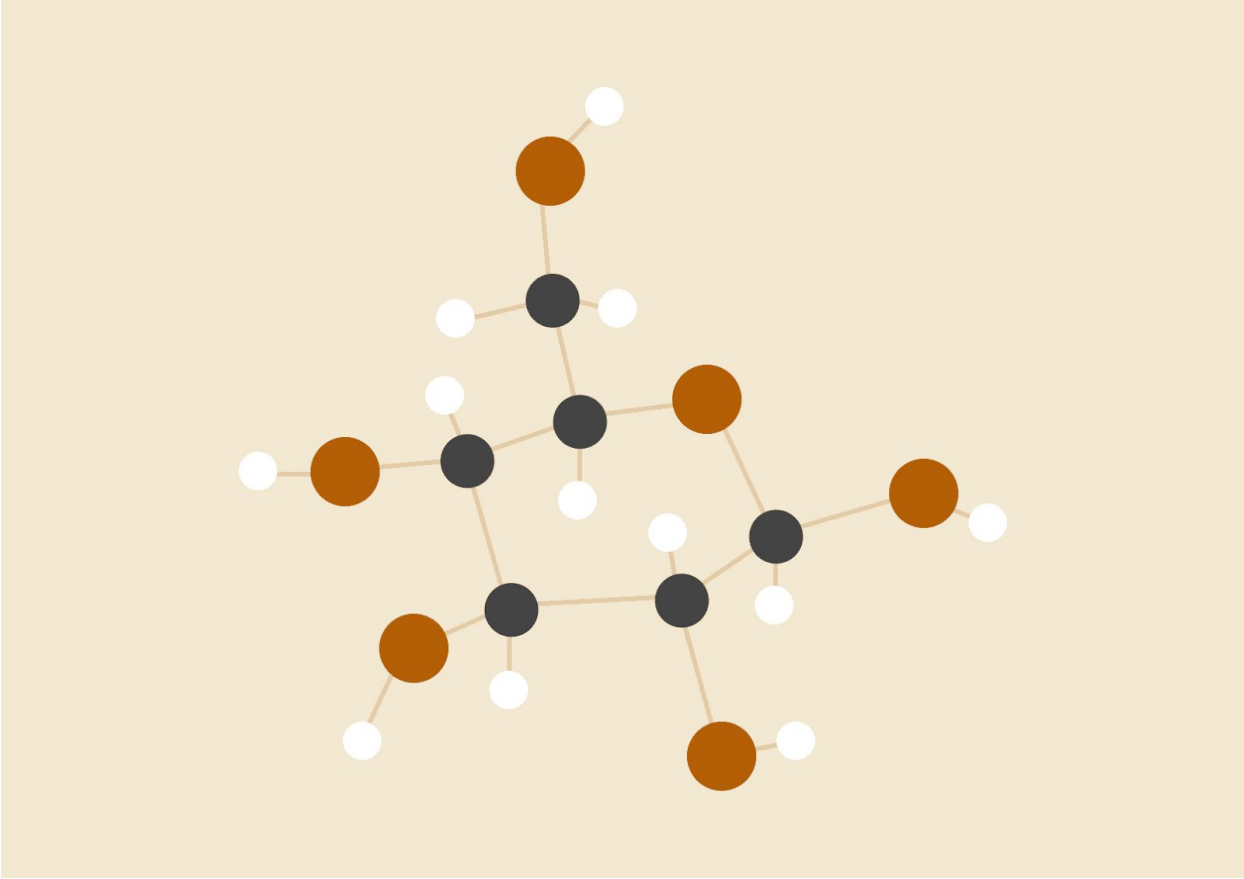
# Data mining

*TP 2*



**Michael Quinto, Jeremy Malera**

625-2

Semestre d'automne 2020

# INTRODUCTION

La banque pour laquelle on travaille, nous a demandé de créer un modèle qui nous servira pour prédire le produit d'investissement qu'un client est plus susceptible de choisir selon son comportement. Ceci est utile pour ne proposer seulement le meilleur produit au client, au lieu de toutes les options possibles. Son comportement est définit par des variables et il faut donc qu'on détermine les plus pertinentes.

Dans le premier TP, nous avons dû déterminer à l'oeil nu ces variables et selon notre étude, celles-ci étaient :

- SE2 : l'âge du client
- IA2 : qui représente le nombre total d'opérations pour ce type d'activité d'investissement

Dorénavant, nous allons utiliser des outils qui vont nous permettre de mieux choisir ces variables et également de les classer dans l'ordre du plus au moins important. Comme outils, nous avons :

