## Partiel Math et R - MSc Data Manangement 20-22 Professeur M. Henri LAUDE

## SERREAU Antoine

## 31/01/2021

Ce document fait l'objet d'une évalutation dans le cadre de mon partiel sur la partie  $Math\'{e}matiques$  pour le Big Data et Programmation R. La réalisation de ce document a été très difficile étant donné la situation actuelle et le travail que l'on doit effectuer pour l'ensemble des matières du semestre 1 + nos projets et nos heures de travaildans le cadre de notre alternance.

## Présentation des "12 travaux d'AstéRix"

Le travail consiste à résumer et évaluer 12 travaux réalisés lors du Premier Semestre de mon MSc Data Management. Ces travaux comportent du code R ou des mathématiques. Tout d'abord, il y a 3 types de travaux : - mathématiques : papiers de recherches, aspects mathématiques, algortihmes - programmation R : différents packages, sujet de programmation - mes travaux personnels

Les 12 travaux sont répartis comme suit : - 5 travaux sur les mahématiques - 5 travaux sur la programmation R - 2 travaux travaux personnels

## Plan du document

- 1. Définition des critères d'évaluations
- 2. Mathématiques :
- Arbre de décision
- DropOut Student
- Algèbre Tropicale
- Prédiction
- Une approche Deep Learning
- 3. Programmation R:
- ggplot2
- leaflet
- dplyr
- Sparklyr
- Shiny
- 4. Evaluation des mes travaux personnels

## 1. Définition des critères d'évaluations

L'évaluation des travaux est libre, nous pouvons choisir les critères que nous souhaitons. J'ai donc décidé de scinder mon évaluation en 2 parties : Structure du document / Contenu

	Libellé du critère	Poids du critère sur 40
Structure du document	Introduction, Sujet, Conclusion	10
	Fluidité de lecture	5
Contenu	Qualité des informations	10
	Connaissances maîtrisées	10
	Références	5

## 2. Mathématiques

#### 2.1. Arbre de décision

-Information concernant ce travail

Titre : Arbre de décision

Lien d'accès au travail présenté : https://github.com/Nicolas-all/PSB1

Auteurs: Rindra LUTZ, Nicolas ALLIX

-Présentation générale du travail

Le travail présente les arbres de décisions. les arbres de décisions modélisent une hiérarchie de test pour prendre une décision ou prédire un résulat en fonction des expériences précédentes. Le but des arbres de décisions est de prédire la variable cible de sortie à partir de plusieurs variables d'entrée connue. Un arbre est constitué de noeuds avec des arrètes qui en découlent jusqu'au noeaud suivant, ainsi de suite et l'arbre se termine lorsque l'ajout de noeuds n'améliore pas la prédiction.

#### -Présentation technique

J'ai décidé de parler plus précisement de la pureté d'un noeud, donc de l'indice de Gini. La pureté d'un noeud se mesure avec l'indice de Gini, plus la valeur de l'indice est proche de 0, plus le noeud est pur.

$$G_i = 1 - \sum_{k=1}^n P_i$$
 ,  $k^2$ 

La variable "pi,k" est la probabilité d'avoir un individu de la classe k parmi la population du ième noeud. Par exemple, calculons la pureté d'un noeud  $\mathbf{a} = [\mathbf{50}, \mathbf{0}, \mathbf{0}]$ 

 $Ga = 1 - (50/50)^2 + (0/50)^2 + (0/50)^2 = 0$  C'est à dire que ce noeud est pur.

#### -Evaluation du travail

	Libellé du critère	Note
Structure du document	Introduction, Sujet, Conclusion	$\overline{\gamma}$
	Fluidité de lecture	4
Contenu	Qualité des informations	7
	Connaissances maîtrisées	7
	Références	2
	Note FINALE	27/40

#### -Conclusion

Ce travail, de mon point de vue, est une énumération de définition. Il y a beaucoup d'informations, c'est très abordable mais j'ai l'impression que le document n'est pas fini. Nous ne savons pas ce sur quoi ce document se base, il n'y a pas de références. Les arbres de décisions sont très intéressants à exploiter, j'espère pouvoir mettre à profit ces connaissances lors des mes projets en entreprise.

