30636 E2020 Devoir individuel

(Ce travail compte pour 30% de la note globale et il est à remettre le 20 juin 2020). Devoir à remettre sur le portail Zone cours. Limitations à 7 pages d'explications. Pour le problème 2, veuillez remettre votre fichier .R

Règles d'association

Problème 1 :

Vous travaillez chez une grande entreprise de distribution vestimentaire et d'accessoires nommée He&Me. L'entreprise est située dans la région de Montréal et s'adresse en général à une clientèle de classe moyenne. De plus, elle possède un service en ligne lui permettant de desservir ses produits à l'international. Tous les produits en ligne se retrouvent en magasin. Dans vos données, vous pouvez avoir des noms de villes. Cela fait référence à la provenance de l'acheteur.

Afin de fructifier les ventes de cet été, l'entreprise effectue mensuellement des promotions dans le but d'inciter leurs clients à acheter une certaine gamme de vêtements et d'accessoires. En tant qu'analyste de données pour He&Me, vous êtes responsable de définir les bonnes promotions à pousser chaque mois. Pour cela, vous utilisez des techniques d'exploitation de données.

Vous avez décidé d'effectuer une analyse de règles d'associations sur votre clientèle en prenant tous les achats effectués sur une période d'un an. Après avoir appliqué des critères, vous obtenez les 10 règles d'associations suivantes :

	Antécédent	Conséquent	Support	Confiance	Lift
1	Chaussures style « Les Canadiens »	Montréal	17 %	99 %	2,1
2	Bracelet en Or	Chapeau de style MA	13 %	97 %	1,1
3	Gants hivernaux	Manteau avec doublure	2 %	82 %	3,4
4	Lunettes de soleil	Maillot de bain	1 %	35 %	1,5
5	Pullover style « Ecolo »	Jeans style « Éléphant »	5 %	88 %	6,2
6	T-shirt en V & achat en magasin	Pantalon gris & achat en ligne	11 %	67 %	3,3
7	Veste en velours	Paris & achat en ligne	17 %	87 %	2,3
8	Pantalon style « NEN »	Tokyo & achat en magasin	23 %	43 %	3,1
9	Londres	Parapluie & achat en ligne	27 %	82 %	1.9
10	Paris & achat en ligne	Chemise blanche	22 %	73 %	2,7

a) Veuillez interprétez la règle d'association numéro 3 (support, confiance, lift). D'après vous, comment vous caractériseriez cette règle et est-ce une règle intéressante pour vos futures promotions? (Expliquez votre réponse).

b) Un de vos partenaires marketing se pose beaucoup de questions sur la règle d'association numéro 2. Est-ce une règle intéressante pour la prochaine promotion? (Expliquez votre réponse en quelques lignes).

c) Parmi l'ensemble des règles d'associations calculées, veuillez choisir la meilleure règle pour la prochaine promotion et veuillez décrire en quelques lignes en quoi consisterait cette promotion.

d) Vous possédez un petit échantillon sur lequel vous voulez effectuer des règles d'association. Cet échantillon comporte des transactions uniques indiquant la présence d'un article (oui ou non). De plus, il comporte la colonne Montréal indiquant si la personne habite à Montréal.

À l'aide du tableau ci-dessous :

transaction_id	manteau	chaussures	pantalon	chemises	ceinture	montréal
#v25654	oui	non	non	non	non	non
#v54542	non	oui	oui	oui	non	non
#v57545	non	non	oui	oui	non	oui
#v86745	non	oui	oui	oui	oui	non
#v64542	oui	oui	oui	oui	oui	non
#v85163	oui	non	oui	oui	non	oui
#v54258	oui	non	oui	non	non	non
#v85963	oui	oui	oui	non	non	oui
#v54212	oui	non	oui	oui	non	non
#v54265	oui	non	non	oui	oui	oui
#v79745	oui	oui	oui	oui	oui	oui
#v65342	non	oui	oui	non	non	non

Veuillez déterminer le support, la confiance et le lift des règles d'association suivantes :

i. Chaussures $\rightarrow$ Pantalon
ii. Pantalon $\rightarrow$ Chaussures
iii. Montréal & Manteau $\rightarrow$ Chemises

Problème 2 :

Vous travaillez maintenant en tant que consultant pour une entreprise dans l'industrie du Retail. Les clients de cette entreprise vous ont demandé vos services dans le but de définir une nouvelle stratégie Marketing.

