TP1 Datamining

# Script Générique R

runAnalysis <- function(datasetName, DiscreteAttributesIndeces, AttrsToRemoveIndeces, ClassAttributeIndex) {  
   
 # Chargez les données  
 data <- *read.csv*(datasetName)  
   
 # Supprimez les attributs inutiles  
 data <- data[, -AttrsToRemoveIndeces] # prendre toutes les lignes et toutes les colonnes sauf celles à supprimer  
  
 # Attributs discrets (qualitatifs) et continus (quantitatifs)  
 discrete\_attributes <- data[, *setdiff*(DiscreteAttributesIndeces, ClassAttributeIndex)] # setdiff: différence de deux ensembles (prends toutes les valeurs de DiscreteAttributesIndeces sauf celles de ClassAttributeIndex)  
 continuous\_attributes <- data[, -*c*(DiscreteAttributesIndeces, ClassAttributeIndex)]  
  
 *cat*("Pourquoi séparer les attributs discrets et continus ?  
 Il est important de séparer les attributs discrets et continus,  
 car l'analyse exploratoire et les visualisations appropriées pour ces types de variables sont différentes.  
 Par exemple, pour les attributs continus, on peut utiliser des histogrammes,  
 des boîtes à moustaches ou des nuages de points pour visualiser la distribution des données.  
 Pour les attributs discrets, on utilise généralement des diagrammes en barres ou des tableaux de fréquence.  
 ")  
  
 # Analyse des attributs continus (quantitatifs)  
 for (i in *colnames*(continuous\_attributes)) {  
 *hist*(continuous\_attributes[[i]], main=i, xlab=i)  
 }  
  
 # Analyse des attributs discrets (qualitatifs)  
 for (i in *colnames*(discrete\_attributes)) {  
 *barplot*(*table*(discrete\_attributes[[i]]), main=i, xlab=i)  
 }  
}  
  
*runAnalysis*("data/HR\_prediction-train.csv", *c*(6, 7, 9, 10), 1, 8)

# Analyse préliminaire

### Combien y a-t-il d'instances et de variables (attributs) ?

Il y a 11 variables et 10’000 instances

1. dim(data) # Connaitre le nombre de lignes et de colonnes

### Quelle est la variable cible et est-elle quantitative ou qualitative ?

La variable cible est “left”, elle signifie si la personne à quitter l’entreprise. C’est une variable qualitative (catégorielle). 0 pour non et 1 pour oui.

### Les autres attributs sont-ils quantitatifs ou qualitatifs ?

Les autres attributs sont un mélange de variables quantitatives et qualitatives. Voici le type de chaque variable:

* Quantitatif
  + satisfaction\_level (niveau de satisfaction de l’employé)
  + last\_evaluation (score dans la dernière évaluation)
  + number\_project (le nombre de projets dans lesquels l’employé participe)
  + average\_montly\_hours (la moyenne des heures mensuelles)
  + time\_spend\_company (temps passé avec l'entreprise)
* Qualitatifs
  + Work\_ accident(accident du travail)
  + Left (quitter l’entreprise - variable cible)
  + promotion\_last\_5years (une promotion au cours des 5 dernières années)
  + department (le département)
  + salary (niveau de salaire)

### Certaines variables doivent-elles être exclues de l'analyse

La variable a exclure est l’ID.

### Pourquoi ? Y a-t-il des données manquantes ?

Non

1. # Savoir si il y a des données manquantes

2. any(is.na(myData)) # return False

# Analyse exploratoire

## Pour chaque attribut **qualitatif f** :

* Calculer la distribution de probabilité *P* (**f**) :

P(f) représente la probabilité de chaque valeur d’un attribut

P(f) = (Nombre d'occurrences de la valeur de l'attribut) / (Nombre total d'instances)

# Attributs qualitatifs  
qualitative\_attributes <- *c*("Work\_accident", "promotion\_last\_5years", "department", "salary")  
  
# Calcul de la distribution de probabilité pour chaque attribut qualitatif  
for (attr in qualitative\_attributes) {  
 freq\_table <- *table*(data[[attr]]) # table de fréquence -> nombre d'occurrences de la valeur f dans la variable / nombre total d'occurrences  
 prob\_distribution <- *prop.table*(freq\_table) # Convertir la table de fréquence en distribution de probabilité en divisant chaque valeur par le nombre total d'occurrences  
 *cat*("\nDistribution de probabilité de", attr, ":\n")  
 *print*(prob\_distribution)  
}

**Résultat**

Distribution de probabilité de Work\_accident :

|  |  |
| --- | --- |
| 0 (pas d’accident) | 1 (accident) |
| 0.8541 | 0.1459 |

Distribution de probabilité de promotion\_last\_5years :

