3	3	3	2	4	4	4	4	4	1	4	4	4	4	4	4		Non
1	4	2	2	3	2	4	4	4	4	3	2	4	2	4	4	4	4		Non
1	5	1	1	3	3	3	4	4	3	3	4	0	0	2	1	4		Oui	
1	6	4	4	4	1	4	4	4	4	4	1	4	4	4	4	4		Non	
1	7	2	2	2	2	4	4	4	4	4	2	4	4	4	4	4		Non	
1	8	1	1	2	3	4	3	3	2	2	4	4	4	4	2	4		Non	
1	9	1	1	1	4	2	2	2	1	1	4	4	3	2	2	4		Oui	

5-4. La visualisation des données des élèves

Pour calculer les notes et trouver l'indicateur d'élève pour chaque élève, on a besoin d'implémenter un widget pour faire le calcul. Ce widget est nommé "CalculGrade". L'entrée de ce widget est la sortie de widget "VisualizeStudentInfo". Pour calculer l'indicateur d'élève, on a besoin des coefficients pour chaque note. Normalement l'indicateur peut être calculé par prendre la moyenne des notes par contre on a un côté des notes de comportements et de l'autre côté des notes de cours. Comme l'importance de chaque note est différente, on a décidé de mettre des coefficients. D'où l'utilisateur doit saisir 15 coefficients (note de français, mathématiques, notes de 13 comportements). Les coefficients serviront pour calculer le résultat des notes. Puis on a besoin de convertir le résultat au format en lettre pour trouver l'indicateur d'élève. Le convertisseur au format en lettre est un des moyens pour classifier des élèves. Avec l'interface, l'utilisateur peut définir les frontières de barème et déterminer les intervalles pour la note Faible, Passable, Moyen, Bien, Très Bien. Pour des élèves qui ont le besoin d'aide, on a mis une colonne spéciale en indiquant si l'élève a un cas spécial (oui/non). Pour ces élèves, on a mis " autre " comme l'indicateur d'élève (sous forme de lettre). Pour la visualisation des notes, on ajoute deux colonnes à la fin du tableau actuel.

Les deux widgets présentés jusqu'à maintenant sont reliés avec des données de comportements. De l'autre côté, on a aussi les données du regard (suivi de regard) c.à.d. les données qui montrent où/combien de temps l'enseignant regarde. Pour ces données, on a besoin d'un widget pour les visualiser et un autre widget pour les analyser. Pour la visualisation, on crée un widget nommé "VisualizeRegardInfo". L'entrée de ce widget sera le fichier ayant le chemin disons "./classe_1/classe_1_pistage.csv". La sortie du widget est tout simplement une table sous forme de "DataTable" ayant les colonnes: Numéro de classe, durée d'ardoise, numéro d'ardoise, durée d'élève, numéro d'élève, durée de cahier, numéro de cahier, durée de tableau, durée d'autre. La Figure 5.5 montre les données de suivi de regard. Si on prend la première ligne, l'enseignant a regardé l'ardoise de l'élève 3 pendant 0.32 sec, l'élève 9 pendant 1.76 sec, le cahier de l'élève 6 pendant 0.16 sec, le tableau de classe pendant 0.64 sec et d'autre endroit pendant 0.25 sec.

No Class	Duration Ardoise	No Ardoise	Duration Eleve	No Eleve	Duration Cahier	No Cahier	Duration Tableau	Duration Autres
1	0.32	3	1.76	9	0.16	6	0.64	0.35
1	0.40	1	0.48	7	0.40	9	0.96	0.24
1	0.40	9	0.12	3	0.76	14	3.20	1.28
1	0.32	8	0.28	1	0.04	7	4.76	0.12
1	0.56	10	0.36	7	0.08	7	0.36	0.16
1	0.40	11	0.32	9	0.48	12	1.12	0.04
1	0.56	13	0.96	9	0.04	10	0.44	0.16
1	0.28	12	0.44	7	0.12	19	0.36	0.24
1	0.24	14	0.40	14	0.20	20	0.52	0.04

5-5. visualisation des données de regard

Après avoir visualisé les données de regard, on a besoin d'un autre widget pour calculer la somme de regard pour chaque élève, c.à.d. la somme de durée d'ardoise, durée d'élève, durée de cahier va présenter la durée de cet élève. On a créé le widget "CalculSumRegard". L'entrée de ce widget sera la sortie du widget précédent pour récupérer les données et la sortie sera tout simplement un tableau (encore sous forme de format "DataTable") avec des colonnes numéro de classe, numéro d'élève, durée d'élève, durée d'ardoise, durée de cahier et la somme.

D'autre part, on a besoin d'un widget pour faire la transformation des données dans un format convenable pour l'entrée de l'algorithme. Par exemple pour l'hypothèse 1, on analyse les événements pédagogiques et pour chaque EVP, on a une liste des regards dans lesquels on cherche l'existence d'un pattern. D'où pour pouvoir utiliser un

algorithme de l'exploration de pattern séquentiel, on doit transformer les données. Pour l'analyse, l'algorithme Hirate-Yamana est utilisé. Comme cet algorithme a une entrée spécifique, avec le widget créé " Convert2HY ", les données sont transformées dans un format acceptable par l'algorithme.

