TP1 Datamining

Script Générique R

```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 $P(\mathbf{f})$:

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 :

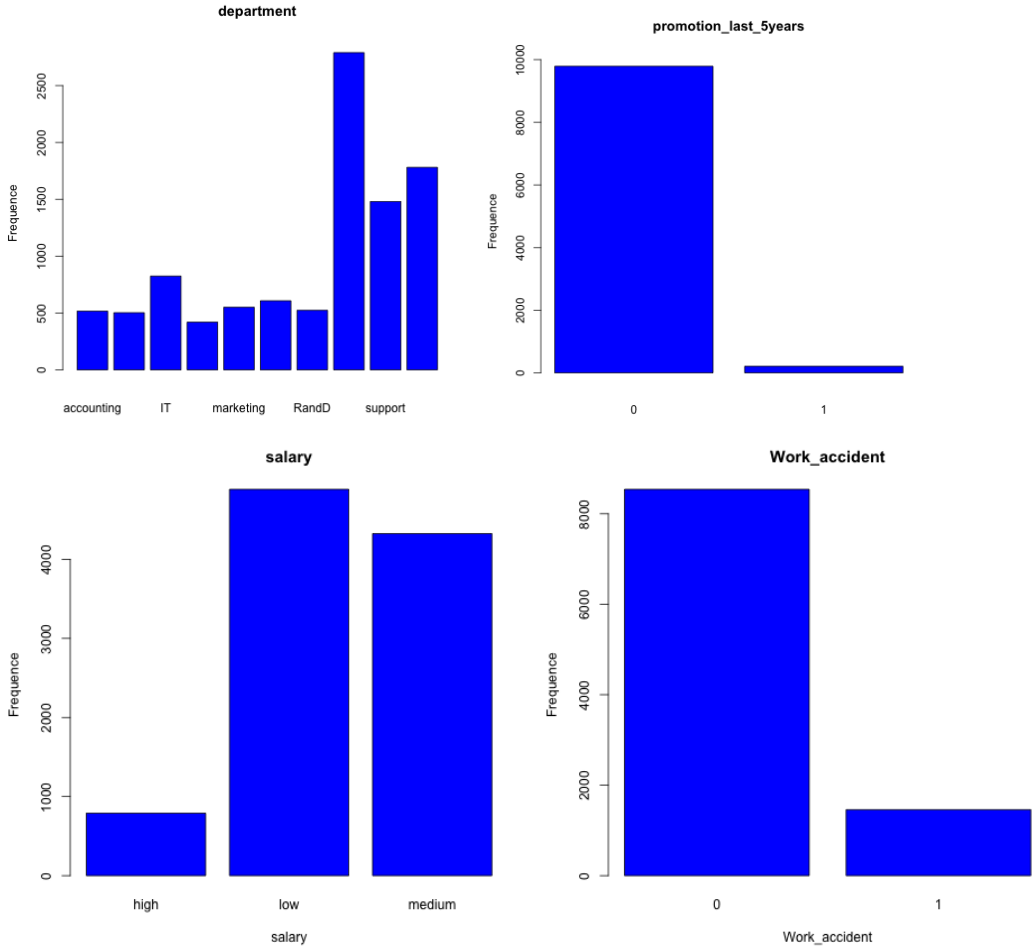
0	1
0.9789	0.0211

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) = (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 :$P(f,y) = P(f|y) * P(y) = P(y|f) * P(f)$
   - 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.
2. Règle de Bayes : $P(y|f) = P(f|y) * \frac{P(y)}{P(f)}$
   - 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 $\mu(\mathbf{f})$ et la variance $\sigma_2(\mathbf{f})$

```r
## 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 $\mu(\mathbf{f}\,|target)$ et la variance $\sigma_2(\mathbf{f}\,|target)$ conditionné par la variable cible

Afin de calculer la moyenne conditionnelle μ(f|target).

1. Crée deux sous-ensembles du tableau avec que les personnes qui ont quitté et qui n'ont pas quitté l'entreprise.
2. Calculer la moyenne et la variance de l'attribut quantitatif f de chacun des sous-ensembles.

```r
# Calcul de la moyenne et de la variance conditionnelles pour chaque attribut quantitatif

quitter <- data[data[[target_var]] == 1,] # extrait les lignes de l'ensemble de données 'data' pour
lesquelles la condition précédente est vraie
pas_quitter <- data[data[[target_var]] == 0,]

for (attr in quantitative_attributes) {
 mean_quitter <- mean(quitter[[attr]])
 mean_pas_quitter <- mean(pas_quitter[[attr]])
 variance_quitter <- var(quitter[[attr]])
 variance_pas_quitter <- var(pas_quitter[[attr]])

 # Afficher les résultats
 cat("\n",attr, ":\n")
 cat("Moyenne conditionnelle de ", attr, " pour les employes qui ont quitte l'entreprise :",
mean_quitter, "\n")
 cat("Moyenne conditionnelle de ", attr, " pour les employes qui sont restes :", mean_quitter, "\n")
 cat("Variance conditionnelle de ", attr, " pour les employes qui ont quitte l'entreprise :",
variance_quitter, "\n")
 cat("Variance conditionnelle de ", attr, " pour les employes qui sont restes :", variance_quitter, "\n")
}
```