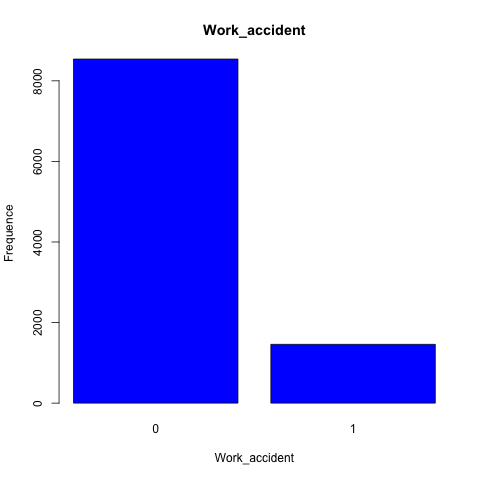
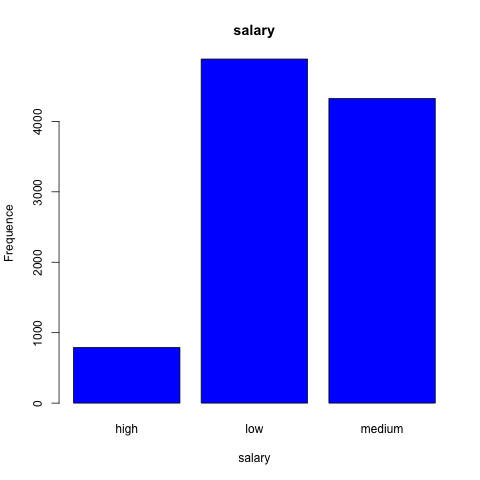
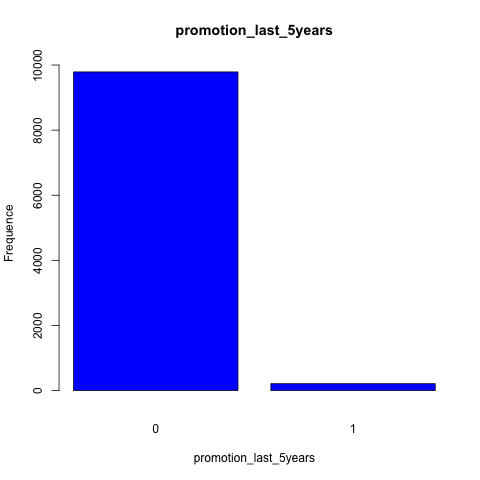
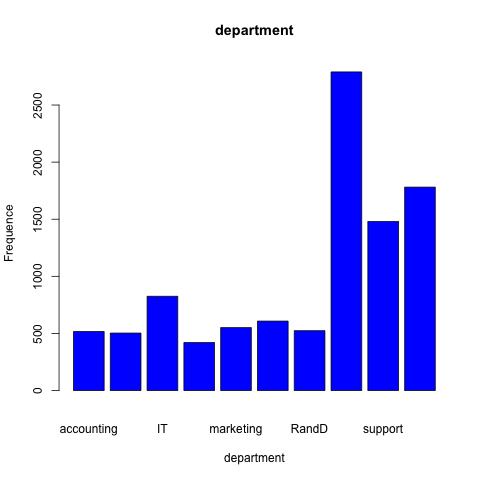
|  |  |
| --- | --- |
| 0 | 1 |
| 0.9789 | 0.0211 |

Distribution de probabilité de department :

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| accounting | hr | IT | management | marketing | product\_mng | RandD | sales | support | technical |
| 0.0517 | 0.0503 | 0.0825 | 0.0420 | 0.0551 | 0.0608 | 0.0524 | 0.2790 | 0.1481 | 0.1781 |

Distribution de probabilité de salary :

|  |  |  |
| --- | --- | --- |
| High | Medium | Low |
| 0.0790 | 0.4325 | 0.4885 |



Les barplots (ou diagrammes à barres) sont utilisés pour représenter les données catégorielles (qualitatives) car ils permettent de visualiser facilement la distribution des fréquences de chaque catégorie. Dans un barplot, chaque catégorie est représentée par une barre et la hauteur de la barre indique la fréquence ou le pourcentage d'occurrences de cette catégorie dans l'ensemble des données.