5.5. Les analyses réalisées pour chaque hypothèse

5.5.1. Analyse pour l'hypothèse 1

Dans l'hypothèse 1, on essaie de prouver l'existence de patterns dans les comportements des enseignants pour les mêmes EVP. L'algorithme Hirate-Yamana est choisi pour trouver des patterns séquentiels grâce à la plateforme SPMF. Comme on n'a pas de données réels, on a créé des données synthétiques avec une méthode de cette plateforme. Le nom de la méthode est " generate a sequence database with timestamp ". L'interface de cette méthode est donnée dans la Figure 5.6.

Choose an algorithm:	Generate_a_sequence_database_with_timesta... ▼	
Set output file	output.txt	...
Choose sequence count:	100	(e.g. 100)
Choose max. distinct items:	22	(e.g. 1000)
Choose item count by itemset:	1	(e.g. 3)
Choose itemset count per sequence:	40	(e.g. 7)

5-6. La génération des données synthétiques

On saisit quatre paramètres pour la génération. Le premier paramètre est le nombre de séquence à générer. Le deuxième paramètre est l'intervalle pour générer des valeurs. Dans notre cas, on a mis 22 pour ce paramètre, car chaque classe contient 22 élèves. Ainsi, après avoir généré les données on pourra déterminer quel élève l'enseignant regarde le plus souvent. Le troisième paramètre sert à limiter le nombre de valeur associé à un timestamp. Dans notre cas, l'enseignant ne peut pas regarder au même temps deux élèves, d'où on a mis " 1 " pour ce paramètre. Donc pour chaque timestamp on doit avoir un identifiant d'élève. Le dernier paramètre est la longueur d'une

séquence, c.à.d. la longueur de la séquence nous donne le nombre d'élèves que l'enseignant a regardé pendant cette séquence. Les membres de la séquence correspondent aux identifiants des élèves.

L'utilisation de cette génération a deux points négatifs. Le premier est que la longueur de chaque séquence est la même. En réalité, pendant chaque EVP, l'enseignant ne regarde pas le même nombre d'élèves. Deuxièmement, la génération des valeurs (entre 1 et 22 pour nous) est bien distribuée c.à.d. dans toutes les données, le nombre d'occurrence de chaque élève est presque la même. On ne peut pas extraire des patterns dans les données de regard bien distribuées. Pour ces raisons, on a pris la sortie de cette méthode mais modifié les longueurs des séquences aléatoirement donc dans la nouvelle version la longueur des séquences différentes, c.à.d. le nombre d'élève que l'enseignant a regardé est devenu différents. En outre, on a essayé d'ajouter les mêmes valeurs pour trouver un pattern.

Pour la continuité de travail, on a du faire la transformation des données dans un format convenable pour l'algorithme de Hirate-Yamana. Pour l'entrée de transformation, on a le fichier .csv dans lequel on a les types d'EVP bien séparés (voir les types des EVP dans TDOP). Pour chaque EVP, on a une séquence des identifiants des élèves que l'enseignant a regardés dans l'ordre. Un exemple de fichier csv est donné dans la Figure 5.7. Dans ce fichier on a tous les EVP de type TRANS. Si on prend la première ligne, dans la TRANS1, l'enseignant a regardé d'abord l'élève 3, puis l'élève 6 ensuite 7 etc.

TRANS1,3,6,7,8,3,9,10,21,1,4,6,14,17,19,9,20,21,6,8,9,19,10,2
TRANS2,4,8,10,12,4,21,11,12,12,16,17,3,8,12
TRANS3,21,20,6,7,2,6,6,9,12,18,23,19,12,18,19,6,5,3,9,11,17,19,2,1,3
TRANS4,2,4,14,15,19,10,9,14,11,19,20,5,1,11,16,17

5-7. Les séquences des identifiants des élèves

On utilise le widget " Convert2HY " qu'on vient de créer. Avec le widget, nos données sont devenues comme dans la Figure 5.8. L'algorithme prend chaque l'identifiant d'élève et met dans un format " < indice > ID -1 ". Ici ID signifie l'identifiant d'élève. A la fin de chaque séquence, l'algorithme met -2 pour terminer.

```

<0> 3 -1 <1> 6 -1 <2> 7 -1 <3> 8 -1 <4> 3 -1 <5> 9 -1 <6> 10 -1 <7> 21 -1 <8> 1 -1 <9> 4 -1 <10> 6 -1 <11> 14 -1 <12> 17 -1
<13> 19 -1 <14> 9 -1 <15> 20 -1 <16> 21 -1 <17> 6 -1 <18> 8 -1 <19> 9 -1 <20> 19 -1 <21> 10 -1 <22> 2 -1 -2
<0> 4 -1 <1> 8 -1 <2> 10 -1 <3> 12 -1 <4> 4 -1 <5> 21 -1 <6> 11 -1 <7> 12 -1 <8> 12 -1 <9> 16 -1 <10> 17 -1 <11> 3 -1 <12>
8 -1 <13> 12 -1 -2
<0> 21 -1 <1> 20 -1 <2> 6 -1 <3> 7 -1 <4> 2 -1 <5> 6 -1 <6> 6 -1 <7> 9 -1 <8> 12 -1 <9> 18 -1 <10> 23 -1 <11> 19 -1 <12>
12 -1 <13> 18 -1 <14> 19 -1 <15> 6 -1 <16> 5 -1 <17> 3 -1 <18> 9 -1 <19> 11 -1 <20> 17 -1 <21> 19 -1 <22> 2 -1 <23> 1 -1
<24> 3 -1 -2
<0> 2 -1 <1> 4 -1 <2> 14 -1 <3> 15 -1 <4> 19 -1 <5> 10 -1 <6> 9 -1 <7> 14 -1 <8> 11 -1 <9> 19 -1 <10> 20 -1 <11> 5 -1 <12>
1 -1 <13> 11 -1 <14> 16 -1 <15> 17 -1 -2

