Informatique Décisionnelle

Découverte d'un outil pour l'intégration de données

Objectifs

Lors de ce TP, vous allez découvrir un outil pour l'intégration de différentes sources de données nommé KNIME. Cet outil, disponible gratuitement pour tout étudiant et enseignant, est multiplateforme : voir le site https://www.knime.com/ pour télécharger le logiciel. Cet outil permet principalement :

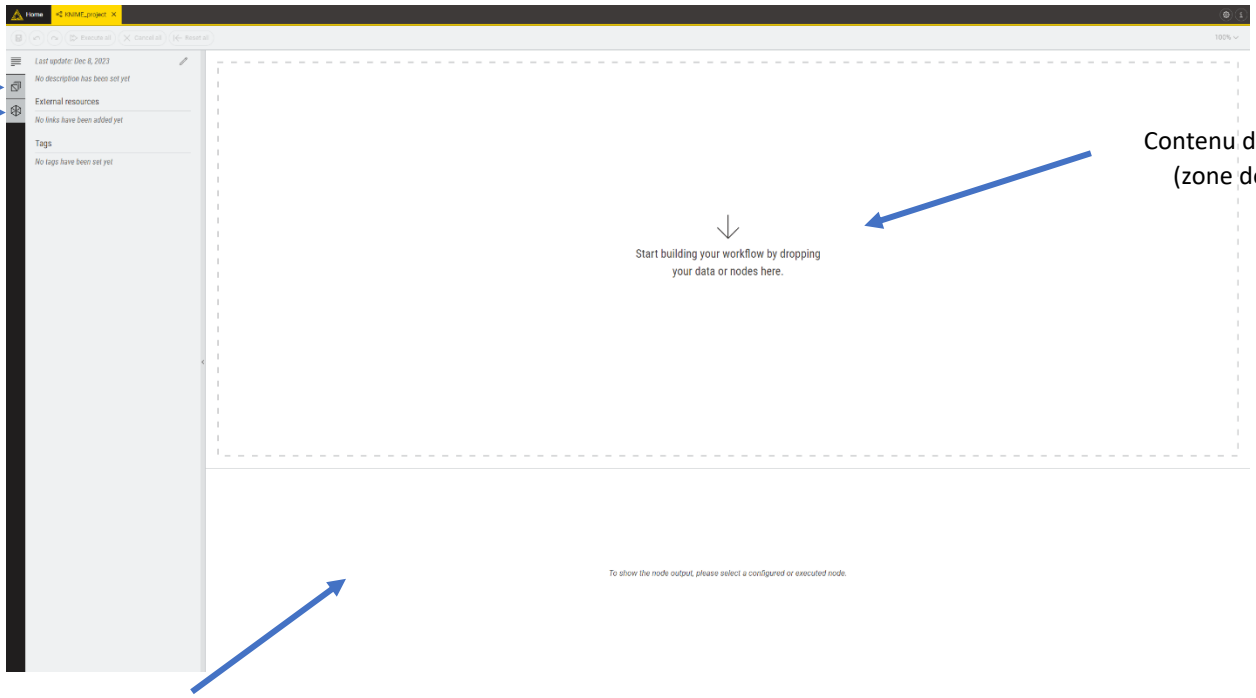
- De se connecter à différentes sources de données, locales ou distantes
- D'intégrer et préparer facilement des données provenant de différentes sources
- D'explorer et visualiser les données intégrées
- De créer divers graphiques (distribution, nuage de points…)

Un tel outil est destiné aux personnes s'occupant régulièrement du nettoyage et de l'intégration des données afin qu'elles puissent être analysées par les analystes et décideurs pour prendre des décisions. L'utilisation des composants intégrés dans KNIME permettent la réalisation de la plupart des tâches de préparation habituelles.

Partie I – Démarrage et présentation de l'interface

Vous trouverez l'outil « knime analytics platform » dans le menu « démarré » des PC de la salle TP. Lancez-le.