#### 2.2. DropOut Student

-Information concernant ce travail

# Titre: Dropout Prediction over Weeks in MOOCs via Interpretable Multi-Layer Representation Learning

Lien d'accès au travail présenté : https://github.com/clairemazzucato/PSBX/tree/main/Mathe%CC% 81matiques%20pour%20le%20Big%20Data

## Auteurs: AUFRERE Thuy, MAZZUCATO Claire, REN Claude

#### -Présentation générale du travail

Ce travail porte sur la prédiction d'abandon d'un étudiant lors de son engagement dans un cours ouverts en ligne (MOOC). Les auteurs du papier de recherche vont s'appuyer sur l'algorithme des branches et des liaisons (BB) en utilisant les données de la plateforme Coursera. Ils s'appuient sur cet algorithme car le clickstream des utilisateurs peut être différent sur l'intervalle et sur la longueur. Ils souhaitent obtenir un modèle non supervisé. A partir de l'algorithme BB, ils vont pouvoir énumérer les solutions du problème et celui-ci va explorer les branches de l'arbre puis faire des comapraisons. Suite à cette étude, les auteurs ont remarqués que le clickstream (flux de clique) impacter le résulat en fonction du niveau de l'étudiant. C'est à dire qu'un étudiant qui ne compte pas abandonner mais qui va rester sur les mêmes sujets toute la semaine va avoir une variance de flux très élevée.

#### -Présentation technique

$$\min_{W_{a,b},b_{a,o}} \frac{1}{2wT} \sum_{t=1}^{T} \sum_{-w \le i \le w, i \ne 0} (\hat{y}_t - x_{t+1})^2$$

L'équation ci-dessus est une double somme sur l'ensemble des semaines et sur un intervalle de 2w autour de la semaine t. La fonction est mise au carré pour avor un résultat prenant en compte des écarts positifs. Le terme "yt" dans la parenthèse au carré représente l'action finale à la semaine t.

#### -Evaluation du travail

	Libellé du critère	Note
Structure du document	Introduction, Sujet, Conclusion	7
	Fluidité de lecture	5
Contenu	Qualité des informations	$\gamma$
	Connaissances maîtrisées	g
	Références	3
	Note FINALE	31/40

#### -Conclusion

Je trouve ce sujet particulièrement intéressant. Pour avoir déjà pratiqué des cours cours ouverts en ligne (MOOC), il est vrai que l'on peut facilement arrêter celui-ci, ce qui a été mon cas... L'utilisation du click-steam peut sembler au premier abord être une bonne solution pour prédire l'abandon d'un élève. Cependant l'étude réalisée montre quela méthodologie sur laquelle ils ont décidés de mettre en place n'a pas eu le succès attendu. Ce travail est maitrisé et l'apprentissage de ce papier de recherche n'a pas l'air d'avoir posé trop de problème.

#### 2.3. Algèbre Tropicale

-Information concernant ce travail

Titre : Algèbre de groupe en caractéristique 1 et distances invariantes sur une groupe fini

Lien d'accès au travail présenté : https://github.com/MarionD436/MATHS

## Auteurs: Marion DANYACH, Imen DERROUICHE, Olfa LAMTI

#### -Présentation générale du travail

Ce travail porte sur l'algèbre tropicale, elle re-définit l'addition et la multiplication ainsi que toutes les opérations qui y sont associées. Deux concepts importants a retenir: L'Algèbre min-plus définie avec le minimum pour addition et l'addition pour multiplication . L'Algèbre max-plus, d"finie avec le maximum pour addition et l'addition pour multiplication. L'algèbre tropicale sert dans la modélisation, l'analyse, l'évaluation de performance et la synthèse de lois de commande pour des classes bien réepertoriées de systèmes à évènements discrets déterministes ou stochastiques (ateliers flexibles, réseaux de transport, réseaux de communications, syst'emes multi-processeur, etc.)

## -Présentation technique

Il n'y a pas d'expression mathématiques utilisées pour répondre à un problème ou non, j'ai donc décidé d'utiliser le guide sur l'algèbre tropicale fournit en Introduction du travail.

Exemple de nouvelles propriétés :

• L'addition

L'addition est donc commutative.

- La multiplication qui est l'addition 'normale'. Comme l'addition normale est commutative, le multiplication tropicale l'est aussi.
- Le parallélisme Nous dirons que deux droites (tropicales) sont parallèles si elles ne coupent pas. Nous allons essayer de trouver s'il existe une ou plusieurs droites parallèles à une droite donnée d et passant par un point A extérieur à cette droite.