```

5-8. La sortie du widget "Convert2HY" et l'entrée de l'algorithme Hirate-Yamana

Après avoir préparé les données, on passe à l'application de l'algorithme Hirate-Yamana pour chercher des patterns séquentiels. L'interface est donnée dans la Figure 5.9. L'algorithme prend 5 paramètres. Le premier paramètre est le nombre minimum de séquence qu'il doit contenir ce pattern. Pour qu'on trouve un pattern dont les valeurs sont consécutives, on donne "1 " pour l'intervalle de temps. Le quatrième et le cinquième paramètre sont reliés avec la longueur de pattern désiré.

Choose an algorithm:	HirateYamana	
Choose input file	input.txt	...
Set output file	output.txt	...
Choose minsup (%):	0.8	(e.g. 0.5 or 50%)
Min time interval:	0	(e.g. 0 itemsets)
Max time interval:	1	(e.g. 2 itemsets)
Min whole time interval:	2	(e.g. 0 itemsets)
Max whole time interval:	50	(e.g. 2 itemsets)

5-9. L'interface de l'algorithme Hirate-Yamana

Une des sorties de l'algorithme Hirate-Yamana est présentée dans la Figure 5.10. Selon la sortie, l'enseignant a regardé respectivement les élèves 17, 16, 5, 12, 21 et 11. Ce pattern a été trouvé dans 89 TRANS.

```

<0> 17 -1 <1> 16 -1 <2> 5 -1 <3> 12 -1 <4> 21 -1 <5> 11 -1 #SUP:89

```

5-10. La sorties de l'algorithme Hirate-Yamana

Comme on a aussi les indicateurs d'élève de ces élèves (sujet de l'hypothèse 2), on peut mettre leurs indicateurs au lieu de leurs identifiants. D'où on peut connaître quel type d'élève l'enseignant regarde. La Figure 5.11 montre la version avec les indicateurs.

Dans l'exemple donné, l'enseignant regarde d'abord l'élève qui a l'indicateur " T " puis il regarde l'élève avec l'identifiant " O " qui a besoin d'aide etc. A la fin de ligne l'expression #SUP : 89 montre le nombre de séquence qui contient ce pattern.

```
<0> T -1 <1> O -1 <2> O -1 <3> T -1 <4> B-1 <5> T -1 #SUP:89
```

5-11. Une autre version de séquence

Si l'on prend un point de vue différent, on connaît les notes de ces élèves et surtout dans l'hypothèse 2, on calcule leurs indicateurs d'élèves. D'où l'on peut mettre comme entrée les indicateurs d'élèves au lieu de leur numéro d'identifiants. D'autre part en connaissant les élèves les plus regardés, on peut voir aussi leurs positions dans la classe. Avec cette connaissance, il est possible d'avoir l'idée sur le regard de l'enseignant.

Dans notre cas, on ne peut pas donner les données de temps à l'algorithme Hirate-Yamana, il n'y a pas de différence entre un regard de 10 sec et 1 min. Pour résoudre ce problème, on peut construire une liste de regard en utilisant une proportion entre les données de regard. Par exemple, l'enseignant regarde l'élève no 9 pendant 20 sec et une autre élève no 4 pendant 60 sec, donc on peut construire la ligne de regard comme: <0> 9 -1 <1> 4 -1 <2> 4 -1 <3> 4 -1 -2.

D'autre part, dans l'hypothèse 1, il est indiqué que chez les enseignants qui ont des scores élevés par CLASS pour un EVP, plus grand nombre d'élèves seront scrutées avec une fréquence plus importante. Pour le prouver, on a utilisé les données du regard de l'enseignant. Parmi ces données, on a analysé la somme de fixation de l'enseignant et la distribution de cette fixation sur des élèves d'une manière équilibrée. Le widget "CalculSumRegard" contient la durée de fixation dans sa dernière colonne. Selon le nombre total de fixation et aussi la distribution de cette fixation, on peut comparer ces résultats avec des notes de CLASS pour prouver notre hypothèse. Pour le prouver, on regarde le tableau comme donné dans la Figure 5.12. La première colonne est l'identifiant d'élève, la deuxième colonne donne la durée de regard par l'enseignant. La dernière colonne montre combien de fois il était regardé. L'enseignant dont les regards

sont comme dans la figure n'a pas distribué son regard bien équilibré parce qu'il a regardé un élève pendant 60.39 sec tandis que l'autre élève pendant 4.64 sec.