* Calculer la probabilité conditionnelle de la variable cible, *y*, compte tenu des valeurs d'attribut

*P* (**y|f** )

P(y|f) représente la probabilité de quitter l’entreprise selon f

P(y|f) = (Nombre d'occurrences de la valeur de l'attribut y pour une valeur spécifique de f) / (Nombre d'occurrences de cette valeur de f)

# Pour chaque attribut qualitatif  
for (attr in qualitative\_attributes) {  
 # Crée un tableau croisé des fréquences pour chaque combinaison de valeurs de l'attribut qualitatif et de la variable cible  
 conditional\_table <- *table*(data[[attr]], data[[target\_var]])  
  
 # Convertit le tableau croisé des fréquences en probabilités conditionnelles en divisant chaque fréquence par le nombre d'occurrences de cette valeur de f (en utilisant margin = 1 pour diviser par les totaux des lignes)  
 conditional\_prob <- *prop.table*(conditional\_table, margin = 1)  
  
 # Affiche la probabilité conditionnelle P(y|f) pour chaque combinaison de valeurs de l'attribut qualitatif et de la variable cible  
 *cat*("\nProbabilite conditionnelle P(", target\_var, "|", attr, "):\n")  
 *print*(conditional\_prob)  
}

Probabilite conditionnelle P( left | Work\_accident ):

|  |  |  |
| --- | --- | --- |
|  | 0 (pas quitter) | 1 (quitter) |
| 0 (pas d’accident) | 0.73492565 | 0.26507435 |
| 1 (accident) | 0.91980809 | 0.08019191 |

Probabilite conditionnelle P( left | promotion\_last\_5years ):

|  |  |  |
| --- | --- | --- |
|  | 0 (pas quitter) | 1 (quitter) |
| 0 (pas de promotion) | 0.75799367 | 0.24200633 |
| 1 (promotion) | 0.94312796 | 0.05687204 |

Probabilite conditionnelle P( left | department ):

|  |  |  |
| --- | --- | --- |
|  | 0 (pas quitter) | 1 (quitter) |
| accounting | 0.7272727 | 0.2727273 |
| hr | 0.7335984 | 0.2664016 |
| IT | 0.7793939 | 0.2206061 |
| Management | 0.8452381 | 0.1547619 |
| Marketing | 0.7676951 | 0.2323049 |
| Product\_mng | 0.7976974 | 0.2023026 |
| RandID | 0.8587786 | 0.1412214 |
| Sales | 0.7494624 | 0.2505376 |
| support | 0.7461175 | 0.2538825 |
| technical | 0.7422796 | 0.2577204 |

Probabilite conditionnelle P( left | salary ):

|  |  |  |
| --- | --- | --- |
|  | 0 (pas quitter) | 1 (quitter) |
| High | 0.93924051 | 0.06075949 |
| Low | 0.70030706 | 0.29969294 |
| Medium | 0.79907514 | 0.20092486 |

* Choisissez une variable qualitative, **f ,** et la variable cible *y* et montrez des exemples de la façon dont les règles de probabilité suivantes s'appliquent à elles :

Prenons la variable qualitative f = « département » et la variable cible y = « Left ».

1. Règle de multiplication :

* P(f, y) représente la probabilité conjointe que l'employé appartienne à un certain département et qu'il ait quitté l'entreprise.
* P(f|y) est la probabilité conditionnelle que l'employé appartienne à un certain département, étant donné qu'il a quitté l'entreprise.
* P(y) est la probabilité que l'employé ait quitté l'entreprise.

1. Règle de Bayes :

* P(y|f) représente la probabilité conditionnelle que l’employé ait quitter son entreprise, en sachant qu’il appartient à un certain département
* P(f|y) est la probabilité conditionnelle que l'employé appartienne à un certain département, étant donné qu'il a quitté l'entreprise.
* P(y) est la probabilité que l’employé ait quitter l’entreprise
* P(f) est la probabilité que l’employé appartienne à un certain département

## Pour chaque attribut **quantitatif f** :

* Calculer la moyenne *μ*(**f** ) et la variance *σ*2 **(f** )
* ## Calculer la moyenne μ(f ) et la variance σ2 (f )  
  # Attributs quantitatifs  
  quantitative\_attributes <- *c*("satisfaction\_level", "last\_evaluation", "number\_project", "average\_montly\_hours", "time\_spend\_company")  
  # Calcul de la moyenne et de la variance pour chaque attribut quantitatif  
  for (attr in quantitative\_attributes) {  
   mean <- *mean*(data[[attr]])  
   variance <- *var*(data[[attr]])  
   *cat*("\nMoyenne de", attr, ":", mean)  
   *cat*("\nVariance de", attr, ":", variance)  
  }

Moyenne de satisfaction\_level : 0.613989

Variance de satisfaction\_level : 0.0612014

Moyenne de last\_evaluation : 0.717581

Variance de last\_evaluation : 0.02956431

Moyenne de number\_project : 3.799

Variance de number\_project : 1.515551

Moyenne de average\_montly\_hours : 200.6863

Variance de average\_montly\_hours : 2484.202

Moyenne de time\_spend\_company : 3.4939

Variance de time\_spend\_company : 2.137577

* Calculer la moyenne *μ*(**f** *|target*) et la variance *σ*2 **(f** *|target*) conditionné par la variable cible
* Classez les variables en fonction de leur importance en calculant la différence de moyenne conditionnelle de la classe mise à l'échelle par l'écart-type.