#### -Evaluation du travail

	Libellé du critère	Note
Structure du document	Introduction, Sujet, Conclusion	7
	Fluidité de lecture	4
Contenu	Qualité des informations	8
	Connaissances maîtrisées	6
	Références	5
	Note FINALE	30/40

#### -Conclusion

Je ne connaissais pas l'algèbre tropicale. Je trouve cela intéressant d'avoir repensé les mathématiques et de voir l'impact sur la géométrie notamment. une chose a retenir en algèbre tropicale : deux points ne sont pas forcément alignés! Concernant le travail en lui-même. Il aurait été intéressant d'insérer des exemples concrets et d'aller plus loin dans la recherche.

#### 2.4. Prédiction

-Information concernant ce travail

## Titre: What is a Good Prediction - Issues in Evaluating General Value Functions Through Error

Lien d'accès au travail présenté : https://github.com/clairemazzucato/PSBX/tree/main/Mathe%CC% 81matiques%20pour%20le%20Big%20Data

## Auteurs: AUFRERE Thuy, MAZZUCATO Claire, REN Claude

#### -Présentation générale du travail

Ce travail présente les "back stage" de la prédiction afin d'évaluer si une prédiction est bonne ou non. La prédiction est actuellement au centre de l'intelligence artificielle. Ici, nous allons nous pencher l'approche GVF (General Value Functions), c'est une approche de prédiction. Afin d'évaluer une prédiction, il existe une approche, RUPEE (Recent Unsigned Projected Error Estimate). Cependant, cette approche n'est pas efficace d'après les auteurs du papier de recherche. Pour les auteurs, une bonne prévision est celle dont les caractéristiques sont bien alignées avec le problème de prédiction. L'objectif de ce papier est d'améliorer les methodes d'avaluations actuelles notamment en évaluant la prédiction en fonction de sa propre erreur et de l'erreur des autres.

#### -Présentation technique

$$G_t = E_{\Pi}(\sum_{k=0}^{\infty} (\prod_{j=1}^{k} (\gamma_{t+j}) C_{t+k+1})$$

Je concentre mon étude sur l'expression mathématique de GVF. Je n'arrive pas à comprendre plus de choses que les personnes qui ont réalisées ce travail...

#### -Evaluation du travail

	Libellé du critère	Note
Structure du document	Introduction, Sujet, Conclusion	8
	Fluidité de lecture	5
Contenu	Qualité des informations	7
	Connaissances maîtrisées	8
	Références	2
	Note FINALE	30/40

#### -Conclusion

Je trouve ce sujet intéressant et à la fois abstrait car il n'y a pas d'exemple. Le travail est bien présenté et le groupe émet une critique qui correspond à ma pensée. Il manque cependant des références pour pouvoir faire un parallèle avec quelque chose de concret. Après quelques recherche, j'ai trouve une page internet avec différents types d'indicateurs pour valider un modèle de prédiction. https://www.aspexit.com/comment-valider-un-modele-de-prediction/

#### 2.5. Une approche Deep Learning

-Information concernant ce travail

Titre: Une approche Deep Learning pour la classification de modèle BIM

Lien d'accès au travail présenté: https://github.com/Thomas-MAS/PSB1/tree/main/Maths

Auteur: Thomas MASSE

-Présentation générale du travail

Ce travail présente un article autour de l'utilisation du BIM (Building Information Modeling). Un objet BIM est composé de sa géométrie en 3D ainsi que d'attributs, qui seront différentes selon la phase du projet ou l'objectif de la maquette (ex: type de béton, ratio de ferraillage). L'objectif de cet article estde savoir

si nous pouvons utiliser une approche de Depp Learning afin de classifier des modèles BIM. Nous allons parler d'auto-encodeur, c'est un réseau de neurones artificiels exploité dans un contexte d'apprentissage non supervisé.

## -Présentation technique

Dans cette partie je concentre mon attention sur l'expression mathématique dans la cas ou il y aune seule couche cachée.

$$z \in \mathbf{R}^p = F$$
$$z = \sigma(Wx + b)$$

Il y a donc ici une fonction d'activiation ainsi qu'une matrice de poids et un vecteur de biais b. La fonction d'activation est une fonction mathématique appliquée à un signal de sortie.