1. A l'ouverture, On vous demande de sélectionner l'espace de travail (Workspace) dans lequel vous allez pouvoir développer vos flux de traitements dans un workflow. Sélectionner l'emplacement qui vous convient. Vous pouvez ensuite cliquer sur « Launch ».
2. Une fois le logiciel démarré, vous devrez créer un nouveau workflow. Dans la rubrique « Local space », vous trouverez une icône pour créer un nouveau workflow. Vous pouvez aussi naviguer dans le « Local space » et le créer à partir de là. Nommez votre premier workflow « Amazon Sales ».
3. Une fois le workflow créé, vous devriez avoir cette interface :



4. Avant de passer à la partie suivante, assurez-vous que dans « Preferences > KNIME Modern UI », que « All nodes » soit activé. Vous aurez ainsi accès, à tous les composants proposés par KNIME.

Partie II – Intégration de données sur des produits d'Amazon

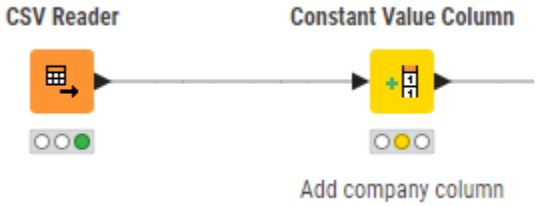
Dans cette section, nous supposons que vous faites partie d'une boite de consulting et que votre objectif est d'intégrer différentes sources de données de sites E-commerces afin qu'elles soient analysables par l'analyste de votre client. Ici, nous allons nous concentrer sur la préparation de données provenant du site Amazon. Voici les principales colonnes du jeu de données :

- PRODUCT ID : Identifiant du produit
- PRODUCT NAME : Nom du produit
- CATEGORY : Catégorie du produit représentée en hiérarchie de sous-catégories.
- DISCOUNTED PRICE : Remise sur le produit
- ACTUAL PRICE : Prix de base du produit
- RATING : Note moyenne du produit
- RATING COUNT : Nombre d'évaluations du produit
- ABOUT PRODUCT : Description du produit
- USER ID : Liste d'identifiants d'utilisateurs qui ont mis un commentaire
- USER NAME : Liste des noms d'utilisateurs qui ont mis un commentaire
- REVIEW ID : Liste d'identifiants des commentaires
- REVIEW TITLE : Liste des titres des commentaires
- REVIEW CONTENT : Liste de descriptions des commentaires

1. Sur Celene, téléchargez le jeu de données « amazon_sales ». Ce dernier contient des produits provenant d'Amazon. Nous allons réaliser quelques préparations classiques dessus.
2. Sur KNIME, dans la liste des composants, sélectionnez le nœud « csv reader » et glissez le dans la zone de travail.
3. Une fois le nœud déposé, double cliquez dessus pour l'ouvrir. Dans « File », sélectionnez le fichier csv « amazon_sales » afin de le charger dans la zone de travail pour y travailler dessus. Enfin, appuyez sur « Apply » pour appliquer les modifications. Vous pouvez ensuite cliquer sur « OK » afin d'enlever le menu du composant.
4. Exécutez le composant afin de charger le fichier. Normalement dans la sortie du nœud, vous devriez voir le jeu de données. Combien de produits (lignes) sont présents dans ce jeu de données ? Plusieurs problèmes de qualité de données sont présents. En regardant le jeu de données, quels sont selon vous, les principaux problèmes de qualité présents qui vont poser des difficultés pour réaliser des analyses ?

   _____
   _____
   _____
   _____
   _____
   _____
   _____

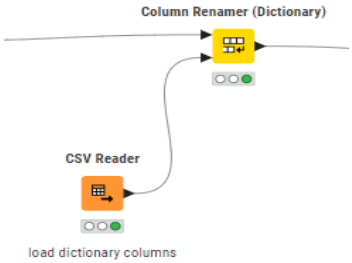
5. Dans un premier temps, il est important d'ajouter une colonne pour préciser de quel site E-commerce les produits proviennent. Pour ce faire, ajoutez un composant « Constant Value Column » à la suite de la chaîne de traitement. Pour ce faire, attachez la sortie du composant « CSV Reader » à l'entrée « Constant Value Column ».



6. Ouvrez ce composant. Ajoutez une nouvelle colonne nommée « company » avec l'option « Append ». Comme valeur, mettez « Amazon » pour préciser de quel site les produits proviennent. Exécutez le composant. Que voyez-vous comme changement à la sortie du nœud « Constant Value Column » ?

