Cégep de Sainte-Foy
Formation continue et services aux entreprises

Spécialisation technique en intelligence artificielle– AEC (LEA.D1)
420-A55-SF Analyse exploratoire des données

Examen Intra (30%) – Analyse exploratoire des données

Date : 18 octobre 2021, 18h30 à 21h30
Durée : 3 heures
Documentation : Vous avez droit aux notes de cours et aux notes personnelles (avec les exemples de scripts Python).
Évaluation : Individuelle. Vous devez utiliser Python pour résoudre les problèmes.
Remise : Vous devez remettre vos réponses et vos scripts Python sur le site web Brio du cours à la fin de l'examen (travail individuel).

Question 1 (30%) – Indices de qualité et prétraitement des données

Vous pouvez télécharger le jeu de données suivant sur le diabète[1] (**DonneesDiabete_v0r2.csv**), fournissant des données anonymisées sur la santé de patients, à savoir si ces patients vont développer un diabète ou non.

Q1.1 (5%) – Chargez les données, et convertissez la variable catégorique (afin d'avoir des données numériques).

Q1.2 (5%) – Calculez les indices de qualité de complétude et de validité.

Q1.3 (5%) – Effectuez un prétraitement pour enlever les variables ayant un degré de complétude inférieur à 60%. Enlevez les points (instances/lignes) qui ont des observations invalides et/ou qui ont des valeurs manquantes.

Q1.4 (5%) – Calculez la matrice de corrélation. Quelles sont les variables qui (selon les données) semblent avoir le plus d'impact sur le développement du diabète chez un patient? Est-ce que ces coefficients de corrélations semblent faire du sens (au point de vue de leur signe)?

Q1.5 (5%) – Calculez pour chacune des variables (dans la matrice de données prétraitées) les statistiques (métriques) descriptives : moyenne, écart-type, minimum, Q1, Q2, Q3 et maximum.

Q1.6 (5%) – Portez en graphique (dans un histogramme de fréquence) la distribution de la variable « ***Âge patient*** » (dans la matrice de données prétraitées). Est-ce que cette distribution est symétrique? Calculez l'asymétrie ou « skewness » de cette distribution.

---

[1] Adapté/modifié du jeu de données de l'article suivant : *Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press*. Site : https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset?resource=download

Question 2 (30%) – Analyse d'un jeu de données

Vous pouvez télécharger le jeu de données suivant sur la consommation d'eau de différentes villes ou municipalités[2] (**DonneesConsomEau_v0r2.csv**), notamment en fonction de la population.

Q2.1 (5%) – Proposez une représentation graphique permettant de voir rapidement (d'un coup d'œil) les proportions de chaque catégorie de la variable « ***Classe de population desservie*** » dans le jeu de données.

Q2.2 (10%) – Identifiez, parmi les variables catégoriques, la variable ordinale et remplacez les catégories par des valeurs, de 1 à 6 (ordre relatif).

Q2.3 (5%) – Calculez la matrice de corrélation. Quelles sont les trois variables les plus corrélées avec le volume d'eau total distribué pour chaque ville/municipalité (« ***Volume d'eau distribué (ML/an)*** »)?

Q2.4 (5%) – Proposez une représentation graphique permettant de voir la distribution des consommations résidentielles (« ***Consommation résidentielle (L/pers/j)*** »), selon la région administrative (« ***# Région Adm*** »). Est-ce qu'il semble y avoir des différences entre les régions administratives?

Q2.5 (5%) – Produisez un diagramme en nuage de points qui montre la relation entre la population desservie (« ***Population desservie*** ») et le volume total d'eau distribué (« ***Volume d'eau distribué (ML/an)*** »). Est-ce qu'il semble y avoir une tendance? Est-ce qu'il y a un point qui sort du lot?

Question 3 (20%) – Analyse des distributions et analyse Monte Carlo

Vous pouvez télécharger le jeu de données suivant sur le fonctionnement d'un restaurant (**Restaurantv0r2**), donnant des durées typiques pour la préparation d'une commande et les durées typiques du repas (les personnes mangeant à une table).

Q3.1 (5%) – À l'aide d'un histogramme (10 intervalles), montrez la distribution de chaque variable.

Q3.2 (10%) – Effectuez une simulation Monte Carlo pour obtenir la distribution du temps total passé au restaurant, soit pour le modèle (équation) suivant (en assumant que les variables d'entrées sont indépendantes) :

Temps Total (min) = Durée préparation (min) + Durée repas (min)

Q3.3 (5%) – À l'aide d'un histogramme (10 intervalles), montrez la distribution obtenue pour le temps total passé au restaurant. Quelle est la moyenne et l'écart-type pour le temps total passé au restaurant?

[2] Adapté des données du jeu de données ouvertes provenant du Bilan annuel de la Stratégie municipale d'économie d'eau potable complété par les municipalités participantes (2019), https://www.donneesquebec.ca/recherche/dataset/strategie-quebecoise-d-economie-d-eau-potable-2019/resource/2fb02886-d81c-4ca7-be42-2ce8bef3b853 .

Question 4 (20%) – Questions générales

Q4.1 (5%) Vous participez à un projet d'entreprise visant à traiter de larges banques de données afin de faire de l'apprentissage supervisé (pour une future application en IA). Un collègue propose d'intégrer dans le code de prétraitement une ligne de code qui remplacerait automatiquement les valeurs manquantes (dans la matrice de données X):

```
X.replace(nan, 0)
```

Que fait cette ligne de code? Êtes-vous d'accord avec cette façon de procéder?

Q4.2 (5%) Votre collègue Marcus cherche à estimer la consommation moyenne des maisons d'un quartier en électricité (en kWh/j). Les données suivantes ont été compilées dans une matrice de données dans Spyder:

X - DataFrame

ID	Consommation (kWh/j)
1	68
2	82
3	93
4	65
5	73
6	1000
7	93
8	65
9	85
10	-1

stats - DataFrame

Index	Consommation (kWh/j)
count	10
mean	162.3
std	295.567
min	-1
25%	65.75
50%	77.5
75%	91
max	1000

Proposez une méthode qui permettrait de retirer les valeurs aberrantes et/ou extrêmes. Quel sera l'effet sur la moyenne et l'écart-type?

Q4.3 (5%) Un script de traitement de données en Python a les lignes de code suivantes :

```
donnee= pd.read_csv('DonneesTemperatures.csv',index_col="Dates")

X = pd.concat([donnee, donnee.shift(1), donnee.shift(2)], axis=1)

X = X.dropna()
```

En général, ce type de traitement est utilisé pour quel type de données? Que font exactement ces lignes de code?

Q4.4 (5%) Un script de traitement d'image vous est fourni (le fichier[3] **Fleur1.jpg** est fourni). Expliquez succinctement ce que fait chaque ligne de code.

```
import numpy as np
import skimage as sk

image_RGB = sk.io.imread("Fleur1.jpg")
sk.io.imshow(image_RGB)
image = np.uint8(sk.color.rgb2gray(image_RGB)*255)
sk.io.imshow(image)

dimensions=image.shape
Segmentation_mat_bin=np.uint8(np.zeros(dimensions))
Seuil=20
Segmentation_mat_bin[image > Seuil] = 1
sk.io.imshow(Segmentation_mat_bin*255)
```

[3] Fichier tiré du site https://cdn.pixabay.com/photo/2014/02/28/05/36/hibiscus-276496_1280.jpg