HEC MONTRÉAL

EXAMEN FNAL
Programmes de Certificat **ÉTÉ 2020**

Intelligence d'affaires

30-636-00

Enseignant : *Hervé Mensah, chargé de cours* *(S01)*

DIRECTIVES

- L'examen final est un devoir maison **à rendre le samedi 27 Juin avant minuit**
- Toute documentation est permise
- L'examen devra être rendu électroniquement sur la plateforme Zone Cours dans la section 'Remise de travaux'
- Cet examen comporte 4 parties, totalisant 100 points :

    - **Partie 1 :** Arbre de classification
    - **Partie 2 :** Méthode d'ensemble
    - **Partie 3 :** Réseaux de neurones
    - **Partie 4 :** Application sur R

    - ATTENTION : Toute forme de plagiat sera sévèrement sanctionnée pour ce devoir

Contexte

Vous travaillez chez une grande agence de voyages qui se nomme Traveligo. L'entreprise est un fournisseur de voyages, implanté dans la région de Montréal. Par ce fait même, votre clientèle est composée de 70 % de québécois. Nous sommes présentement en hiver et l'organisation possède des objectifs

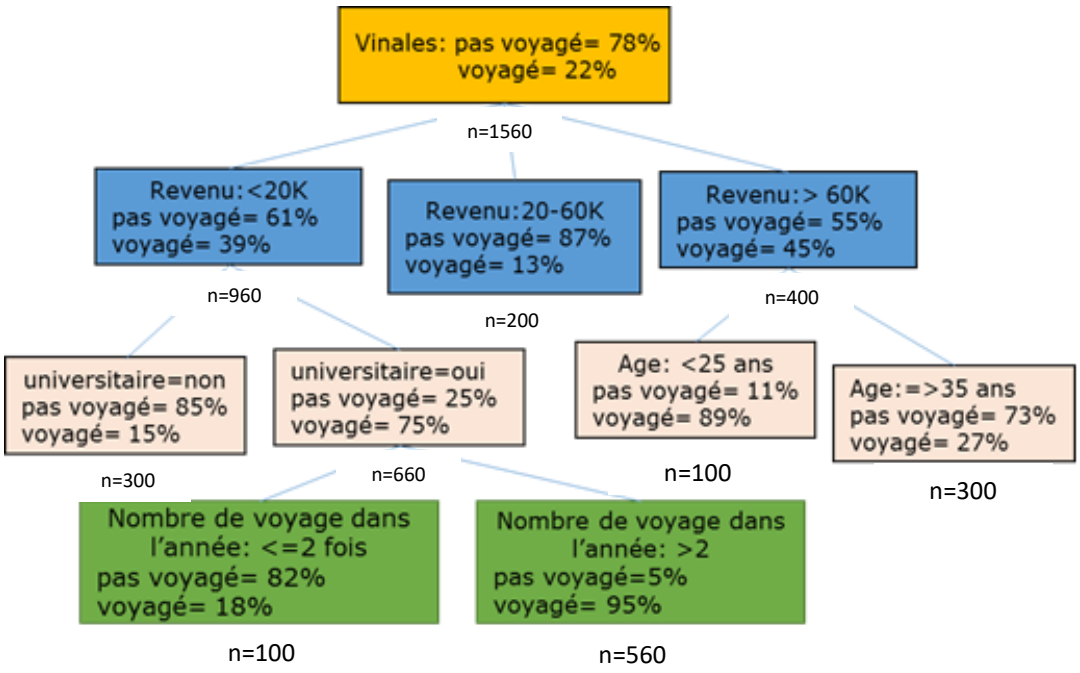
Afin de fructifier les ventes à l'international, l'entreprise effectue mensuellement des promotions dans le but d'inciter leurs clients à acheter un billet d'avion pour une certaine destination. En tant

qu'analyste de données pour Traveligo, vous êtes responsable de définir les bonnes promotions à offrir chaque mois. Pour cela, vous utilisez vos connaissances en techniques de *data mining.*

PARTIE 1: Arbre de classification

Toujours au sein de la compagnie, vous décidez de bâtir un modèle prédictif afin de mieux cibler les personnes qui ont le plus de chance d'aller à Vinales. Vous construisez alors 3 modèles; un modèle d'arbre de classification, un modèle de forêt aléatoire et un modèle de réseaux de neurones.