Pour cela, ils ont mis à votre disposition la table de données que vous avez extrait contient les données suivantes :

Online_Retail.csv

InvoiceNo : Le numéro de la facture faisant référence à l'identifiant unique de la transaction
StockCode : Le code du produit lui faisant référence d'identifiant unique
Description : La description du produit
Quantity : La quantité acheté pour un produit en particulier
InvoiceDate : La date à laquelle a eu lieu la transaction
Unitprice : Le prix unitaire du produit
CustomerID : L'identifiant unique de l'acheteur
Country : Le pays de provenance de l'acheteur

Toutes les questions suivantes devront être effectuées à l'aide du langage R

1) Les directeurs vous posent quelques questions afin de mieux comprendre le contexte :
   - **a.** Quel est le produit qui se retrouve le plus au sein des transactions?
   - **b.** Quelle est la provenance engendrant le plus de transactions ?
   - **c.** Quel est le produit le plus rentable au sein des transactions ?
   - **d.** Quel est le client ayant fait le plus de visites ?

2) À l'aide des règles d'association et de l'algorithme Apriori, veuillez déterminer 3 règles que vous pensez intéressantes pour la compagnie. Veuillez donner une interprétation à ces règles (sur toutes les mesures) et justifier en quoi elles sont intéressantes. Enfin veuillez proposer une action avec chacune d'entre elles.

Calcul de distance

Problème 3 :

L'algorithme du K plus proche-voisins (KNN), est un algorithme d'apprentissage supervisé visant à effectuer de la classification ou de la régression (comme les arbres de décision). Le principe repose sur le concept de distance que nous avons abordé lors de la segmentation. En effet, il est question de trouver les voisins les plus proches d'une observation avec les variables explicatives puis d'attribuer la classe la plus prédominante de la variable cible. Le K représente le nombre de voisins d'une observation que nous devons utiliser pour attribuer la prédiction.

Nom	Provenance	Nombre de produit	Age	Revenu	Sexe	Depense	Fidélité
Louisa	Québec	7	18	30000	F	154	eleve
Annie-claire	Québec	3	32	32745	F	678	moyen
Francoise	Québec	8	22	45854	F	1567	eleve
Chantale	Québec	4	18	48983	F	567	moyen
Francois	Montréal	5	36	54748	H	1457	moyen
Jonathan	Montréal	5	33	55759	H	2312	moyen
Jeremy	Toronto	11	27	43234	H	2456	eleve
Axel	Toronto	12	29	4398	F	165	eleve
Kevin	Québec	2	19	67493	F	1282	faible
Thiago	Québec	2	35	89498	H	738	faible
Clara	Montréal	3	39	44356	F	993	faible
Bruno	Toronto	5	42	54789	H	1923	moyen
Pascal	Toronto	1	29	93456	H	2129	eleve
Abou	Montréal	7	39	34678	H	1546	moyen
Mohamed	Toronto	7	33	75355	H	3987	eleve
Michael	Toronto	14	31	45000	H	4277	eleve
Raphaelle	Montréal	12	29	34453	F	2466	eleve
Karine	Montréal	12	18	65789	F	1055	eleve
Sandra	Montréal	5	26	89765	F	745	moyen
Alice	Montréal	4	26	50549	F	1028	?
Steve	Toronto	5	31	47894	H	1283	moyen
Fabien	Toronto	4	33	39776	H	2213	moyen
Ted	Québec	3	27	43465	H	2419	moyen
Victor	Québec	2	40	17975	H	1836	moyen
Julia	Montréal	1	45	44356	F	4277	faible
Bazia	Montréal	1	18	54789	F	2466	faible
Lucas	Toronto	6	21	93456	H	1055	faible
Thomas	Toronto	7	30	34678	H	745	faible