satisfaction_level :
Moyenne conditionnelle de satisfaction_level pour les employes qui ont quitte l'entreprise : 0.4434985
Moyenne conditionnelle de satisfaction_level pour les employes qui sont restes : 0.6672687
Variance conditionnelle de satisfaction_level pour les employes qui ont quitte l'entreprise : 0.06960998
Variance conditionnelle de satisfaction_level pour les employes qui sont restes : 0.04665847

last_evaluation :
Moyenne conditionnelle de last_evaluation pour les employes qui ont quitte l'entreprise : 0.718194
Moyenne conditionnelle de last_evaluation pour les employes qui sont restes : 0.7173894
Variance conditionnelle de last_evaluation pour les employes qui ont quitte l'entreprise : 0.039348
Variance conditionnelle de last_evaluation pour les employes qui sont restes : 0.02651144

number_project :
Moyenne conditionnelle de number_project pour les employes qui ont quitte l'entreprise : 3.842923
Moyenne conditionnelle de number_project pour les employes qui sont restes : 3.785274
Variance conditionnelle de number_project pour les employes qui ont quitte l'entreprise : 3.281199
Variance conditionnelle de number_project pour les employes qui sont restes : 0.9633378

average_montly_hours :
Moyenne conditionnelle de average_montly_hours pour les employes qui ont quitte l'entreprise : 206.8274
Moyenne conditionnelle de average_montly_hours pour les employes qui sont restes : 198.7672
Variance conditionnelle de average_montly_hours pour les employes qui ont quitte l'entreprise : 3702.801
Variance conditionnelle de average_montly_hours pour les employes qui sont restes : 2088.346

time_spend_company :
Moyenne conditionnelle de time_spend_company pour les employes qui ont quitte l'entreprise : 3.877782
Moyenne conditionnelle de time_spend_company pour les employes qui sont restes : 3.373934
Variance conditionnelle de time_spend_company pour les employes qui ont quitte l'entreprise : 0.966149
Variance conditionnelle de time_spend_company pour les employes qui sont restes : 2.443379

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.

```r
## Classez les variables en fonction de leur importance en calculant la différence
## de moyenne conditonnelle de la classe mise à l'échelle par l'écart-type
# Crée un vecteur pour stocker les résultats
importance <- c()

# Boucles sur les attributs quantitatifs
for (attr in quantitative_attributes){

  # Calcule la moyenne conditionnelle de la classe pour l'attribut
  mean_left <- mean(quitter[[attr]])
  mean_not_left <- mean(pas_quitter[[attr]])

  # Calcule la différence de moyenne conditionnelle
  mean_diff <- abs(mean_left - mean_not_left) # abs = valeur absolue, comme ca c'est toujours positif

  # Calculer l'écart-type pour chaque classe de variable cible
  sd_left <- sd(quitter[[attr]])
  sd_not_left <- sd(pas_quitter[[attr]])

  # Calculer la moyenne de l'ecart type
  mean_sd <- mean(c(sd_left, sd_not_left))

  # Calcule la différence de moyenne conditionnelle mise à l'échelle par l'écart-type
  scaled_diff <- mean_diff / mean_sd

  # Ajoute le résultat au vecteur d'importance
  importance <- c(importance, scaled_diff)
}

# Crée un dataframe avec les noms des attributs quantitatifs et leur importance
importance_df <- data.frame(Attribute = quantitative_attributes, Importance = importance)

# Trie les attributs par importance décroissante
importance_ranked <- importance_df[order(-importance_df$Importance), ]

# Affiche le classement des attributs par importance
print(importance_ranked)
```

1. Nous avons initialisé un vecteur vide importance pour stocker les valeurs d'importance calculées pour chaque attribut quantitatif.
2. Nous avons parcouru chaque attribut quantitatif avec une boucle for, effectuant les opérations suivantes pour chaque attribut :
    1. Nous avons calculé la moyenne conditionnelle de la classe pour l'attribut, séparément pour les employés ayant quitté et ceux n'ayant pas quitté l'entreprise.

    2. Nous avons calculé la différence absolue entre ces deux moyennes conditionnelles.

    3. Nous avons calculé l'écart-type de l'attribut pour chaque classe de variable cible (quitter et pas quitter).

    4. Nous avons calculé la moyenne des écarts-types pour les deux classes.

5. Nous avons divisé la différence de moyenne conditionnelle par la moyenne des écarts-types pour obtenir la différence de moyenne conditionnelle mise à l'échelle par l'écart-type, qui représente l'importance de l'attribut quantitatif.