Pour ce qui est de l'arbre de classification, voici l'arbre final :



a) Pour cet arbre de classification, veuillez décrire les règles du modèle qui conduisent à la prédiction d'un voyage à Vinales. Avec ces résultats, veuillez décrire brièvement en quoi consisterait la campagne pour attirer de nouveaux clients vers cette destination.

b) À l'aide des informations contenues dans la sortie de l'arbre de classification, veuillez déterminer le taux de mauvaise classification. En comparant ce taux de mauvaise classification avec une autre mesure, veuillez-vous prononcer sur l'utilité du modèle pour la compagnie Traveligo.

Voici deux tableaux récapitulatifs des mesures pour les trois modèles :

Apprentissage **(75 000 observations)**	Taux de mauvaise classification	Lift cumulé 4ème décile	ROC	Sensitivité

Réseaux de neurones	15 %	3,2	0.81	51 %
Arbre de classification	18 %	2,1	0.86	63 %
Forêts aléatoires	21 %	2,7	0.80	76 %
Validation **(40 000 observations)**	Taux de mauvaise classification	Lift cumulé 4ème décile	ROC	Sensitivité
Réseaux de neurones	18 %	2.7	0.77	45 %
Arbre de classification	18 %	2,0	0.85	63 %
Forêts aléatoires	21 %	2,7	0.80	75 %

c) Sachant que vous avez une plus grande importance à prédire les personnes qui ont voyagé, quel est le meilleur modèle parmi les 3 modèles (justifiez votre réponse)?

PARTIE 2: Méthodes d'ensemble

Pour cet exercice, nous vous demandons de construire une forêt aléatoire comportant 2 arbres de classification. Chaque arbre aura pour complexité une profondeur maximal de 2 ainsi qu'un indice d'impureté de Gini. Pour la construction des arbres, nous vous demandons de choisir aléatoirement les variables à chaque embranchement (faîte une pile ou face pour savoir quelles seront les variables retenues). Après la construction de la forêt aléatoire, veuillez calculer le taux moyen d'erreur OOB grâce aux fichier EXCEL **data_examen_final_foret_alea** sur Zone-Cours. Enfin, pour Marc-André, un jeune homme montréalais de 26 ayant 6 produits à son actif et des dépenses de 250 $ pour un revenu annuel de 64 000$, veuillez déterminer son niveau de fidélité avec l'organisation. Veuillez expliquer chacune de vos étapes pour la construction et la prédiction de Marc-André. Notes : Pour la construction des arbres avec les variables explicatives quantitatives, nous vous demandons d'utiliser la méthode de point de coupure avec la moyenne. Cela signifie que vous devez faire deux groupes par variable quantitative qui seront 1- en dessous de la moyenne et 2- au-dessus de la moyenne

PARTIE 3 : Application du concept des réseaux de neurones

Toujours étant analyste de données pour la même compagnie, l'organisation vous demande d'être dans les discussions en ce qui concerne les différentes stratégies d'acquisition. En effet, vous avez produit un modèle de réseau de neurones afin de prédire si une personne à une forte propension de devenir un client fidèle ou pas. Vous avez décidé de faire un réseau de neurones comprenant une couche d'entrée, une couche cachée qui elle contient deux neurones puis une couche de sortie.

Nom	Taille	Nombre d'enfants	Nombre d'année d'expérience	Nombre de carte de crédit
Claire	5.7	0	5	1
Marc	5.9	2	4	1
Sébastien	6.2	1	6	3

1) Veuillez représenter graphiquement l'architecture du réseau de neurones.
2) Un de vos collègues de travail vous demandent pourquoi avoir utiliser dans ce modèle une fonction logistique et non linéaire. Veuillez argumenter en quelques lignes sur pourquoi la fonction d'activation logistique et la plus appropriée pour ce genre de problème.
3) Afin de donner une probabilité pour chacune de nos personnes, nous utilisons les poids suivants :

Couche d'entrée vers la couche cachée

	Taille	Nombre d'enfants	Nombre d'année d'expérience	Nombre de carte de crédit
1er neurone	-0.2	0.55	0.2	-0.02
2ème neurone	0.15	0.75	-0.1	0.4

Couche cachée vers la couche de sortie

1^{er} neurone	-0.21
$2^{ème}$ neurone	0.27

De plus, vous pouvez prendre pour acquis que les trois personnes font parties d'un échantillon. En prenant une fonction d'activation logistique et un point de coupure de 50% (0.5), veuillez donner pour Claire, Marc et Sébastien une prédiction si oui ou non la personne deviendra membre de l'organisation. En ce qui concerne l'explication des calculs, veuillez seulement ceux en rapport à la prédiction de Claire.