#### -Evaluation du travail

	Libellé du critère	Note
Structure du document	Introduction, Sujet, Conclusion	8
	Fluidité de lecture	4
Contenu	Qualité des informations	8
	Connaissances maîtrisées	8
	Références	2
	Note FINALE	30/40

#### -Conclusion

Ce travail est plutôt clair dans l'ensemble, il n'y a pas de références malheureusement. Il n'y a pas eu de test réalisé mais l'auteur à l'air d'avoir acquis les connaissances nécessaire et il a bien compris le sujet. La lecture est fluide mais attention à la mise en page.

## 3. Programmation R

#### 3.1. GGPLOT2

-Information concernant ce travail

Titre : GGPLOT 2

Auteurs: Claude REN

-Présentation générale du travail

La librairie ggplot2 est une librarie R du package tidyverse. Ggplo2 permet donc la construction de graphiques complexes de manière efficace avec une syntaxe cohérente et puissante. Ggplot2 part du principe que les données relatives à un graphique sont stockées dans un dataframe (tableau de données).

Il y a de nombreuses possibilités d'utilisation de ggplot2, il est possible de faire des graphiques de différents types. Vous pouvez insérer des boites moustaches par exemple, génrer des fonctions, modifier les couleurs et pleins d'autres paramètres. Ce package est très très vaste.

-Présentation technique

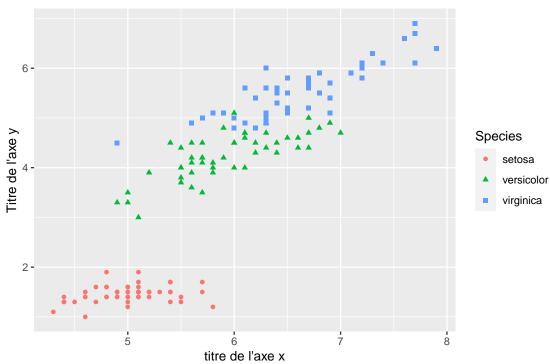
Afin de vous présenter un exemple, je vais utiliser le data frame Iris et générer un graphique.

```
library(ggplot2)
```

## Warning: package 'ggplot2' was built under R version 4.0.3

```
scatter <- ggplot(data=iris, aes(x = Sepal.Length, y = Petal.Length))
scatter + geom_point(aes(color=Species, shape=Species)) +
xlab("titre de l'axe x") + ylab("Titre de l'axe y") +
ggtitle("Ceci est un titre")</pre>
```

## Ceci est un titre



## Explication du code :

data = iris : permet d'extraire les données du dataset Iris aes(x = , y = ) : permet de définir les valeurs attribuées aux axes x et y geom\_point : génère le graphique à points aes(color=Species, : attribuer des couleurs en fonction des espèces shape=Species) : attribuer une forme aux points en fonction des espèces ggtitle : attribuer un titre.

## -Evaluation du travail

	Libellé du critère	Note
Structure du document	Introduction, Sujet, Conclusion	9
	Fluidité de lecture	5
Contenu	Qualité des informations	10
	Connaissances maîtrisées	g
	Références	5
	Note FINALE	38/40

#### -Conclusion

Ce travail est très bien réalisé, il est complet et très agréable à lire. Le package Ggplot2 est à avoir dans ses compétences. Il offre énormément de possibiliés. Allez voir le travail de Claude pour plus d'exemples et de connaissances.

#### 3.2. Leaflet

-Information concernant ce travail

Titre: Forecast Leaflet

Lien d'accès au travail présenté : https://github.com/ArnaudFrsc/PSBX

#### Auteurs: Arnaud FORASACO

#### -Présentation générale du travail

Ce travail présenter le package Leaflet. C'est un pakcage JavaScript open source très populaires pour faire des réaliser des cartes interactives. Leaflet permet d'intégrer et de contrôler facilement les cartes des brochures dans R.

-Présentation technique

```
library(leaflet)
map <- leaflet() %>%
addTiles() %>%
setView(lng= 78.0419, lat=27.1750 ,zoom=15)
addMarkers(map= map,lng= 78.0419, lat=27.1750 ,popup="Taj Mahal")
```

Ici, setView est une méthode permettant de manipuler le widget de la carte. Il est possible d'utiliser colorNumeric afin de définri des couleurs pour la carte.