6. Nous avons ajouté cette valeur d'importance au vecteur importance.

3. Nous avons créé un dataframe importance_df contenant les noms des attributs quantitatifs et leur importance respective.

4. Nous avons trié ce dataframe par ordre décroissant d'importance pour obtenir le classement des attributs quantitatifs en fonction de leur importance relative pour prédire la variable cible.

Attribute	Importance
satisfaction_level	0.932681105
time_spend_company	0.395787098
average_montly_hours	0.151295977
number_project	0.041282799
last_evaluation	0.004455397

Choisissez une variable quantitative, **f** , et

- Écrivez la distribution normale pour les cas suivants : $P(\mathbf{f})$, $P(\mathbf{f}/y)$, expliquez ce qui change entre les différentes distributions.
- Tracez ces distributions normales, comparez-les aux histogrammes respectifs que nous allons créer ci-dessous, discutez des similitudes et des différences.

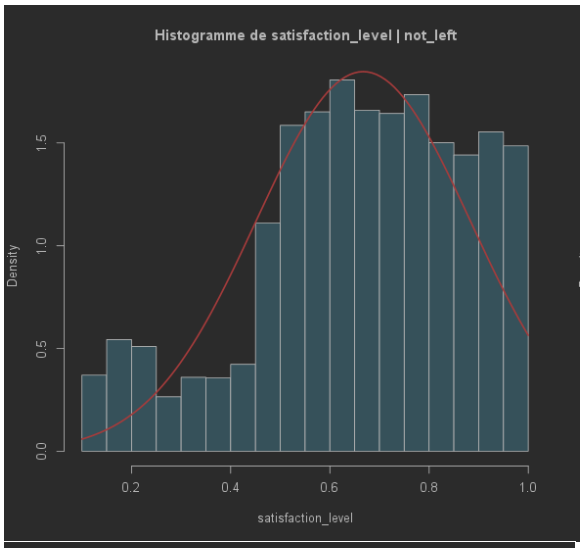
Nous allons choisir la variable quantitative **f** = « satisfaction_level ».

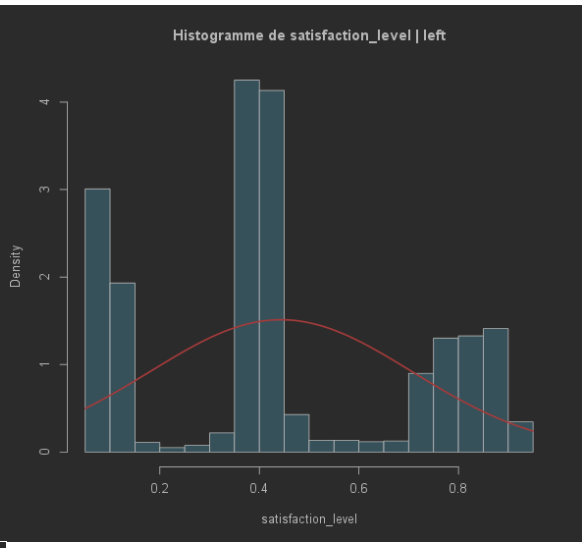
- $P(\mathbf{f})$: Il s'agit de la distribution de la probabilité du niveau de satisfaction des employés (qui sont resté et qui ont quitté l'entreprise)

