EXERCICE 1 : ANALYSE DE LA CONSOMMATION DE SUCRE

Objectifs : Traiter et analyser un jeu de données.

1. Créer un projet intitulé « Conso_Sucre ».
2. Importer le fichier « sugar_consumption.csv » présent sous l'emplacement « C:\dataiku\source », et conserver le nom par défaut. Que constatez-vous ?

Ce fichier contient une enquête réalisée sur la quantité de consommation alimentaire de sucre et d'édulcorants (en grammes par personne et par jour).

3. Combien y a-t-il de variables ? Combien y a-t-il d'observations ?
4. Pour commencer à traiter le fichier, créer une analyse visuelle (*Lab*)
5. Renommer la première variable « NA. » en « Pays ».
6. Filtrer la variable « X2004» pour n'en extraire que les valeurs erronées. Quel est le nombre d'observations erronées ?
7. Nettoyer le fichier pour n'en conserver que les valeurs exploitables selon la variable « X2004 », et désélectionner le filtre.
8. Répéter l'opération 6 sur la variable « Pays ». Le résultat obtenu vous parait-il justifié ? Comment pourrions-nous résoudre le problème ? Une fois le problème résolu, désélectionner le filtre de la variable « pays »
9. Nous souhaitons à présent analyser uniquement les années 2000, pour ce faire supprimer toutes les variables pour ne conserver que le nom des pays et les années 2000 (*Columns View*), combien y a-t-il de variables et d'observations ?
10. Au fur et à mesure des opérations réalisées, des actions ont été enregistrées dans le menu « Script » (sur le côté gauche de l'écran). Pour rendre notre travail effectif, nous avons besoins de déployer le script et de l'exécuter.
11. Notre jeu de données étant nettoyé, nous allons analyser ce dernier via les graphiques (*Datasets*  *> Charts*). Construire un histogramme, avec X = nom des pays / Y = année 2000.

12. Trier l'histogramme par ordre décroissant et ne conserver que les 10 premiers pays. Que pouvez-vous dire concernant les pays les plus consommateurs de sucre ?
13. Nous souhaitons analyser l'évolution de la consommation de sucre de ces 10 pays entre l'année 2000 et 2004. Que constatez-vous ?
14. Nous souhaitons à présent avoir une vision globale de l'évolution de la consommation mondiale de sucre entre 2000 et 2004 (*Nuage de points*). *Que pouvez-vous dire ?*
15. Nous souhaitons connaitre la consommation moyenne par année, pour se faire lancer la console R intégrée (*Notebook*), essayer la commande « sapply ». Comparer le résultat avec l'option « analyze » présent dans le menu de chaque « variable ».
16. Nous souhaitons à présent faire une comparaison entre la consommation de sucre et la qualité dentaire. Pour ce faire importer le fichier « badteeth ».
17. Fusionner le fichier « sugar_consumption_prepared » avec « badteeth » via l'interface « Worklow » (*join with ...*). Nous souhaitons fusionner uniquement la colonne « X2004 » du fichier badteeth, en la renommant en « X2004_bad ». Vérifier le schéma sur le Workflow.

L'étude porte sur la mauvaise qualité des dents par pays.
Cette étude est destinée à refléter la prévalence des carries dans une population.
La prévalence est un outil de mesure statistique médicale. Elle renseigne sur le nombre de personnes atteintes par une maladie, elle comptabilise à la fois les nouveaux cas et ceux diagnostiqués plus anciennement à un instant précis, contrairement à la notion d'incidence, qui ne recense que les nouveaux cas sur un intervalle de temps donné.
La variable X2004 contient la moyenne pondérée du nombre de dents infectées, manquantes, plombées, parmi les enfants de 12 ans dans chaque pays.