No Eleve	Total	Counter
3	11.28	26
9	32.60	68
6	17.52	28
1	10.96	32
7	25.24	67
14	11.76	37
8	31.68	58
10	9.60	31
11	15.36	60

5-12. Distribution des regard

5.5.2. Analyse pour l'hypothèse 2

Pour la deuxième hypothèse, on parle d'existence de groupes de pilotage. Selon cette hypothèse, l'enseignant regarde fréquemment à ce groupe pour le déroulement et la continuité de cours. Pour prouver cette groupe, on a besoin des données de regard et aussi l'indicateur d'élève. L'indicateur d'élève sera calculé par des notes de comportement et des notes de cours. Avec l'analyse de regard, on va déterminer à quel élève et avec quel périodicité l'enseignant regarde. Quand on utilise ces deux types de données en utilisant la méthode clustering l'algorithme k-mean, il sera possible de trouver quels élèves constituent ce groupe de pilotage.

Pour trouver l'existence d'un tel groupe, il faut faire un prétraitement avant d'appliquer des méthodes d'exploration de données d'éducation. Dans notre cas, on analyse deux types de données. Le premier type sera relatif aux données du regard venant d'annotation des vidéos par ANVIL. Ces données seront des données quantitatives. L'autre type de données sera des données qualitatives venant des enseignants. On rappelle que les enseignants remplissent un questionnaire pour chaque élève. Avec le widget "CalculSumRegard" on a additionnée la durée d'ardoise (l'intervalle de temps que l'enseignant a regardé l'ardoise de l'élève donc le numéro d'ardoise est en même

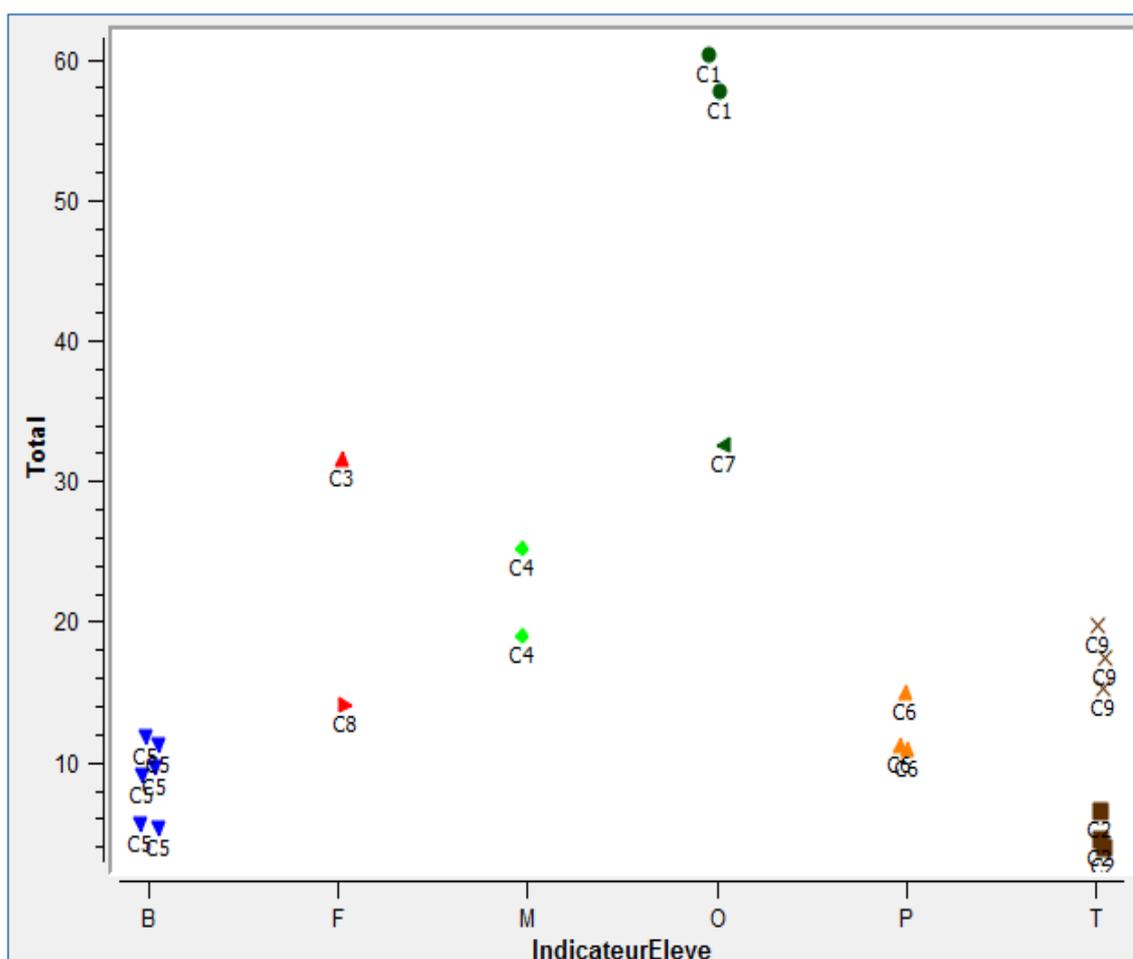
temps le numéro d'élève), la durée d'élève et la durée de cahier. Cette somme nous montre l'intérêt de l'enseignant sur l'élève. D'où à la fin, aux sorties des widgets "CalculGrade" et "CalculSumRegard" on a deux tableaux. Pour trouver une liaison entre les données et construire une relation on a créé un nouveau tableau qui contient les deux avec un widget Multi-Key Merge Data déjà existant dans la plateforme. En le faisant on a choisi le numéro de classe et le numéro des élèves comme les attributs communs dans les deux tableaux. Dans le tableau final, pour chaque élève, on a toutes ses notes, son indicateur d'élève, la durée que l'enseignant a regardé son ardoise, son cahier, soi-même, la totale durée de regard et le compteur de regard.