PARTIE 4 : Mise en pratique avec R : Prédiction du dépôt en direct

Contexte

Vous avez récemment changé d'emploi et vous travaillez maintenant pour l'institution financière HMBC. L'organisation vous a mandaté d'être scientifique de données dans le département analytique et votre premier projet est de déterminer un modèle de classification afin de déterminer les personnes qui s'inscriront à votre nouvelle offre : Le dépôt direct. Pour cela, vous possédez la table de données data devoir final possédant les informations suivantes

Variables dans la table de données **data_examen_final.csv**:

1 - age (numeric)
2 - job : type de travail (categorical: admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")
3 - marital : statut marital (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
4 - education (categorical: "unknown","secondary","primary","tertiary")
5 - default: Y a-til un crédit en défaut? (binary: "yes","no")
6 - balance: Moyenne annuelle de la balance du compte (numeric)
7 - housing: A-t-il une hypothèque? (binary: "yes","no")
8 - loan: A-t-il un prêt personel? (binary: "yes","no")

9 - contact: Type de communication (categorical: "unknown","telephone","cellular")
10 - day: dernier jour de contact dans le mois (numeric)
11 - month: dernier mois de contact dans l'année (categorical: "jan", "feb", "mar", ..., "nov", "dec")
12 - duration: durée de du dernier contact en secodes (numeric)
13 - campaign: nombre de contacts effectués avec le client durant la campagne (numeric, includes last contact)
14 - pdays: nombre de jours qui se sont écoulés après le dernier contact lors de la dernière campagne (numeric, -1 means client was not previously contacted)
15 - previous: Nombre de contacts effectués avant la dernière campagne (numeric)
16 - poutcome: résultat de la dernière campagne (categorical: "unknown","other","failure","success")
17 - y – Le client a-t-il souscrit au dépôt en direct (binary: "yes","no")

À l'aide et **seulement à l'aide** du logiciel R, veuillez répondre aux questions suivantes. Nous vous demandons aussi d'envoyer le code R (en txt ou en word) afin de valider l'ensemble de vos réponses. Les réponses peuvent être à l'intérieur de votre code mises en commentaire ou bien directement sur votre document de réponses.

1) Pouvez-vous déterminer l'âge moyen ainsi que la médiane concernant la balance du compte de notre clientèle?
2) Dans la base de données se trouve la variable y représentant si la personne à souscrit au dépôt direct. Pouvez-vous changer le nom de cette variable en 'deposit'?
3) Pour le reste de l'analyse, nous nous intéresserons seulement aux clients qui possèdent une balance strictement positive. Veuillez créer la table de données data_bank_deposit qui ne possédera que les clients qui ont une balance strictement supérieure à 0.
4) Avant de rentrer dans la prédiction, vous vous demandez à quoi ressemblerait un arbre de classification sur la table de données data_bank_deposit en prenant toutes les variables explicatives. Cet arbre de classification sera un arbre à base de la mesure de Gini, possédant une profondeur maximale de 5 ainsi qu'un nombre minimal de 50 observations dans les feuilles terminales.
   Veuillez effecteur un tel arbre de classification et afficher le graphique de l'arbre en question.
5) Pouvez interpréter la règle menant à la première feuille terminale
6) À partir de cet arbre, pouvez citer les 3 variables explicatives ayant la plus grande importance dans la construction de l'arbre.
7) Vous décidez maintenant de passer à la construction d'un modèle de prédiction. Veuillez déterminer le taux naturel d'erreur que nous possédons dans notre table de données data_bank_deposit.

8) Veuillez construire trois modèle, boosting, bagging, et forêt aléatoire tous de 50 arbres et ayant une profondeur maximale de 4. Pour ce qui est de la forêt aléatoire, veuillez choisir un tirage aléatoire de variables de 5.
9) En comparant le taux de mauvaise classification de tous les modèles, lequel est le plus performant?
10) Vous avez un très fort intérêt à bien prédire les personnes qui vont prendre le dépôt direct car des frais seront associés lors de la prise de contact. Avec cette information, quel est le modèle que vous retenez.
11) Dans ce projet, il y a un concept qui n'a pas été appliqué pour s'assurer que les modèle de contiennent pas de sur-apprentissage. Veuillez expliquer en quoi consisterait ce concept