   _____
   _____
   _____

7. Une fois que vous avez exécuté le composant, la nouvelle colonne est ajoutée automatiquement à la fin. Pour la positionner au début du jeu de données, attachez à la suite de la chaîne de traitements, un composant « Column Resorter ». Ouvrez le composant et positionnez la colonne « company » au tout début. Utilisez ce composant quand vous souhaitez réorganiser les colonnes.
8. Le typage des colonnes numériques contenant des chiffres ne sont pas de type numérique. Nous allons commencer par les colonnes DISCOUNTED PRICE, ACTUAL PRICE. Ajoutez à la chaîne de traitements un composant « String Manipulation » et ouvrez-le. Dans la partie « Expression », utilisez la fonction « removeChars(str, chars) » pour enlever le caractère ₹. Remplacez « str » par le nom de la colonne (double cliquez sur la colonne dans « Column List » pour l'ajouter dans la partie « Expression ») et « chars » par le caractère à supprimer. Commencez par la colonne « DISCOUNTED PRICE ». Faites la même chose pour la colonne « ACTUAL PRICE » avec un autre composant « String Manipulation ». Si d'autres caractères ("â,ì" par exemple) sont présents au lieu du caractère ₹, réalisez la même opération mais en supprimant les caractères correspondants.
9. La colonne RATING, contient un caractère « | » qui peut poser des problèmes lors de la conversion de cette colonne en type entier par la suite. Utilisez un « String Manipulation » pour supprimer ce caractère.
10. Nous allons maintenant utiliser le composant « String to Number » pour convertir les colonnes contenant des chiffres en type numérique. Ouvrez ce composant. Pour les colonnes citées dans les tâches 8 et 9, utilisez ce composant pour les convertir en type « Number (Double) ». Vérifiez bien qu'il y a seulement ces colonnes dans la partie « Includes ». Les autres doivent être dans la partie « Excludes ». Le séparateur décimal est un «.». Le séparateur des milliers est «,».
11. Faites exactement la même chose pour la colonne « RATING COUNT », mais convertissez-le en type « Number (Integer) ».
12. Pour des raisons de simplicité, votre boite vous demande maintenant d'arrondir les notes d'évaluation (colonne « RATING »). Utilisez le composant « Round Double » pour les arrondir en entier. Une fois ce composant ouvert, pensez à décocher l'option « Append as new columns » et de fixer « Precision » à la valeur « 0 ».
13. On vous a fourni un fichier csv pour changer les noms des colonnes qui ne plaisent pas au client. L'idée est donc de remplacer le nom actuel des colonnes par ceux fournis dans le fichier csv « dictionary_columns » que vous trouverez sur Celene. Chargez-le avec un nœud « CSV Reader ».



14. Pour mettre à jour les noms des colonnes, vous allez utiliser le composant « Column Renamer (Dictionary) ». Ce nœud accepte deux entrées. Attachez à la première entrée, la chaîne de traitements. Attachez à la seconde, le nouveau composant « CSV Reader » contenant le fichier csv « dictionary_columns ».
15. Ouvrez le composant de renommage. A partir du dictionnaire que vous lui avez fourni, il vous demande de préciser la colonne contenant les noms des colonnes à renommer « Lookup column ». Vous devez aussi préciser la colonne contenant les nouveaux noms « Names column ». Enfin, décochez l'option « Fail if no assignment in dictionary table ». Vous pouvez fermer le composant. Exécutez le nœud. Normalement, les colonnes sont renommées à leur sortie.
16. Nous allons maintenant ajouter des informations supplémentaires sur ces produits en ajoutant deux nouvelles colonnes. La première correspond au pourcentage de remise du produit. La seconde, à une note pondérée entre la note moyenne et le nombre d'évaluations du produit. Pour réaliser ces calculs, vous allez utiliser un nœud « Math Formula ». Attachez-le, à la suite de la chaîne et ouvrez-le. Cochez « Append Column » et ajoutez le nom de la nouvelle colonne à calculer « discount_percentage ». Dans la partie « Expression », diviser la remise « discount » par le prix de base « base_price ». Pour le représenter en pourcentage, multiplier cette division par 100. Une fois l'expression écrite, appliquez, quittez et exécutez le composant. Faites la même chose pour la note pondérée qui correspond au produit entre la note moyenne « average_rating » et le nombre d'évaluations « number_of_ratings ». Nommez cette nouvelle colonne « rating_weighted ». Quels sont les produits contenant la note pondérée la plus importante ?