En vous basant sur la table de données présentée ci-haut, veuillez déterminer la valeur de fidélité (la variable cible dans ce problème) d'Alice en analysant quelles sont les 6 femmes (K=6) les plus proches d'elle. Pour calculer, la distance nous vous demandons d'utiliser seulement les variables quantitatives.

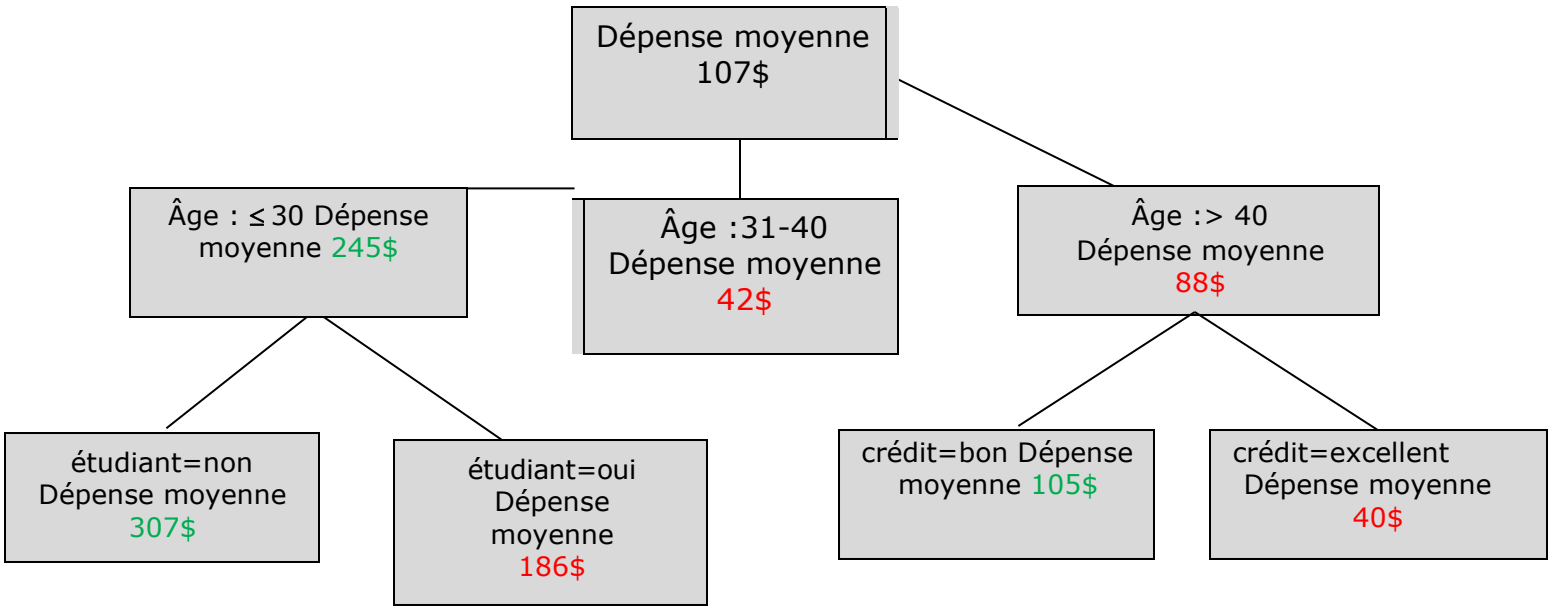
Arbre de régression

Problème 4 :

Description de l'algorithme :

Un arbre de régression est une technique d'apprentissage supervisée ayant exactement le même principe que l'arbre de classification ; cette technique va chercher à créer des règles avec les variables explicatives du jeu de données pour comprendre une variable cible. La grande différence se trouve au niveau du type de la variable à expliquer. En effet, dans le cas d'un arbre de régression, la variable cible est une variable de type continue.