Parmi des données, on choisit les durées totales du regard et l'indicateur d'élève. Pour prouver l'existence de groupe de pilotage, on a appliqué l'algorithme de k-mean. Comme k-mean est un algorithme non supervisé, il faut donner le nombre de groupe pour diviser les données ou bien c'est possible de choisir l'option " optimized " pour que l'algorithme décide le nombre de classe soi-même de façon optimale. Dans nos expériences, on a choisi cette option et l'algorithme de clustering k-mean a divisé nos données en 9 classes. Parce que l'algorithme répète le clustering 100 fois à l'arrière (100 est le nombre de tour maximum dans Orange). Il donne comme le résultat le plus optimum. La sortie de l'algorithme k-mean est donnée dans la Figure 5.13. La première colonne indique l'indicateur d'élève sous forme de lettre. La deuxième colonne donne la durée moyenne de ce groupe (par exemple la durée moyenne pour le cluster 7 est 59.07 et la dernière colonne montre à quel cluster l'élève appartient.

IndicateurEleve	Total	Cluster
O	59.07	C1
O	32.60	C7
F	31.68	C3
M	22.18	C4
T	17.57	C9
F	14.16	C8
P	12.41	C6
B	8.79	C5
T	5.11	C2

5-13. Le résultat de clustering k-mean

Après avoir appliqué le clustering k-mean, il est possible d'analyser les clusters plus précisément en utilisant les widgets de visualisation qui se trouvent dans la catégorie " Visualisation ". Pour chaque cluster, il permet de montrer les clusters avec leurs membres sur des graphes. La Figure 5.14 montre 9 groupes avec différents couleurs. L'axe x représente l'indicateur d'élève et l'axe y montre la durée total (combien de second l'enseignant regarde a l'élève).

