République Tunisienne

Ministère de l'Enseignement Scientifique et de la Recherche Scientifique

Université de Carthage - Ecole Supérieure de la Statistique et de l'Analyse de l'Information

Rapport de Projet de Fin d'Etudes soumis afin d'obtenir le titre d'

Ingénieur en Statistique et Analyse de l'Information

par

FirstName LastName

Le titre du rapport
Titre

Soutenu le 12/06/2018 devant le Jury composé de :

M. Ben Foulen Foulenia	Président
Mme Ben Foulena Foulen	Rapporteur
M. Ben Foulen Fouleni	Rapporteur
M. Ben Foulen Fouleni	Encadrant
M. Ben Foulen Fouleni	Encadrant

Projet de Fin d'Etudes fait à

(Entreprise d'accueil)

Dédicace

A ... pour son(leur) sacrifice et son(leur) soutien,
en témoignage de mon infinie reconnaissance et mon profond attachement

A tous ceux qui me sont chers...

Remerciements

Je n'aurais jamais pu réaliser ce projet sans la précieuse aide et sans le soutien d'un grand nombre de personnes dont la générosité, la bonne humeur et l'intérêt manifestés à l'égard de mon PFE m'ont permis de progresser.

Ma reconnaissance va à ceux qui ont plus particulièrement assuré le soutien affectif de ce travail : ma famille ainsi que mes amis. Mes parents...

Table des matières

Table des figures

Liste des tableaux

Liste des Algorithmes

Introduction

Voici une référence à l'image de la Figure 1.1 page 10 et une autre vers la partie 2 page 12. On peut citer un livre [Caillois, 1991] et on précise les détails à la fin du rapport dans la partie références. Voici une note [1] de bas de page [2]. Nous pouvons également citer l'Algorithme 1, la Définition 2.1, le Théorème 2.1 ou l'Exemple 2.1...

Le document est déatillé comme suit : le chapitre 1 introduit le cadre général de ce travail. Il s'agit de présenter l'entreprise d'accueil et de détailler la problématique. Le chapitre 2 introduit les données ainsi que les modèles choisies. Le chapitre 3 donne les principaux résultats et la comparaison entre divers modèles (courbe de ROC, indice de Gini). Nous clôturons ce travail par une brève conclusion résumant le travail accompli ainsi que des perspectives qui pourraient enrichir ce travail.

1. Texte de bas de page
2. J'ai bien dit bas de page

Chapitre 1

Données étudiés

tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo

1.1 Cadre du projet et Motivation

And your chapter one goes here [et Nom, 2012a, et Nom, 2012b].
Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod

1.2 Outils utilisés

Il faudrait citer les principaux packages, outils informatiques utilisés lors de ce PFE... Avec les bonnes références bibliographiques.
This is a second subsection[Genette, 1972], [Schaeffer, 1999].
Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam,

1.3 Description des données

consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

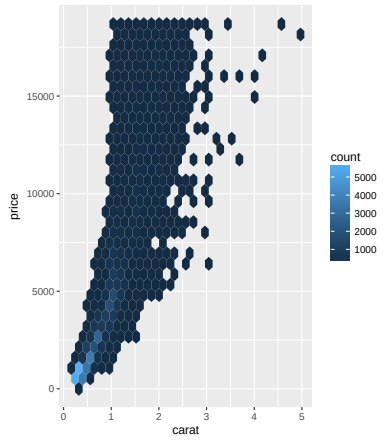

FIGURE 1.1 – Test Image

Entrée	Sortie
A	B
C	D

TABLE 1.1 – Test Table

quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

- **Menu Item**

  Menu Description.

  Focus topics : *Topic one, topic two, topic three, ...*
- **Menu Item**

  Menu Description.

  Focus topics : *Topic one, topic two, topic three, ...*
- **Menu Item**

  Menu Description.

  Focus topics : *Topic one, topic two, topic three, ...*

Also bullets such as :

— One
— Two
— Three

— Four
— ...

