

# **EXPOSE DE FOUILLE DE DONNEES ET RECHERCHE D'INFORMATION,**

## **THÈME : MÉTHODE D'APPRENTISSAGE SUPERVISÉ : PERFORMANCE DES ÉTUDIANTS**

Réalisé par :  
ADOUM OKIM BOKA  
Encadreur : Mme Nguyen THI MINH HUYEN  
ntmhuyen@gmail.com



INSTITUT  
DE LA FRANCOPHONIE  
POUR L'INNOVATION

**August 20, 2019**

# SOMMAIRE



## INTRODUCTION

## PRÉSENTATION DES DONNEES

Présentation de données

## ANALYSE EXPLORATOIRE

Methode Factorielle

Méthode de classification

## CHOIX D'UNE MÉTHODE D'APPRENTISSAGE SUPERVISE

Méthodes

## PRÉPARATION DES DONNÉES

Pré traitement des données

## CONSTRUCTION ET ÉVALUATION DES MODÈLES

Matrice de confusion

Indicateurs de performances

Description de l'arbre

Performances de la solution  $\theta = 0.7$

## COMPARAISON AVEC UNE AUTRE MÉTHODE

## CONCLUSION

L'analyse des données est une famille de méthodes statistiques dont les principales caractéristiques sont d'être multidimensionnelles, descriptives et prédictives.

## Présentation de données

- ☐ Les attributs de données comprennent les notes des élèves;
- ☐ Les caractéristiques démographiques, sociales et scolaires ont été collectés -> rapports et de questionnaires;
- ☐ Performance dans deux matières distinctes (Por - Maths).

**Résumé :** Prédisez les performances des élèves dans l'enseignement secondaire (lycée).

<b>Caractéristiques du jeu de données:</b>	Multivarié	<b>Nombre d'instances:</b>	649	<b>Surface:</b>	Social
<b>Caractéristiques d'attribut:</b>	Entier	<b>Nombre d'attributs:</b>	33	<b>Date du don</b>	2014-11-27
<b>Tâches associées:</b>	Classification, régression	<b>Valeurs manquantes?</b>	N / A	<b>Nombre de visites sur le Web:</b>	425362

Cette section est divisé en deux parties:

- ☐ Réduction du nombre de variables (dimension) : méthodes factorielles;
- ☐ Réduction du nombre d'individus en regroupant par classes : méthodes de classification.

### Methode Factorielle:

Les données sont classifiées en deux types qui sont:

- ☐ les valeurs discrètes dont nous allons appliquer l'analyse de correspondance multiples;
- ☐ les valeurs continues pour l'analyse en composante principal.

## Méthode de classification:

La classification (clustering) est une méthode mathématique d'analyse de données qui permet de faciliter l'étude d'une population d'effectif important (personnes, animaux, plantes, malades, gènes,...), dans l'optique de les regrouper en plusieurs classes de telle sorte que les individus d'une même classe soient le plus semblables possible et que les classes soient le plus distinctes possibles.

## Analyse exploratoire

- ☐ méthodes factorielles;
- ☐ méthodes de classification.

L'attribut cible G3 a une forte corrélation avec les attributs G2 et G1.

- ☐ Méthode de Bayes naïf;
- ☐ K plus proches voisins;
- ☐ Arbre de décision;
- ☐ Réseaux de neurones;
- ☐ algorithme K-NN.





## Méthodes

- ☐ Arbre de décision;
- ☐ CART(Classification And Régression Trees);
- ☐ algorithme K-NN;

Objectif est de prédire la performance des élèves en s'appuyant principalement sur l'attribut G3. **pour ce faire, nous allons transformer les attributs G1, G2, G3 en des valeurs discrètes. Afin quelles soit facilement segmentale**

## Pré-traitement des données

## Pré traitement des données:

FICHIERACCUEILINSERTIONMISE EN PAGEFORMULESDONNÉESRÉVISIONAFFICHAGE

Calibri11A A

</

## Arbre de décision – La méthode CART

Default title

- Dataset (student-por.xls)
  - Discrete select examples 1
    - Define status 1
      - Supervised Learning 1 (C-RT)