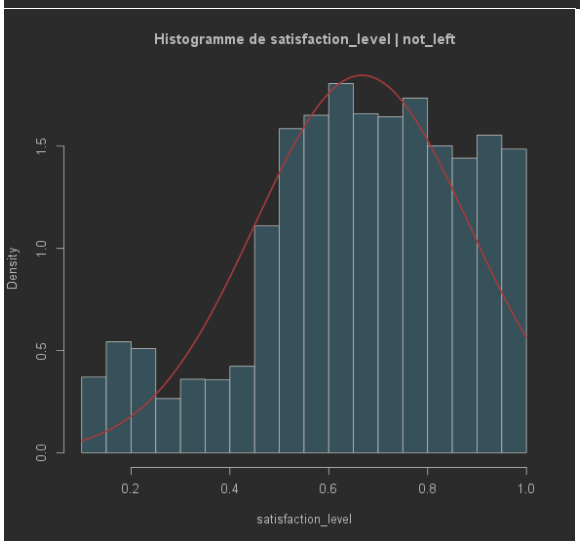
- $P(\mathbf{f}/y)$: Il s'agit de la distribution de la probabilité du niveau de satisfaction des employés, selon la variable cible « left ».

```r
## P (f |y), expliquez ce qui change entre les différentes distributions.
mean_f <- mean(data$satisfaction_level, na.rm = TRUE)
sd_f <- sd(data$satisfaction_level, na.rm = TRUE)

mean_f_left <- mean(quitter$satisfaction_level, na.rm = TRUE)
sd_f_left <- sd(quitter$satisfaction_level, na.rm = TRUE)

mean_f_not_left <- mean(pas_quitter$satisfaction_level, na.rm = TRUE)
sd_f_not_left <- sd(pas_quitter$satisfaction_level, na.rm = TRUE)


# P(f)
hist(data$satisfaction_level, main = "Histogramme de satisfaction_level",
    xlab = "satisfaction_level", col = "lightblue", border = "black", freq = FALSE)

curve(dnorm(x, mean = mean_f, sd = sd_f), add = TRUE, col = "red", lwd = 2)

# P(f|y)
hist(quitter$satisfaction_level, main = "Histogramme de satisfaction_level | left",
    xlab = "satisfaction_level", col = "lightblue", border = "black", freq = FALSE)

curve(dnorm(x, mean = mean_f_left, sd = sd_f_left), add = TRUE, col = "red", lwd = 2)

hist(pas_quitter$satisfaction_level, main = "Histogramme de satisfaction_level | not_left",
    xlab = "satisfaction_level", col = "lightblue", border = "black", freq = FALSE)

curve(dnorm(x, mean = mean_f_not_left, sd = sd_f_not_left), add = TRUE, col = "red", lwd = 2)

# print P(f)
cat("P(f) = N(", mean_f, ", ", sd_f, ")\n")
# print P(f|y=left)
cat("P(f|y=left) = N(", mean_f_left, ", ", sd_f_left, ")\n")
# print P(f|y=not_left)
cat("P(f|y=not_left) = N(", mean_f_not_left, ", ", sd_f_not_left, ")\n")
```

Proportion P(satisfaction_level) : 1
Proportion P(satisfaction_level|left) : 0.2381
Proportion P(satisfaction_level|not_left) : 0.7619


Histogramme de satisfaction_level | not_left
Density
0.0
0.5
1.0
1.5
0.2
0.4
0.6
0.8
1.0
satisfaction_level


Histogramme de satisfaction_level | left
Density
0
1
2
3
4
0.2
0.4
0.6
0.8
satisfaction_level


Histogramme de satisfaction_level | not_left
Density
0.0
0.5
1.0
1.5
0.2
0.4
0.6
0.8
1.0
satisfaction_level

Pour chaque attribut **qualitatif f** : visualiser la distribution de probabilité $P(\mathbf{f})$ et les probabilités conditionnelles $P(\mathbf{y}|\mathbf{f})$ à l'aide de diagrammes à barres

Pour chaque attribut **quantitatif f** : visualiser la distribution $p(\mathbf{f})$ et les distributions conditionnelles $p(\mathbf{f}|cible)$ par des histogrammes.

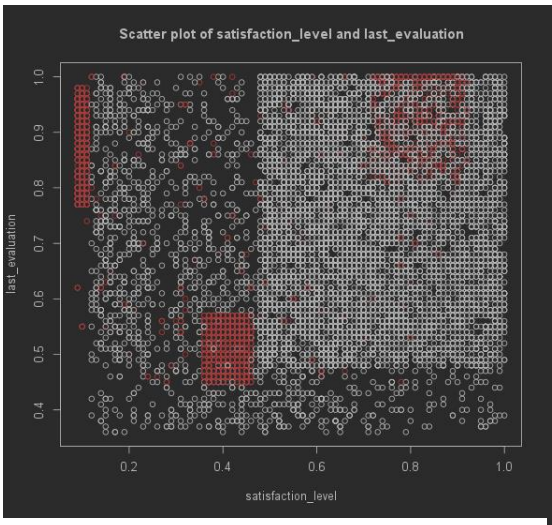
Quels sont les attributs les plus et les moins utiles pour prédire la variable cible ?

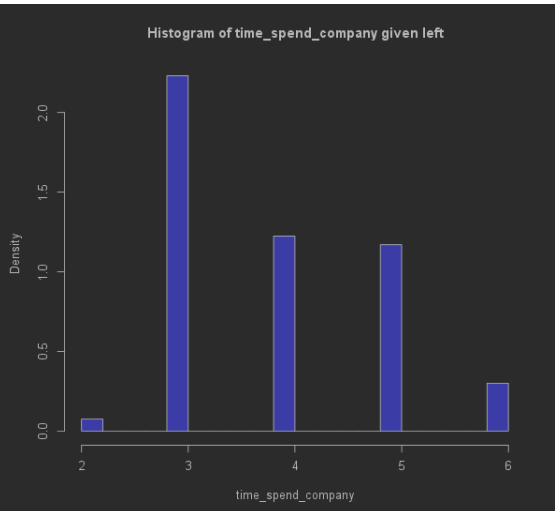
Les attributs les plus utiles sont les attributs quantitatifs, car ils sont plus discriminants que les attributs qualitatifs.