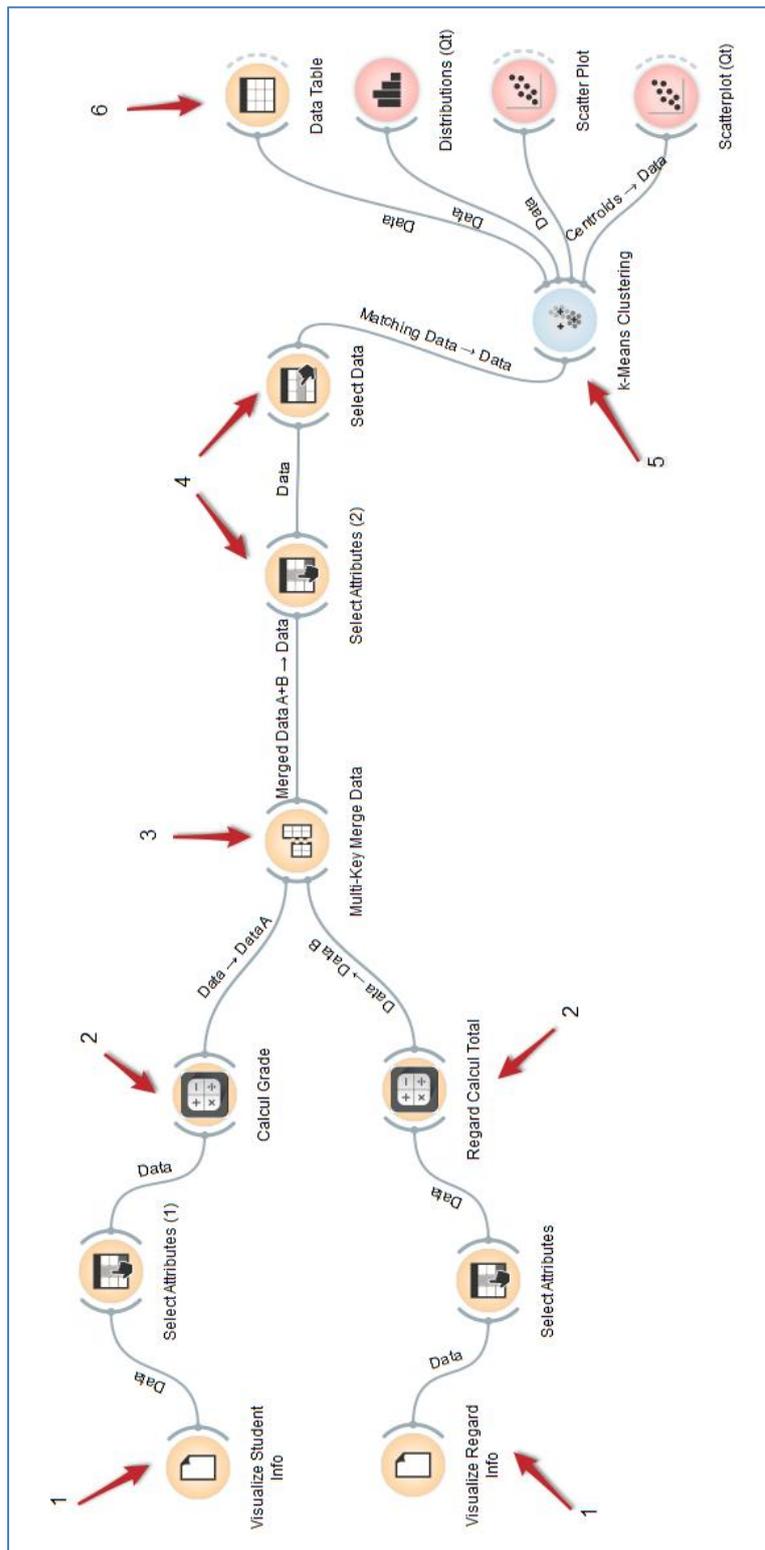


5-14. La visualisation de k-mean

IndicateurEleve	Total	Cluster
O	57.76	C1
O	60.39	C1
T	6.68	C2
T	4.00	C2
T	4.64	C2
F	31.68	C3
M	19.12	C4
M	25.24	C4
B	11.28	C5
B	9.60	C5
B	5.68	C5
B	11.76	C5
B	5.32	C5
B	9.12	C5
P	10.96	C6
P	11.32	C6
P	14.96	C6
O	32.60	C7
F	14.16	C8
T	17.52	C9
T	15.36	C9
T	19.84	C9

5-15. Les membres de clusters

La sortie de l'algorithme k-mean nous montre la relation entre l'indicateur d'élève des élèves et les durées des regards. Par exemple, si l'on considère les groupes 7 et 4 créés par le k-mean, on voit que les moyennes des durées du regard de ces groupes sont plus grandes que les autres groupes. Les membres de ces groupes sont présentés dans la Figure 5.15. Quand on analyse les membres de ces groupes, on voit que plupart de ces élèves ont des cas spéciaux (avoir un cas spécial avait été indiqué dans le fichier de comportement). Si on analyse le groupe C2, on voit que l'indicateur des élèves de ce groupe est totalement "Moyen". Par ces données et ces résultats, on prouve l'existence de groupe de pilotage. Ce groupe de pilotage est constitué par des élèves qui sont regardés le plus souvent par l'enseignant. Tous les traitements réalisés dans cette hypothèse sont présentée dans la Figure 5.16.



5-16. Les traitements complets pour cette hypothèse

Les 6 étapes à exécuter sont les suivants:

Etape 1: Récupérer les données (données de comportements et le regard) par des fichiers et les visualiser.

Etape 2 : Calcul de note finale pour chaque élève et aussi calcul la somme de durée de regard de l'enseignant (élève+son ardoise+son cahier)

Etape 3 : Faire ensemble les deux tableaux en utilisant le widget Multi-Key Merge Data.

Etape 4 : Choix des données pour l'algorithme k-mean avant d'appliquer

Etape 5 : L'application de clustering avec l'algorithme k-mean

Etape 6 : La visualisation de la sortie de k-mean

6. CONCLUSION ET DISCUSSION

Pendant la période de stage, on a travaillé sur le sujet de l'exploration des données de l'éducation qui est un domaine de recherche en émergence. Le but au début était de traiter et d'analyser des données expérimentales pour prouver les hypothèses du projet qui étaient déjà déterminées. Pourtant les données ne sont pas totalement collectées. Pour cette raison, pour certaines hypothèses on a dû générer des données synthétiques afin de tester nos processus.

Tout d'abord on a commencé notre travail avec l'état de l'art pour chercher des papiers similaires au tour de notre sujet. Puis on a passé à la visualisation des données expérimentales. Le premier outil qu'on a utilisé est GazeTracker. Cet outil nous permet de visualiser des données d'oculomètre et de déterminer "area of interest". Par contre comme notre environnement est dynamique, la détermination de "area of interest" devient impossible. Ensuite, on a commencé à utiliser l'outil ANVIL pour faire l'annotation et la création des canaux sur la vidéo. En fait faire l'annotation sur ANVIL est un travail qui doit être réalisé par d'autres personnes dans le projet. On a touché la plateforme ANVIL seulement pour comprendre le processus d'annotation et le type des données générées. Ce processus prend du temps. L'annotation de 10 min peut être réalisée dans 40 heures. C'est pourquoi on n'a pas pu recevoir des sorties d'annotation et pris seulement 10 min de partie de vidéo annoté pour notre analyse. Puis, on a commencé à travailler sur nos hypothèses après avoir généré des données synthétiques.

Dans ce travail, on a bien analysé les données des élèves et essayé de prouver les hypothèses. En faisant l'implémentation sur la plateforme Orange, la structure est créée de façon modulaire, c.à.d. dans le futur, s'il y aura des classes, des enseignants et des élèves supplémentaires, ils peuvent très bien être intégrés dans le système.