1.4 powers series

$$\sum_{i=0}^{\infty} a_i x^i \tag{1.1}$$

The equation 1.1 is a typical power series.

Chapitre 2

Modèles utilisés et Applications

Ici, il s'agit de l'utilisation de TB A contrived acronym (ABC) et Another acronym (EFG) sont des acronyms et des abbréviations... La méthode Support Vector Machines (SVM) est également couramment utilisée.

Exemple 2.1. *On considère le cas particulier...*

— The individual entries are indicated with a black dot, a so-called bullet.
— The text in the entries may be of any length.

Théorème 2.1. *Soit n un entier naturel. Si n est premier alors il n'est divisible que par 1 et par lui-même.*

Démonstration. Here is my proof. □

Definition 2.1. *Soit A une courbe...*

2.1 Modèles retenus

Décrire ici les principales méthods de classification utilisées. Nous donnons l'exemple de CAH dans la section 2.1.1. La section 2.1.2 donne la méthode k-means, etc.

2.1.1 Méthode CAH

Avec cette méthode, nous sommes assurés de l'indépendance du choix initial des centres []. Le choix du nombre de classe se fait aussi à posteriori []. Nous pouvons utiliser la répartition de l'inertie par nombre de classe pour déterminer ce nombre.

Algorithme 1 : Classification Hiérarchique

Input : Write here the input

Output : Write here the output

```
1 while While condition do
2     instructions
3     if condition then
4         instructions1
5         instructions2
6     else
7         instructions3
8     end
9 end
```

Mais attention, car cette méthode est coûteuse en terme de temps de calculs. Elle est aussi limitée et elle ne peut pas traiter un grand nombre d'observations [].

2.1.2 Méthode K-means

Cet algorithme est applicable à des données de grandes tailles []. La réaffectation d'observation d'une classe à une autre au cours de l'itération améliore la qualité des classes. Ceci n'est pas possible avec la classification hiérarchique pour laquelle une affectation est irréversible [].

Algorithme 2 : k-means

Entrée : Write here the input

Sortie : Write here the output

```
1  x ← 0
2  y ← 0
3  foreach ForEach condition do
4      /* comments on code                                          */
5      foreach ForEach condition do
6          if If condition then
7              instruction(s) like below :
8              increase x by 1
9              decrease y by 2
10         end
11         if If condition then
12             instruction
13         else if ElseIf condition then
14             instruction
15         else
16             instruction
17     end
18 end
```

Avec cette méthode, le nombre de classes est fixé d'avance et la partition finale dépend du choix initial de ces derniers [].

2.2 Tests de validation

Filtre de Tuckey (valeurs extrêmes), Régression linéaire, ...

Chapitre 3

Indicateurs de performances et Résultats

Citons ici l'algorithme 1 et l'algorithme 2.Nous proposons ici les indicateurs pour vérifier la qualité du modèle.

3.1 Qalité du modèle

Afin d'évaluer un modèle, nous utilisons les techniques de performances suivantes : la matrice de confusion [], la courbe de gain ROC [], l'Area Under Curve [], la courbe de Lift [], etc.

3.1.1 Matrice de confusion

Il s'agit d'évaluer la qualité du modèle en confrontant les valeurs prédites avec les vraies valeurs prises par la variable à expliquer. Pour cela, nous définissons plusieurs mesures de performances comme l'accuracy, la précision, le rappel, la sensibilité et la spécificité. Ces indicateurs sont définis dans la Table 3.1 où les **Vrais positifs(TP)** sont les individus dont la classe réelle est 1 (Positif) et leur classe prédite est également 1 (Positif). Les **Vrais négatifs (TN)** sont les individus dont la classe réelle est 0 (Négatif) et leur classe prédite est également 0 (Négatif). Les **Faux positifs (FP)** sont les individus dont la classe réelle est 0 (Négatif) et leur classe prédite est 1 (Positif). Dans ce cas, le modèle prédit incorrectement la classe de ces individus. Enfin, les **Faux négatifs (FN)** sont les individus dont la classe réelle est 1 (Positif) et leur classe prédite est 0 (Négatif).