Discrete select examples 1

Parameters

Attribute selection : statut  
Value selection : learning

Results

520 selected examples from 649

Computation time : 0 ms.  
Created at 26/06/2019 18:57:50

Composants							
Data visualization	Statistics	Nonparametric statistics	Instance selection	Feature construction	Feature selection	Regression	Factorial analysis
PLS	Clustering	Spv learning	Meta-spv learning	Spv learning assessment	Scoring	Association	
<ul style="list-style-type: none"> <li>Binary logistic regression</li> <li>BVM</li> <li>C4.5</li> <li>C-PLS</li> <li>C-RT</li> </ul>	<ul style="list-style-type: none"> <li>CS-CRT</li> <li>CS-MC4</li> <li>C-SVC</li> <li>CVM</li> <li>Decision List</li> </ul>	<ul style="list-style-type: none"> <li>ID3</li> <li>K-NN</li> <li>Linear discriminant analysis</li> <li>Log-Reg TRIRLS</li> <li>Multilayer perceptron</li> </ul>	<ul style="list-style-type: none"> <li>Multinomial Logistic Regression</li> <li>Naive bayes</li> <li>Naive bayes continuous</li> <li>PLS-DA</li> <li>PLS-LDA</li> </ul>	<ul style="list-style-type: none"> <li>Prototype-NN</li> <li>Radial basis function</li> <li>Rnd Tree</li> <li>Rule Induction</li> <li>SVM</li> </ul>			



### Matrice de confusion:

La matrice de confusion confronte les vraies valeurs et les valeurs prédites de GC sur les 520 observations ayant participé à l'apprentissage (growing + pruning). Elle est accompagnée du taux d'erreur qui est de 0.0577 dans notre exemple.

## Indicateurs de performances:

Trees sequence (# 7)

N°	# Leaves	Err (growing set)	Err (pruning set)	SE (pruning)
7	1	0,1121	0,1163	
6	2	0,0776	0,1105	
5	4	0,0489	0,0756	
4	5	0,0374	0,0756	
3	7	0,0259	0,0814	
2	9	0,0201	0,0872	
1	11	0,0172	0,1047	

## Data partition

Growing set

Pruning set

## Classifier performances

Error rate	0,0577				
Values prediction			Confusion matrix		
Value	Recall	1-Precision		BON	MAUVAIS
BON	0,9544	0,0200	BON	440	21
MAUVAIS	0,8475	0,2958	MAUVAIS	9	50
			Sum	449	71

## Description de l'arbre:

### Tree description

Number of nodes	7
Number of leaves	4

### Decision tree

- GB in [BON] then GC = BON (99,30 % of 286 examples)
- GB in [MAUVAIS]
  - famrel < 4,5000
    - GA in [MAUVAIS] then GC = MAUVAIS (61,54 % of 26 examples)
    - GA in [BON] then GC = BON (77,78 % of 18 examples)
  - famrel >= 4,5000 then GC = MAUVAIS (94,44 % of 18 examples)

---

Computation time : 63 ms.

Created at 26/06/2019 20:11:26

## Performances de la solution $\theta = 0.7$

Nous obtenons un taux d'erreur de 0.1163, avec un intervalle de confiance à 88% égal à [0.1312 ; 0.1398]. En comparant les différents taux d'erreur du TEST par rapport à (0, 0.7 et 1), nous constatons que l'arbre à 4 feuilles suffit largement pour assurer un niveau de performances satisfaisant



## Algorithme K-NN

### Classifier performances *CART*

Error rate			0,0577			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		BON	MAUVAIS	Sum
BON	0,9544	0,0200	BON	440	21	461
MAUVAIS	0,8475	0,2958	MAUVAIS	9	50	59
			Sum	449	71	520

### Classifier performances *K-NN*

Error rate			0,0635			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		BON	MAUVAIS	Sum
BON	0,9913	0,0597	BON	457	4	461
MAUVAIS	0,5085	0,1176	MAUVAIS	29	30	59
			Sum	486	34	520

# CONCLUSION



- Choix de la méthode;
- Stratégie pour obtenir un bon résultat;
- Nos résultats.

Merci pour votre aimable attention!



INSTITUT  
DE LA FRANCOPHONIE  
POUR L'INNOVATION