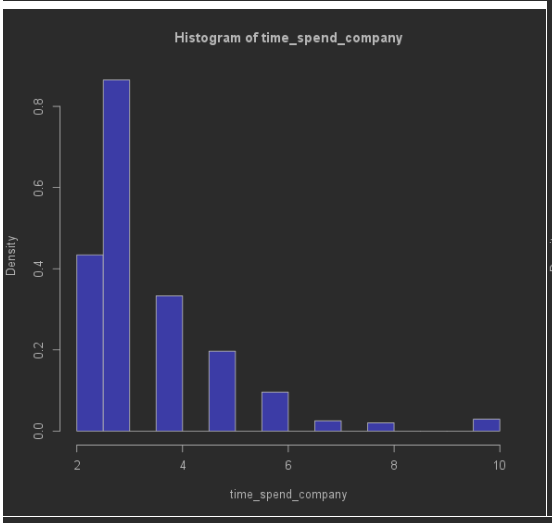
Choisissez deux variables continues que vous jugez les plus utiles et visualisez leur effet sur la cible dans un nuage de points.

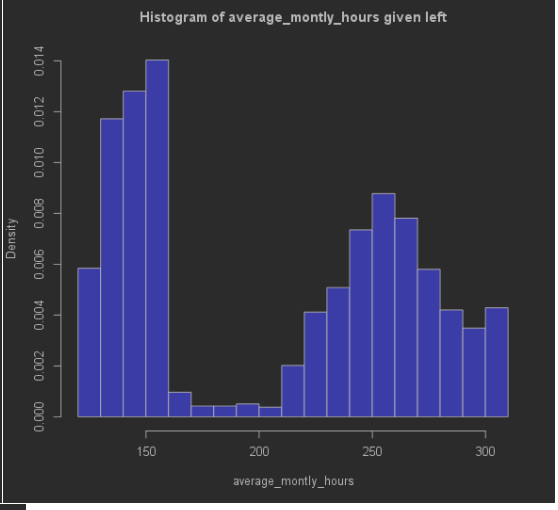
Les attributs les moins utiles sont les attributs qualitatifs, car ils ne sont pas discriminants.

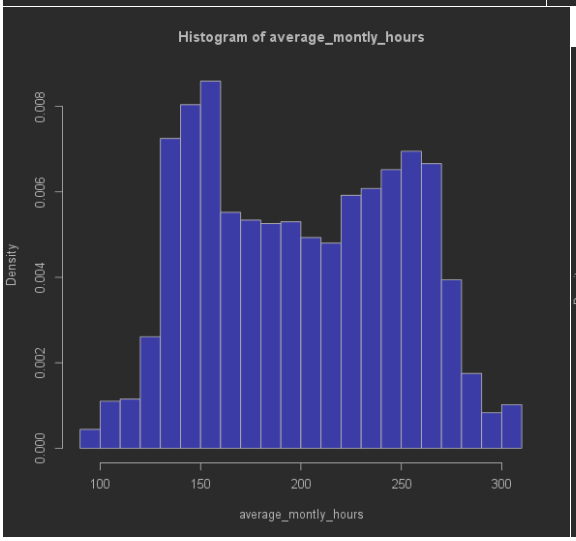
```
for (attr in qualitative_attributes) {
  # Probability distribution
  barplot(table(data[,attr])/nrow(data), main=paste("Probability distribution of", attr), xlab=attr,
ylab="Probability")
  # Conditional probabilities
  barplot(table(data[,attr], data[,target_var])/nrow(data), main=paste("Conditional probabilities of",
target_var, "given", attr), xlab=attr, ylab="Probability")
}

# For each quantitative attribute, visualize the distribution and the conditional distributions by histograms
for (attr in quantitative_attributes) {
  # Distribution
  hist(data[,attr], freq=FALSE, breaks=20, col="blue", main=paste("Histogram of", attr), xlab=attr)
  # Conditional distributions
  hist(data[data[,target_var]==1,attr], freq=FALSE, breaks=20, col="blue", main=paste("Histogram of", attr,
"given", target_var), xlab=attr)
}

# Sélectionner les deux attributs les plus importants
# satisfaction_level et last_evaluation
plot(data[,c("satisfaction_level", "last_evaluation")], col=data[,target_var]+1, main="Scatter plot of
satisfaction_level and last_evaluation", xlab="satisfaction_level", ylab="last_evaluation")
```

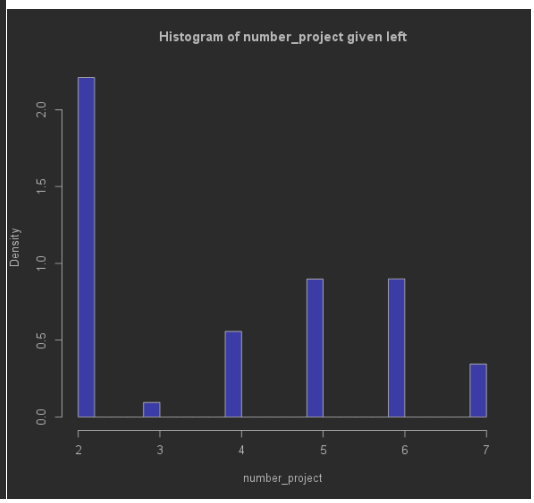

Scatter plot of satisfaction_level and last_evaluation
last_evaluation
satisfaction_level
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.2
0.4
0.6
0.8
1.0