		Classe réelle	
		Negative	Positive
Classe prédite	Negative	Vrai négatif (TN)	Faux négatif (FN)
	Positive	Faux positif (FP)	Vrai positif (TP)

TABLE 3.1 – Matrice de confusion

L'accuracy, notée ACC, est généralement utilisée pour mesurer la performance, qui se traduit par le rapport entre les bonnes prédictions faites par le modèle et le nombre total de prédiction, voir équation (3.1).

$$ACC = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}} \tag{3.1}$$

3.1.2 Courbe de gain Roc et AUC

La courbe ROC (en anglais Receiving Operating Characteristics) est une méthode d'évaluation d'un modèle de prédiction en visualisant son pouvoir séparateur.
Si cette courbe coïncide avec la diagonale alors le modèle st tout aussi performant qu'un modèle aléatoire. Plus la courbe ROC s'approche du coin supérieur gauche, meilleur est le modèle. En effet, cela permet de capturer le plus possible les vrais positifs avec le moins possible de faux positifs.

Dans [**?**] nous trouvons une définition de l'échelle d'interprétation de l'AUC (en anglais Area Under Curve) : c'est l'aire sous la courbe ROC). Lorsque l'AUC est supérieure à 0.9 on parle d'une qualité excellente du modèle. Lorsque l'AUC est comprise entre 0.7 et 0.9, le modèle est de qualité satisfaisante. La qualité du modèle est dite faible lorsque l'AUC est comprise entre 0.5 et 0.7.

3.1.3 Critère de choix de modèle

Ces critères visent à comparer des modèles entre eux en mesurant leurs qualités. Nous pouvons citer l'AIC pour un modèle de paramètre p, appelé aussi « Akaike Information Criterion » est défini dans l'équation (3.2).

$$AIC = . \tag{3.2}$$

L'*AIC* sert à pénaliser en fonction du nombre de paramètre afin d'obtenir un modèle de taille raisonnable.

Le critère de Schwarz *BIC* ou « Bayesian Informative Criterion » est défini par l'équation (3.3).

$$BIC = . \tag{3.3}$$

En présence de plusieurs modèles, ces critères sélectionnent celui qui aura la plus faible valeur.

3.2 Application et Principaux résultats

Conclusion et Perspectives

And a very interesting conclusion here.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Annexe A

Code R pour résoudre la problématique

A.1 Pré-traitement des données

A.2 Code R pour les modèles

An appedix if you need it.

```
Insérer ici le code !
```

A.3 Librairies utilisées

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo.

Bibliographie

[Caillois, 1991] Caillois, R. (1991). *Les jeux et les hommes*. Gallimard, Paris.

[et Nom, 2012a] et Nom, P. (2012a). *Mon livre*. Editeur.

[et Nom, 2012b] et Nom, P. (2012b). *Mon livre*. Editeur.

[Genette, 1972] Genette, G. (1972). *Figure III*. Seuil, Paris.

[Huizinga, 1938] Huizinga, J. (1951 [1938]). *Homo Ludens. Essai sur la fonction sociale du jeu*. Gallimard, Paris.

[Jenkins, 2004] Jenkins, H. (2004). Game design as narrative architecture. In Harrigan, P. and Wardrip-Fruin, N., editors, *First Person : new media as story, performance, and game*. MIT Press, Cambridge.

[Schaeffer, 1999] Schaeffer, J.-M. (1999). *Pourquoi la fiction ?* Seuil, Paris.

Acronymes

ABC A contrived acronym. 10

EFG Another acronym. 10

SVM Support Vector Machines. 10

Index

Entries, 10

Résumé

Insérer un résumé en Français

Mots clés— Ici, Mettre, Cinq, Mots, Clés.