Pendant ce projet, on a réussi à travailler sur seulement les deux premières hypothèses. Sur les données, on est arrivé à proposer une analyse pour déterminer à quelle période l'enseignant regarde ses élèves et on a essayé de prouver que pendant le même EVP, l'enseignant regarde avec un ordre. Pour cela, on a utilisé des algorithmes pour trouver des patterns séquentiels.

De l'autre côté, dans l'explication de l'hypothèse c'était indiqué que les enseignants ayant des niveaux élevés de "sensibilité" et de "gestion des comportements" auront un pattern de comportements oculaires spécifiques", par contre comme on n'a pas des vraies données expérimentales, on n'a pas pu le prouver. D'où pour l'hypothèse 1, les méthodes pour analyser sont trouvées, le chemin qu'il faut prendre est proposé et la base du travail dans le futur est constituée.

Pour l'hypothèse 2, on a utilisé seulement 10 min de vidéo déjà annoté. En utilisant la méthode clustering et l'algorithme k-mean on a trouvé que l'enseignant se focalise sur certains élèves dans la classe. On appelle ce groupe le groupe de pilotage. Dans ces analyses on a utilisé des vraies données d'où la conclusion à laquelle on est arrivé est juste. Par contre, le manque de richesse de données nous empêche de faire la généralisation.

De l'autre côté, il existe quelques défis. Tout d'abord le nombre des participants, le nombre des classes ne sont pas suffisants. On avait seulement quatre classes qui ne sont pas nombreux. Ça cause manque des données à analyser. Dans la littérature on parle de l'existence de "area of interest". Par contre dans notre cas, notre environnement est constitué par les élèves et l'enseignant qui ne sont pas statiques. Malgré le manque de "area of interest", on peut trouver des solutions. Par exemple, pour l'hypothèse 1, comme l'entrée de l'algorithme, on peut donner la location des élèves dans la salle de cours. On peut numéroter les élèves selon leur distance au tableau (on peut mettre 2 pour les élèves qui sont à l'arrière de la classe, 1 pour les élèves qui sont en face du tableau). D'où il est possible de déterminer les endroits que l'enseignant regarde le plus fréquemment. Dans notre cas, on a analysé les données du regard qui sont en ordre mais il manquait la durée du regard c.à.d. on n'a pas donné les durées du regard dans

l'analyse. Donc pour une autre version d'analyse on peut éliminer les durées très petites et construire une liste du regard en utilisant la proportion entre les données du regard. Par exemple l'enseignant regarde l'élève no 5 pendant 1 min et un autre élève no 7 pendant 2 min alors on peut construire la ligne du regard comme $\langle 0 \rangle 5 -1 \langle 1 \rangle 7 -1 \langle 2 \rangle 7 -1 -2$. On écrit l'élève 7 deux fois à cause de sa durée.

Comme souligné dans la partie d'état de l'art, dans la littérature il existe peu de travaux intéressants à l'analyse automatique de l'activité de l'enseignant. Pendant ce travail, on a utilisé plusieurs sources de données quantitatives et qualitatives. On pense que le nombre de sources est suffisant par contre la masse des données n'est pas suffisante pour avoir un bon résultat et pour généralisation des idées. Pour la continuité du projet, il est possible d'ajouter de nouvelles hypothèses.

D'autre part, on a aussi quelques propositions pour le futur. Dans le cas actuel on a quatre classes c.à.d. quatre enseignants et dans chaque classe, les élèves sont différents. On croit que, analyser le même enseignant devant différentes classes peut donner des résultats plus évidents. En outre, ce serait bien d'augmenter le nombre de participants pour la richesse de données.

BIBLIOGRAPHIE

Ayres, J., Flannick, J., Gehrke, J., Yiu, T. (2002). Sequential pattern mining using a bitmap representation. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 429-435). ACM.

Baker, R. S. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*, vol.1, no.1, pp.3–17.

Baker, R. S. (2010). Data mining for education. *International encyclopedia of education*, vol.7, pp.112–118.

Brophy, J., McCaslin, M. (1992). Teachers' reports of how they perceive and cope with problem students. *The Elementary School Journal*, 3-68.

Conati, C., Carenini, G., Harati, M., Toker, D., Fitzgerald, N., Flagg, A.(2011). User-Adaptive Visualizations: Can Gaze Data Tell Us When a User Needs Them?. In *Scalable Integration of Analytics and Visualization*.

Courtemanche, F., Aïmeur, E., Dufresne, A., Najjar, M., Mpondo, F. (2011). Activity recognition using eye-gaze movements and traditional interactions. *Interacting with Computers*, 23(3), 202-213.

Day, R. F. (2010). Examining the validity of the Needleman–Wunsch algorithm in identifying decision strategy with eye-movement data. *Decision Support Systems*, 49(4), 396-403.

Demšar, J., Zupan, B., Leban, G., Curk, T. (2004). Orange: From experimental machine learning to interactive data mining (pp. 537-539). Springer Berlin Heidelberg.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (1996). Advances in knowledge discovery and data mining.

Fournier-Viger, P. (2011). Spmf: A sequential pattern mining framework. URL <http://www.philippe-fournier-viger.com/spmf> accessed May, 2014].

Gioia, G. A. (2000). BRIEF: Behavior rating inventory of executive function: Professional manual. Psychological Assessment Resources.