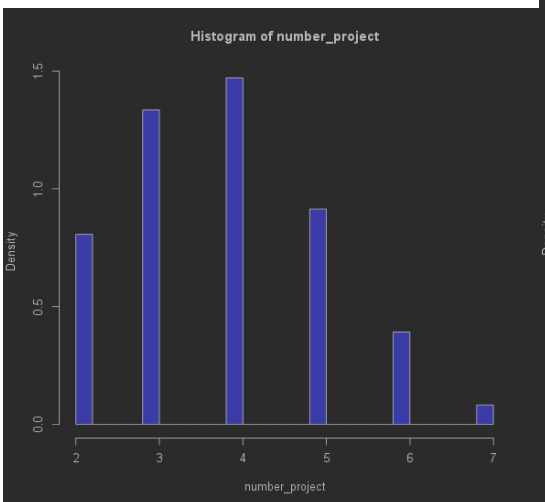

Histogram of time_spend_company given left
Density
time_spend_company
0.0
0.5
1.0
1.5
2.0
2
3
4
5
6

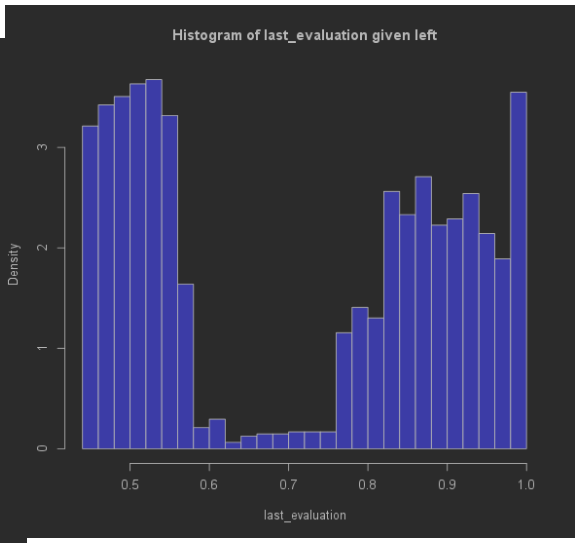

Histogram of time_spend_company
Density
time_spend_company
0.0
0.2
0.4
0.6
0.8
2
4
6
8
10

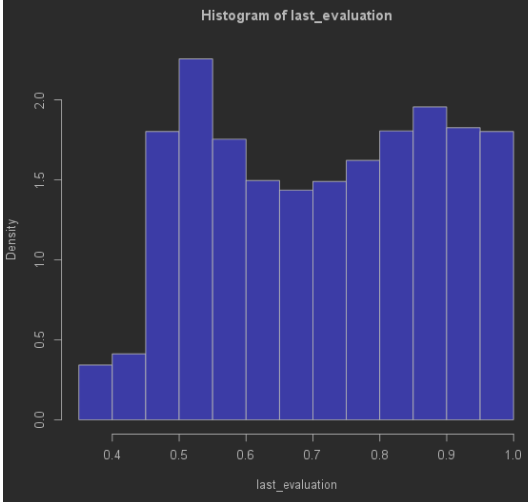

Histogram of average_montly_hours given left
Density
average_montly_hours
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
150
200
250
300

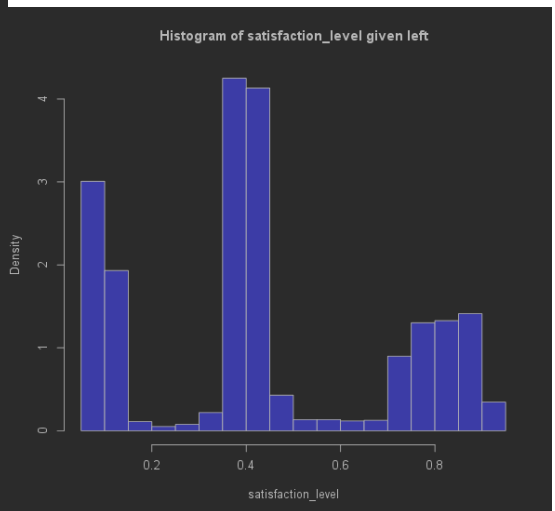

Histogram of average_montly_hours
Density
average_montly_hours
0.000
0.002
0.004
0.006
0.008
100
150
200
250
300