#### -Evaluation du travail

T 11 117 1 148	
Libellé du critère	Note
Introduction, Sujet, Conclusion	6
Fluidité de lecture	3
Qualité des informations	6
Connaissances maîtrisées	6
Références	2
Note FINALE	23/40
	Introduction, Sujet, Conclusion Fluidité de lecture Qualité des informations Connaissances maîtrisées Références

#### -Conclusion

Le travail sur le pakcage Leaflet manque de commentaires. A première vue, on ne comprend pas très facilement le document, cela manque de fluidité. Le package Leaflet a du potentiel mais il n'est pas mis en application ici.

### 3.3. dplyr

-Information concernant ce travail

Titre: Tidyverse/dplyr package

Lien d'accès au travail présenté : https://github.com/soukaina ElGhaldy/PSB-X/tree/main/R\_packages/dplyr\_package

Auteurs: Soukaina EL GHALDY, Jiayue LIU

-Présentation générale du travail

Ce travail présente le package dplyr faisant parti du package tidyverse. Il s'agit d'une extension facilitant le traitement et la manipulation de données contenues dans une ou plusieurs tables, qu'il s'agisse de data frame ou de tibble (tableaux de données). Elle propose une syntaxe claire et cohérente, sous formes de verbes, pour la plupart des opérations de ce type.

Le package dplyr permet de faire différents des sélections de colonnes, de lignes, afficher des variables, créer des variables et les ajouter dans la base de données initiales etc...

dplyr est un package qui contient une grammaire de manipulation des données qui se fait grâce un nombre réduit de verbes.

#### -Présentation technique

Nous allons concentrer notre étude sur le verbe slice.

#### library(dplyr)

## Warning: package 'dplyr' was built under R version 4.0.3

```
library(tibble)
```

## Warning: package 'tibble' was built under R version 4.0.3

```
data<-data("iris")
data <- as_tibble(data) #Conversion en objet tibble
slice(data, 1:5)</pre>
```

```
## # A tibble: 1 x 1
## value
## <chr>
## 1 iris
```

Ici, nous selectionnons les 5 premières lignes.

-Evaluation du travail

	Libellé du critère	Note
Structure du document	Introduction, Sujet, Conclusion	8
	Fluidité de lecture	4
Contenu	Qualité des informations	8
	Connaissances maîtrisées	8
	Références	4
	Note FINALE	32/40

#### -Conclusion

Ce travail est complet, bien organisé et agréable à lire. Il y a beaucoup d'explications et de références ce qui est un bon point afin de comprendre facilement l'utilisation de ce package. Le package dplyr est très intéressant dans son utilisation, celui-ci fait un bon duo avec ggplot2.

## 3.4. Sparklyr

-Information concernant ce travail

Titre: sparklyr

Lien d'accès au travail présenté : https://github.com/RamyaHTDJ/Psb\_Ramya

Auteurs: Florine COMLAN, Ramya HOUNTONDJI

-Présentation générale du travail

Sparklyr fournit une interface R pour Apache Spark. Sparklyr est un moyen populaire permettant aux développeurs d'utiliser Spark. Ce travail explique comment utiliser sparklyr. Spark intègre des librairies

additionnelles : Streaming, traitement des données en flux, accès aux données Spark avec des requêtes SQL, etc. . .

#### -Présentation technique

```
#install.packages("sparklyr")
#install.packages("devtools")
#library(devtools)
#library(tidyr)
#library(sparklyr)
#spark_install(version = "2.1.0")
#spark_conn<- spark_connect(master="local")</pre>
```

La connexion à spark se fait en utilisant la commande spark\_connect. "Spark\_connect" prend une URL qui donne l'emplacement à Spark. Pour un cluster local (lors de l'exécution), l'URL doit prendre la valeur "local". Pour un cluster distant (sur une autre machine, généralement un serveur hautes performances), la chaîne de connexion sera une URL et un port sur lesquels se connecter.

#### -Evaluation du travail

	T *1 11/ 1 */\	NT /
	Libellé du critère	Note
Structure du document	Introduction, Sujet, Conclusion	8
	Fluidité de lecture	4
Contenu	Qualité des informations	8
	Connaissances maîtrisées	8
	Références	2
	Note FINALE	30/40

#### -Conclusion

Ce travail est complet et compréhensible, cependant, peu de références dans le document. Le package sparklyr utilise la syntaxe de dplyr, ce qui permet dans le cadre de l'utilisation de sparlyr ou dplyr d'être capable de l'utiliser.