    _____
    _____
    _____
    _____
    _____

17. Nous allons continuer sur quelques traitements sur les catégories. Dans la colonne « category », certaines hiérarchies commencent par la sous-catégorie « Electronics ». Ce niveau de granularité est trop générique. De plus, plusieurs produits électroniques ne commencent pas forcément par « Electronics ». Utilisez un composant « String Manipulation » pour supprimer cette sous-catégorie. Vous pouvez utiliser la fonction « replace » pour remplacer « Electronics| » par un texte vide.

Pour simplifier de futures analyses sur ces catégories, votre client souhaite seulement garder la première et la dernière sous-catégorie de la hiérarchie.

18. Pour ce faire, nous allons d'abord utiliser un composant « Cell Splitter » pour séparer en plusieurs colonnes les sous-catégories. Attachez ce composant à la chaîne et ouvrez-le. Sélectionnez la colonne « category ». Entrez comme délimiteur « | ». En « Output », sélectionnez « As new columns ». Vérifiez bien que l'option « Guess size and column types » est sélectionné. Appuyer sur « Apply », quitter et exécutez le composant. Vous devriez trouver en sortie plusieurs nouvelles colonnes « category_Arr[i] » où i varie entre 0 et 6. « category_Arr[0] » correspond à la première sous-catégorie tandis que « category_Arr[6] » correspond à la dernière sous-catégorie.

19. Pour chaque hiérarchie nous souhaitons garder la première et la dernière sous-catégorie respectivement en début et fin de la hiérarchie. Pour récupérer ces valeurs, utilisez le composant « Column Aggregator » et attacher le à la chaîne et ouvrez-le. Dans la partie « Aggregation column(s) », sélectionnez seulement les colonnes « category_Arr[i] ». Cliquez maintenant sur la rubrique « Options ». Sélectionnez les méthodes d'aggrégation « First » et « Last ». Renommez les deux colonnes agrégées en les nommant « main_category » pour l'aggrégation « First », et « sub_category » pour l'aggrégation « Last ». Décochez les options « Missing » pour éviter de prendre en compte les valeurs manquantes. Cochez « Remove aggregation columns » pour enlever toutes les colonnes « category_Arr[i] » en sortie du composant. Appliquez et éxécutez le composant. Quelle est la catégorie principale et la sous-catégorie du produit « B0B6F7LX4C » ligne « Row16 » ?

    _____

Votre client vous a fourni un nouveau fichier csv « rating_intervals » avec plusieurs intervalles de notes avec leur statuts respectif (Bad, Medium, Good, Top). L'idée est d'attribuer un statut général à chaque produit afin de catégoriser sa qualité.

20. Chargez ce fichier dans un composant « CSV Reader » et exécutez-le.
21. Nous allons utiliser le composant « Value Lookup » pour attribuer le statut de qualité à chaque produit. Ce composant accepte deux entrées. Attachez à la première la chaîne de traitement. Attachez à la seconde le composant « CSV Reader » contenant le fichier « rating_intervals ». Ouvrez le composant. Dans « Lookup column » sélectionnez la colonne « average_rating ». Dans « Key column », sélectionnez « left_rating » du fichier « rating_intervals ». Sélectionnez l'option « Match next smaller ». Vérifiez bien que les colonnes « left_rating », « right_rating » et « rating_status » sont bien inclues pour les avoir en sortie. Ce paramétrage se base sur l'interval de gauche « left_rating » pour trouver le statut du produit. Par exemple, si la note moyenne « average_rating » d'un produit se situe dans l'interval [2.5,3.5[, alors son statut est « Medium ». Exécutez le composant et vérifiez qu'en sortie le statut est attribué pour chaque produit. Quels sont les produits avec un statuts de qualité « Bad » ?