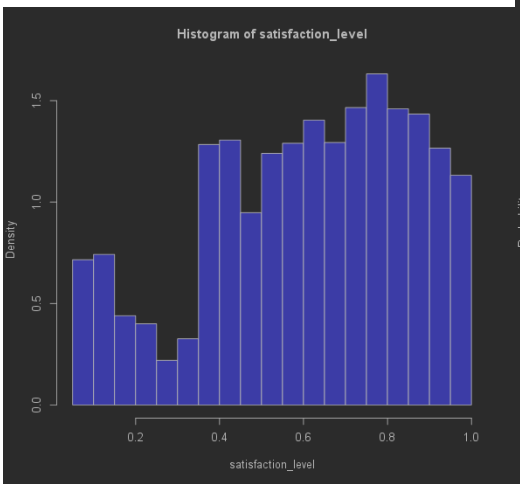

Histogram of number_project given left
Density
number_project
0.0
0.5
1.0
1.5
2.0
2
3
4
5
6
7

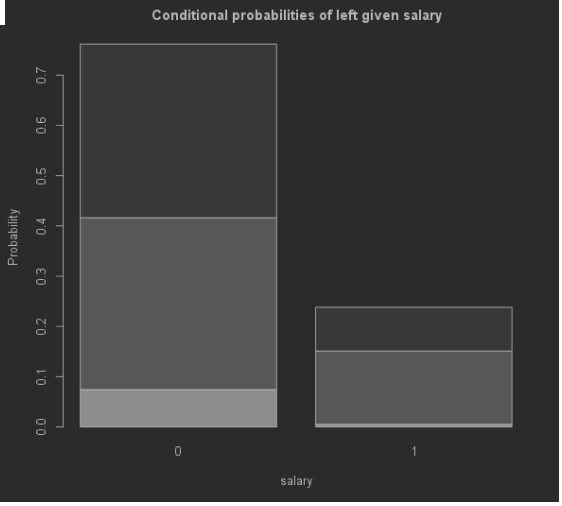

Histogram of number_project
Density
number_project

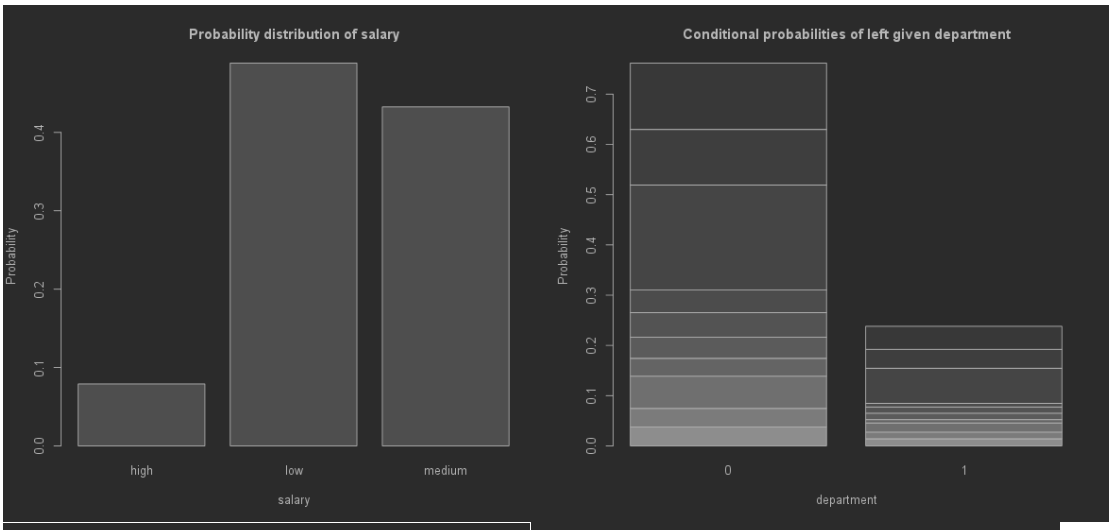

Histogram of last_evaluation given left
Density
last_evaluation


Histogram of last_evaluation
Density
last_evaluation

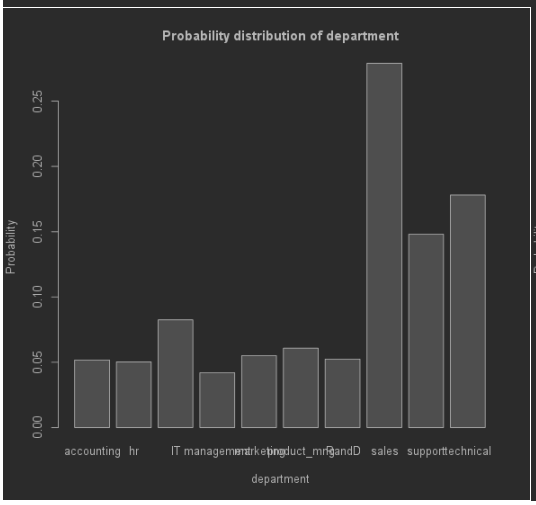

Histogram of satisfaction_level given left
Density
satisfaction_level

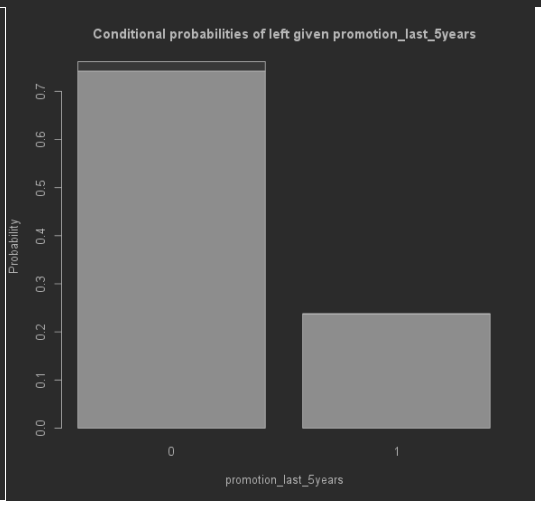

Histogram of satisfaction_level
Density
satisfaction_level

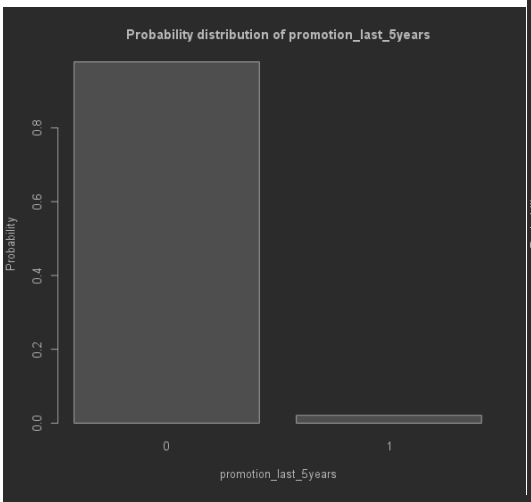

Conditional probabilities of left given salary
Probability
0
1
salary

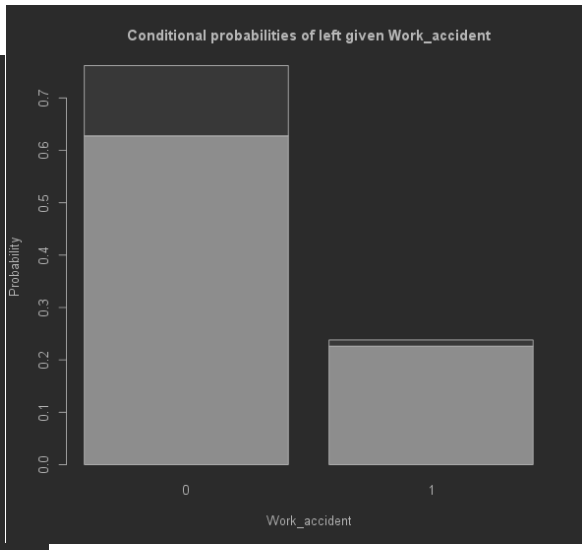

Probability distribution of salary
Probability
0.0
0.1
0.2
0.3