#### 3.5. Shiny

-Information concernant ce travail

Titre: Shiny

Lien d'accès au travail présenté: https://github.com/WilliamRbc/PSBX/tree/main/package

Auteurs: Rindra LUTZ, William ROBACHE

-Présentation générale du travail

Shiny est un package R, développé par RStudio, qui permet la création de pages web interactives sur lesquelles il est possible de réaliser toutes les analyses / actions disponibles sous R. Afin de faire fonctionner le package Shiny, il faut créer une interface utilisateur qui servira d'interface à l'utilisateur.

-Présentation technique

```
# Définition de l'interface utilisateur de notre application
ui <- shinyUI(fluidPage(
```

```
# Le titre de votre application
titlePanel("Aperçu d'un dataset"),
#Indiquer le 'layout' de votre application : autrement dit le squelette visuel de l'application
sidebarLayout(
#Composants de la région gauche de lapplication
sidebarPanel(
#Ici, nous insérons un champ pour entrer un chiffre, ainsi qu'un menu déroulant
textInput(inputId = "lignes",
label = "Combien de lignes voulez-vous voir ? ",
value = 10),
selectInput(inputId = "labs",
label = "Dataset",
choice = c("cars", "rock", "beaver1", "sleep")
)
),
#Ici, nous indiquons l'élement qui sera présent dans la fenêtre principale : l'element "dataset", #qui
mainPanel(
tableOutput("dataset")
)
)))
```

L'exemple de code ci-dessus nous montre la création de l'interface utilisateur. Ci-dessous, la partei serveur où l'on met les codes R qui vont s'éxécuter avant l'affichage de l'application.

```
server <- shinyServer(function(input, output) {
    #On retrouve ici l'élement "dataset", qui communique avec ui par 'output'.
    output$dataset <- renderTable({
    #Nous imprimons les éléments d'après les données en entrée : ces derniers sont appellés avec 'input' p
    if(input$labs == "cars") {
        print(head(cars, input$lignes))
    } else if(input$labs == "rock") {
        print(head(rock, input$lignes))
    } else if(input$labs == "sleep") {
        print(head(sleep, input$lignes))
    } else {
        print(head(beaver1, input$lignes))
    }
}</pre>
```

#### -Evaluation du travail

	Libellé du critère	Note
Structure du document	Introduction, Sujet, Conclusion	8
	Fluidité de lecture	4
Contenu	Qualité des informations	8
	Connaissances maîtrisées	$\gamma$
	Références	3
	Note FINALE	30/40

## $\hbox{-} Conclusion$

Le package Shiny semble avoir beaucoup d'utilité. Celui-ci est bien expliqué dans ce travail. Les explications dans les chunks sont d'une grande aide pour comprendre son utilisation.

## 4. Evaluation des mes travaux personnels

#### 4.1. Quantmod

Le Tutoriel sur le package Quantmod été un de mes premières utilisations de RStudio et de Rmarkdown. De ce fait, je trouve que le document aurait pu être mieux organisé. Les points positifs de mon travail sont les exemples utilisées ainsi que les explications sur l'utilisation de la fonction sur R mais aussi la compréhension de ceux ci dans des cas réels. Je pense que mon document peut être utile pour des personnes voulant faire leurs propres anlayses de cours d'actions.

#### Auto-évaluation:

	Libellé du critère	Note
Structure du document	Introduction, Sujet, Conclusion	7
	Fluidité de lecture	3
Contenu	Qualité des informations	8
	Connaissances maîtrisées	8
	Références	5
	Note FINALE	32/40

#### 4.2. KNN

L'article réalisé sur l'algorithme KNN été très intéressant à faire. Je pense avoir bien introduit son utilisation et ses différents aspects bien que j'aurai pu aller plus loin dans le développement des expressions mathématiques. Cependant, j'ai attténué ce manque en ajoutant différentes sources permettant au lecteur de se renseigner plus profondément sur cet algorithme. Mon travail est complet, comprenant toutes les parties demandées pour la réalisation de cet article et il est bien organisé. Afin d'améliorer son format, l'ajout d'un sommaire serait un plus.

#### Auto-évaluation:

	Libellé du critère	Note
Structure du document	Introduction, Sujet, Conclusion	9
	Fluidité de lecture	4
Contenu	Qualité des informations	8
	Connaissances maîtrisées	8
	Références	5
	Note FINALE	34/40