Heiner, C., Heffernan, N., Barnes, T. (2007). Educational data mining. In Supplementary Proceedings of the 12th International Conference of Artificial Intelligence in Education.

Hirate, Y., Yamana, H. (2006). Generalized sequential pattern mining with item intervals. *Journal of computers*, 1(3), 51-60.

Hora, M. T., Oleson, A., Ferrare, J. J. (2013). Teaching Dimensions Observation Protocol (TDOP) User's Manual.

Iqbal, S. T., Adamczyk, P. D., Zheng, X. S., Bailey, B. P. (2005). Towards an index of opportunity: understanding changes in mental workload during task execution. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 311-320). ACM.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651-666.

Kardan, S., Conati, C. (2012). Exploring gaze data for determining user learning with an interactive simulation. In *User Modeling, Adaptation, and Personalization* (pp. 126-138). Springer Berlin Heidelberg.

Kipp, M. (2005). *Gesture generation by imitation: From human behavior to computer character animation*. Universal-Publishers.

Knoepfle, D. T., Wang, J. T. Y., Camerer, C. F. (2009). Studying Learning in Games Using Eye-tracking. *Journal of the European Economic Association*, 7(2-3), 388-398.

Lewis, R. (1999). Teachers coping with the stress of classroom discipline. *Social Psychology of Education*, vol.3, no.3, pp.155–171.

Loboda, T. D., Brusilovsky, P. (2010). User-adaptive explanatory program visualization: evaluation and insights from eye movements. *User Modeling and User-Adapted Interaction*, 20(3), 191-226.

Loboda, T. D., Brusilovsky, P., Brunstein, J. (2011). Inferring word relevance from eye-movements of readers. In *Proceedings of the 16th international conference on Intelligent user interfaces* (pp. 175-184). ACM.

Lundgren, U. P. (1973). *Pedagogical Frames and the Teaching Process. A Report From An Empirical Curriculum Project*.

Marzano, J. S., Pickering, D. (2003). *Classroom management that works: Research-based strategies for every teacher*. ASCD.

Muldner, K., Christopherson, R., Atkinson, R., Burleson, W. (2009). Investigating the utility of eye-tracking information on affect and reasoning for user modeling. In *User Modeling, Adaptation, and Personalization* (pp. 138-149). Springer Berlin Heidelberg.

Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu, M. C. (2001). Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *2001 IEEE 29th International Conference on Data Engineering (ICDE)* (pp. 0215-0215). IEEE Computer Society.

Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Hsu, M. C. (2004). Mining sequential patterns by pattern-growth: The prefixspan approach. *Knowledge and Data Engineering, IEEE Transactions on*, 16(11), 1424-1440.

Pianta, R. C., La Paro, K. M., Hamre, B. K. (2008). *Classroom assessment scoring system*. Baltimore: Paul H. Brookes.

Rayner, K. (1995). Eye movements and cognitive processes in reading, visual search, and scene perception. *Studies in Visual Information Processing*, 6, 3-22.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3), 372.

Ritter, S., Harris, T. K., Nixon, T., Dickison, D., Murray, R. C., Towle, B. (2009). Reducing the Knowledge Tracing Space. *International Working Group on Educational Data Mining*.

Romero, C., Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1), 135-146.

Romero, C. (2010). Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol.40, no.6, pp.601–618.

Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, 78(1), 153-189.

Simola, J., Salojärvi, J., Kojo, I. (2008). Using hidden Markov model to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research*, 9(4), 237-251.

Van den Bogert, N., van Bruggen, J., Kostons, D., Jochems, W. (2014). First steps into understanding teachers' visual perception of classroom events. *Teaching and Teacher Education*, 37, 208-216.

Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. *Machine learning*, 42(1-2), 31-60.

APPENDICE

13 questions sur les activités et les comportements scolaires des élèves:

A une confiance en soi justifiée dans les activités scolaires (se sent capable).

Manque de confiance en soi (éventuellement renonce).

Est capable d'une attention régulière.

Est rapide dans l'exécution d'une tâche.

Est efficace dans l'exécution d'une tâche.

Fait preuve d'autonomie face à une tâche scolaire.

Dans l'exécution d'une tâche, sait anticiper et organiser son travail.

Présente des signes de fatigue pendant les activités scolaires.

Participe activement aux conversations scolaires.

A un bon niveau de langage par rapport au niveau moyen de la classe.

Comprend les consignes de la classe.

Suit naturellement le rythme de la classe (sans besoin particulier de soutien ou de contrainte).

S'intègre bien parmi les enfants de la classe.

BIOGRAPHIE

Gülay Yıldırım est née en 1988 à Sivas en Turquie. Elle est allée au lycée française, Lycée de Galatasaray et diplômé en 2007. Puis elle a commencé son étude licence a l'Université Galatasaray dans le Faculté d'Ingénierie et de Technologie dans le programme Génie Informatique en 2007.

En 2012, elle est diplômée de l'Université Galatasaray et dans le mois Septembre 2012 elle a commencé son master dans l'Institut des Sciences et d'Ingénierie sur le génie informatique. Pendant la première année de master, elle a appliqué pour le programme deux diplômes et est allée en France pour étudier la deuxième année de master à l'Université Joseph Fourier Grenoble. Elle a passé un an à Grenoble et à la fin elle a diplôme de l'UJF.