18. Corriger les erreurs (variables pays & cases vides) et construire le nouveau dataset.
19. Construire l'histogramme avec comme abscisse les pays et l'ordonnée « X2004_bad ». Trier l'histogramme par valeur Ascendante, et ne conservez que les 10 premiers pays Que constatez-vous ?
20. Pouvons-nous dire que la consommation de sucre afflue sur la qualité des dents ?

EXERCICE 2 : PREDICTION DU DIABETE

Objectifs : Mettre en place un modèle prédictif et le déployer sur de nouvelles données.

1. Créer un nouveau projet nommé « Diabete »
2. Importer les fichiers « *diabète 2015* » et « *diabète 2016* ». Intéressons-nous dans un premier temps aux données de 2015. Vérifier que le fichier importé contient bien 563 individus et 9 variables.
3. Nettoyer vos données si nécessaire (recodage, suppression des données manquantes). Mettre la variable « Outcome » au format binaire.

Notre but est de prédire la présence de diabète chez des femmes enceintes.

4. Réaliser un modèle prédictif (*Models → Prediction*) avec comme variable cible « Outcome ». Puis, modifier les paramètres suivants ;
    a. Choisir 70% de l'échantillon pour l'apprentissage et 30% de l'échantillon pour le test du modèle.
    b. Sélectionner uniquement les modèles « Arbre de décision » et « Régression logistique ».

5. Quel est le meilleur modèle ? Quel critère vous a permis de le déterminer ?

REGRESSION LOGISTIQUE

6. Quelles sont les variables significatives ? Afficher les p-values de ces dernières. Quels sont les coefficients de régression ?
7. Quelle est la valeur seuil au-delà de laquelle la prédiction est considérée comme positive (ici présence de diabète) ? D'un point de vue médical, il est préférable de mettre l'accent sur la

sensibilité (*Recall*). De combien est-elle ici ? Si nous souhaiterions gagner en sensibilité, que faudrait-il faire avec la valeur seuil ? *Essayer de jouer cette dernière … Puis revenez à la valeur optimale.*

8. Afficher la courbe ROC. D'après cette dernière, notre modèle est-il bon ? Vous pouvez vous aider du *Reading Tips*.
9. Dans notre échantillon test, combien avons-nous de mauvaises prédictions ? Visualisez-les.

Le but de l'analyse prédictive est de pouvoir appliquer nos modèles à de nouveaux individus dont nous ignorons les valeurs de la variable cible.

10. Deploier le modèle (*Deploy*) et ensuite l'appliquer (*Apply*) aux données de 2016 depuis le workflow.
11. Combien avons-nous prédit de personnes diabétiques ? de personnes non-diabétiques ?

EXERCICE 3 : CLASSIFICATION DES HOTELS

Objectifs : Appliquer un modèle de classification et l'interpréter.

1. Importer le jeu de données « ESIEADMTD5_EX1.csv »
2. Mettre en place une méthode de classification. Nous choisirons d'utiliser la méthode des k-means, avec une partition en 3 classes. Mettre les variables « Etoile » et « Pays » en variables illustratives. (*Settings → Features*)
3. Combien y a-t-il d'individus par classe ?
4. Quelles sont les caractéristiques de chaque classe ? Pour plus de détails, aller dans « *Numérical heatmap* ».
5. A l'aide de « *Cluster Profiles* » regarder la distribution des pays et des étoiles dans les différents clusters. Comment sont-ils répartis dans les classes ?
6. Visualiser les individus de votre classe, en déployant votre modèle et en l'appliquant sur votre jeu de données (*Deploy*). Vérifier que vous retrouvez le même nombre d'individus par classe.

ANNEXE

Rappel pour se connecter :

- Lancer VirtualBox – Démarrer « Dataiku »
- Se connecter via le navigateur (127.0.0.1 :10000) / Login : admin MDP : admin
- **Si problème de connexion**, dans la fenêtre de VirtualBox :
  - Login : dataiku MDP : dataiku
  - Lancer la commande : « *./dss/bin/dss start* »