    _____
    _____
    _____
    _____

22. Pour finir, une dernière opération vous est demandée sur la colonne « reviews_id ». Cette dernière contient une liste d'identifiants des commentaires sur le produit. Votre client souhaite que vous calculiez le nombre de commentaires écrits pour chaque produit. Cette opération est assez similaire à la tâche 19 où vous allez utiliser les composants « Cell Splitter » et « Column Aggregator » pour répondre au besoin. Cependant, le délimitateur à utiliser est « , », et l'agrégation à utiliser est l'opération « count » pour compter le nombre de commentaires donnés au produit. Pensez bien à décocher l'option « Missing ». Nommez cette colonne « number_of_reviews ». Quels sont les produits avec le moins de commentaires ?

    _____
    _____
    _____

23. Une fois le jeu de données préparé, vous pouvez créer un nouveau fichier csv contenant les données préparées. Utilisez un composant « CSV Writer ». Spécifiez la localisation dans « File » où vous souhaitez sauvegarder ce nouveau jeu de données. Sélectionnez « overwrite » pour écraser le fichier si besoin. Exécutez le composant, et vérifiez qu'il est bien sauvegardé dans le chemin que vous avez spécifié.

Partie III – Intégration de données pour l'analyse des ventes d'une boulangerie

Pour cette partie, vous devez préparer des données d'une boulangerie coréenne afin de faire des analyses sur les ventes réalisées.

1. Pour commencer, créez un nouveau workflow et nommez le « Korean Bakery ».
2. Chargez le fichier csv « korean_bakery_sales.csv » dans KNIME. Chaque ligne représente une commande représentée par Les colonnes :

- Datetime : Contient la date et l'heure de la livraison
- Day of week : Jour de la semaine où la commande a été réalisée
- Total : Montant total de la commande en Won (Monnaie coréenne)
- Place : Lieu de la commande
- Les colonnes restantes correspondent à la quantité commandée pour un produit spécifique :
    - Viennoiseries : produits qui ne sont pas des boissons (plain bread, croque monsieur…).
    - Boissons
        - Caffe latte
        - Cacao deep
        - Milk tea
        - Lemon ade
        - Vanila latte
        - Berry ade
        - Americano
    - Aliments sucrés
        - Jam
        - Croissant
        - Pandoro
        - Tiramisu
        - Tiramisu croissant
        - pain au chocolat
        - almond croissant
        - gateau chocolat
        - cheese cake
        - orange pound
        - merinque cookies

3. Certaines commandes ne contiennent aucun total avec aucuns produits commandés. Elles sont donc inutiles pour les analyses. Utilisez un nœud « Rule-based Row Filter » pour retirer les commandes avec aucun total. Dans la partie expression, ajoutez « MISSING $total$ => TRUE » et cochez « Exclude TRUE matches » pour écarter les commandes sans total. Après exécution, combien de commandes ont été écartées ?

   _____

4. Les colonnes « croque monsieur » et « mad garlic » sont du type string. Convertissez ces colonnes en type entier.

Pour faciliter des analyses temporelles sur les commandes, vous devez extraire différents attributs temporels.

- Date
- Année

- Mois
- Trimestre
- Jour de la semaine
- Heure de la journée
- Moment de la journée (matin ou après-midi)

5. Dans un premier temps, nous allons extraire la date sur la colonne « datetime ». Utilisez un composant « String Manipulation » pour extraire la date dans une nouvelle colonne nommée « date ». Dans la partie expression, vous pouvez utiliser la fonction « substr(str, start, length) » pour extraire une sous chaîne de caractères.
6. Convertissez la colonne « date » en type « Local Date ». Utilisez le composant « String to Date&Time » et ouvrez-le. Assurez-vous que dans la partie « include », il n'y a que la colonne « date » que vous souhaitez convertir. Dans « Date format », ajoutez bien le format souhaité qui est « yyyy-MM-dd ». Sélectionnez « Date » dans « New type ». Dans « Locale », sélectionnez « en-US ». Exécutez le composant, et vérifiez bien que le nouveau type de la colonne soit « Local Date ».
7. Convertissez aussi la colonne « datetime » au format « yyyy-MM-dd HH:mm ». Cette colonne contient deux informations, la date et le moment de la journée représentée par « HH:mm ». Le type approprié est donc « Date&time ».
8. Un composant pratique sur KNIME pour l'extraction de plusieurs attributs temporels est « Extract Date&Time Fields ». Ajoutez-le et ouvrez-le. Cochez les champs pour obtenir les champs temporels demandés dans la liste précédente. Appliquez et vérifier en sortie que vous avez bien tous les attributs temporels demandés dans le composant. Pour récupérer le dernier attribut temporel, passez à la question suivante.
9. A partir de l'heure, vous pouvez obtenir le moment de la journée où la commande a été faite (matin ou après-midi). Utilisez un composant « Rule Engine » et ouvrez-le. Nommez la nouvelle colonne à ajouter « time of day ». Dans la partie « Expression », écrivez les deux règles suivantes :
    a. Si l'heure de la commande est supérieure ou égale à 12 : La valeur de la colonne est « Afternoon ».
    b. Si l'heure de la commande est inférieure à 12 : la valeur de la colonne est « Morning »