Un exemple d'arbre de régression :



Interprétation : Les personnes ayant moins de 30 ans et étant étudiant vont en moyenne, effectuer des dépenses de 186$.

Les données du problème :

Nom	Provenance	Nombre de produit	Age	Revenu	Sexe	Depense	Fidélité
Louisa	Québec	7	18	89765	F	154	eleve
Annie-claire	Québec	3	32	50549	F	678	moyen
Francoise	Québec	8	22	47894	F	1567	eleve
Chantale	Québec	4	18	39776	F	567	moyen
Francois	Montréal	5	36	43465	H	1457	moyen
Jonathan	Montréal	5	33	17975	H	2312	moyen
Jeremy	Toronto	11	27	44356	H	2456	eleve
Axel	Toronto	12	29	54789	F	165	eleve
Kevin	Québec	2	19	93456	F	1282	faible
Thiago	Québec	2	35	34678	H	738	faible
Clara	Montréal	3	39	44356	F	993	faible
Bruno	Toronto	5	42	54789	H	1923	moyen
Pascal	Toronto	1	29	93456	H	2129	eleve
Abou	Montréal	7	39	34678	H	1546	moyen
Mohamed	Toronto	7	33	75355	H	3987	eleve
Michael	Toronto	14	31	45000	H	4277	eleve
Raphaelle	Montréal	12	29	34453	F	2466	eleve
Karine	Montréal	12	18	65789	F	1055	eleve
Sandra	Montréal	5	26	30000	F	745	moyen
Alice	Montréal	4	26	32745	F	1028	?
Steve	Toronto	5	31	45854	H	1283	moyen
Fabien	Toronto	4	33	48983	H	2213	moyen
Ted	Québec	3	27	54748	H	2419	moyen
Victor	Québec	2	40	55759	H	1836	moyen
Julia	Montréal	1	45	43234	F	4277	faible
Bazia	Montréal	1	18	4398	F	2466	faible
Lucas	Toronto	6	21	67493	H	1055	faible
Thomas	Toronto	7	30	89498	H	745	faible

Le choix de variable dans un arbre de régression :

Dans un arbre de régression, le choix de variable pour un embranchement emploie un concept bien connu des analystes : la variance. En effet, lorsque nous voulons prédire une variable continue, nous cherchons à avoir une estimation (par exemple la moyenne). Sachant cela, il est désirable que chaque estimation des feuilles terminales possède la plus petite variance.

1) En quelques lignes, veuillez expliquer pourquoi nous cherchons à minimiser la variance et donc à ne pas la maximiser.

2) En vous basant sur les données présentées plus-haut, nous vous demandons de trouver la première variable importante afin de prédire la variable « Dépense ». Plus précisément, nous vous demandons de déterminer quelle sera la première variable à utiliser entre les variables « Provenance », « Age » et « Sexe ».

   L'idée est de calculer la moyenne de notre variable à expliquer pour chacune des modalités respectives à leurs variables explicatives.
   Ensuite, nous vous demandons de calculer la variance pour chaque modalité d'une variable et d'additionner ces variances en pondérant avec le nombre d'observation de toutes les modalités de la variable. Vous aurez alors « une somme de variances pondérés par modalité » pour chaque variable explicative et vous devrez choisir celle qui en possède la plus petite.

   Pour créer les modalités de la variable « âge », nous vous demandons de vous baser sur la médiane de la table de données. La médiane sera donc le point de coupure pour cette variable explicative continue.

   Après avoir choisi la première variable importante, veuillez donner une interprétation pour chaque modalité.