0.4
high
low
medium
salary

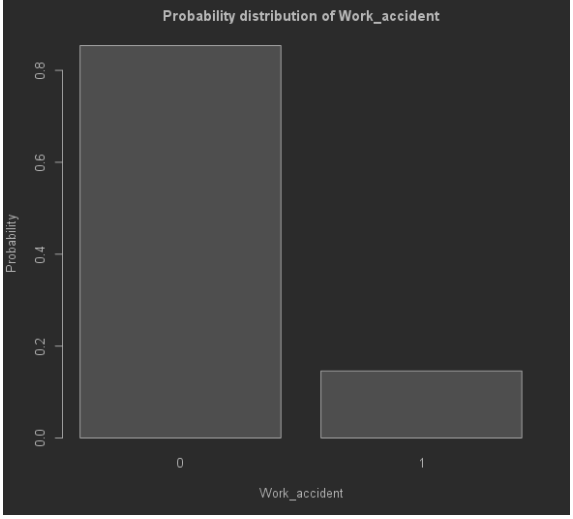
Conditional probabilities of left given department
Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
1
department


Probability distribution of department
Probability
0.00
0.05
0.10
0.15
0.20
0.25
accounting
hr
IT
management
marketing
product_mng
R&D
sales
support
technical
department


Conditional probabilities of left given promotion_last_5years
Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
1
promotion_last_5years


Probability distribution of promotion_last_5years
Probability
0.0
0.2
0.4
0.6
0.8
0
1
promotion_last_5years


Conditional probabilities of left given Work_accident
Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
1
Work_accident


Probability distribution of Work_accident
Probability
0.0
0.2
0.4
0.6
0.8
0
1
Work_accident