   Appliquez et exécutez le composant.
10. Utilisez un composant « column filter » pour enlever la colonne « datetime » devenue inutile.
11. Pour les colonnes contenant la quantité de chacun des produits, nous souhaitons remplacer les valeurs manquantes par des « 0 ». Ajoutez un composant « Missing Value » à la suite de la chaîne et ouvrez-le. Dans « Number (integer) », sélectionnez que vous souhaitez ajouter un entier fixe qui est la valeur 0. Une fois réalisé, appliquez et exécutez. Vérifier bien que les valeurs manquantes soient remplacées par des 0.

Pour réaliser facilement des analyses sur les commandes, nous avons aussi besoin d'indicateurs numériques. Notamment, on vous demande de calculer les indicateurs suivants :

- Nombre total de produits commandés pour chaque commande.
- Nombre total de boissons commandées pour chaque commande.
- Nombre total d'aliments sucrés pour chaque commande.
- Nombre total de viennoiseries pour chaque commande.
- Nombre total d'aliments salés pour chaque commande.

12. Vous pouvez utiliser des composants « Column Aggregator » ou/et « Math formula » pour calculer ces indicateurs. Quelle est la date de la commande contenant le plus de boissons ?

---

13. Il est possible que certaines commandes ne contiennent aucuns produits commandés. Ecartez ces lignes. Combien de commandes ont été écartées ?

Pour faciliter les filtrages des commandes, on vous demande maintenant de calculer deux nouvelles colonnes booléennes :

- Has drinks : True si la commande contient des boissons, False sinon.
- Has both : True si la commande contient les deux (boissons et non boissons), False sinon.

14. Ajoutez ces deux nouvelles colonnes à l'aide du composant « Rule Engine ».

Pour avoir une meilleure visibilité à l'international, votre boulangerie aimerait traduire le prix des commandes en euros. Pour ce faire, nous avons besoin des taux de changes en Won correspondant à 1 euros, pour chaque date de chaque commande. Les commandes de la boulangerie vont de 2019 à 2020.

Sur Celene, vous avez à disposition le fichier « euro_exchange_rates.csv ». Ce dernier contient le taux de change en Won/Euro pour plusieurs pays pour chaque date de "1999-01-04" à "2023-05-26". Le taux de change est mis à jour du lundi au vendredi. Le week-end ce n'est pas mis à jour.

Vous devrez joindre les deux sources de données en associant le bon taux de change à chaque commande en se basant sur la date de la commande, et de la date du taux de change. Notez que si la commande a eu lieu le week-end, vous devez associer la date de la commande à la dernière date où le taux de change a été enregistré (c'est à dire vendredi).

Une fois joint, il vous suffira de convertir le prix total en Won de la commande en euros.

15. Chargez le fichier csv « euro_exchange_rates.csv ».
16. Filtrez les colonnes contenant la date de change et le prix en Won/Euro à l'aide d'un composant « column filter ».
17. Convertissez en « Local Date » les dates de change, et en type Double les Won/Euro.
18. Pour joindre les données, utilisez un composant « Value Lookup » entre la chaîne de traitement et les données de change en respectant les conditions citées précédemment.
19. Calculer une nouvelle colonne contenant le prix convertis en euros. Combien vaut en euros la commande la plus chère ?

    ---

20. Enfin, sauvegardez le jeu de données préparé dans un nouveau fichier csv.