- L'information gain
- L'information gain ratio
- Le Gini gain

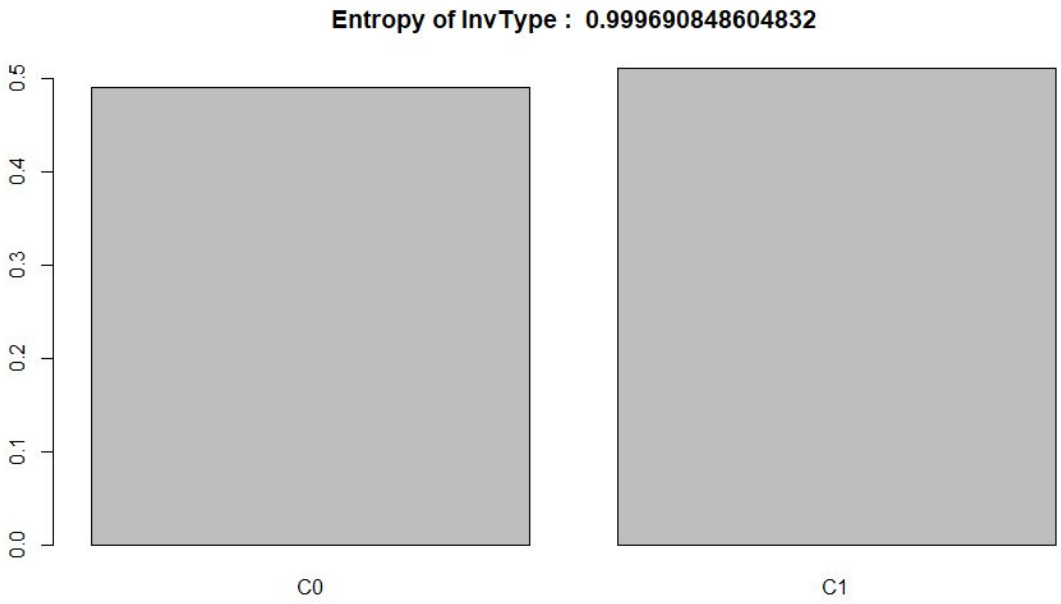
Ces outils vont donc nous aider pour sélectionner les variables selon un critère quantitatif. Cette sélection s'appelle "feature selection".

# VARIABLE CIBLE

## InvType

Notre variable cible est InvType et prend les valeurs C0 et C1. Nous devons étudier comment les autres variables influencent le choix des clients entre les deux produits d'investissement grâce aux outils mentionnés précédemment.

Voici le barplot représentant la distribution de invType, avec la valeur de son entropie:




L'entropie d'une variable permet de mesurer l'incertitude de celle-ci. La valeur peut varier entre 0 et 1, avec 1 signifiant une incertitude complète. Dans le cas ci-dessus, nous pouvons voir que la part des client ayant le produit d'investissement C0 est presque égale à celle des clients ayant C1. Cela veut dire que si on devait piocher un client au hasard, il serait difficile de prédire quel investissement il a choisi, d'où l'incertitude importante.

Dans un premier temps, nous allons calculer l'information gain des variables discrètes, afin de voir la plus importante et la moins importante parmis celles-ci. Dans un deuxième temps, nous calculerons celles des variables continues pour comparer toutes les variables ensemble et sélectionner les plus prédictives.

Le choix de ces variables sera simple et efficace car il suffit de prendre les variables avec les information gain les plus élevés.

# VARIABLES PREDICTIVES DISCRETES

Commençons par les variables discrètes. Au total, nous avons 19 variables discrètes :

- **SE2** : l'âge
- **PE1 - PE15** : historique des investissements
- **IA1 - IA3** : activité d'investissement

Pour savoir laquelle permet de prédire le plus la variable cible, nous devons prendre la variable avec l'information gain plus élevé, et inversement.

Parmis les variables discrètes, nous avons SE2, IA1 et IA3 qui ont plus que deux valeurs distinctes. Sachant que plus le nombre de valeurs distinctes est élevé, plus l'information gain l'est, il faudra normaliser celui-ci afin de ne pas se tromper dans la sélection et d'avoir seulement les meilleurs variables. On fait ceci en divisant l'information gain par l'entropie de la variable prédictive.

## SE2 (localisation géographique du client)

Nombre de valeurs distinctes : 90

Entropie : 5.316317

Entropie normalisée : 0.818921

Entr. conditionnelle : 0.977796217686204

Information gain : 0.021894630918628

Info. gain ratio : 0.0267359390167551

## PE1

Nombre de valeurs distinctes : 2

Entropie : 0.283969226929732

Entr. conditionnelle : 0.99655735563341

Information gain : 0.0000351130414905088

Info. gain ratio : 0.00123650868335805

## PE2

Nombre de valeurs distinctes : 2

Entropie : 0.0308693294663661

Entr. conditionnelle : 0.999616519214385

Information gain : 0.0000743293904470477

Info. gain ratio : 0.00240787188228477

## PE3

Nombre de valeurs distinctes : 2

Entropie : 0.064457144787588

Entr. conditionnelle : 0.998986185778891

Information gain : 0.0000704662825940838

Info. gain ratio : 0.0109322686920586

## PE4

Nombre de valeurs distinctes : 2

Entropie : 0.0538413123037941

Entr. conditionnelle : 0.9979514057465

Information gain : 0.00173944285833216

Info. gain ratio : 0.0323068436467063

## PE5

Nombre de valeurs distinctes : 2

Entropie : 0.889347808279643

Entr. conditionnelle : 0.996969192764592

Information gain : 0.00272165584023987

Info. gain ratio : 0.00306028284423913

## PE6

Nombre de valeurs distinctes : 2

Entropie : 0.548436904499226

Entr. conditionnelle : 0.998163837668814

Information gain : 0.00152701093601826

Info, gain ratio : 0.00278429646781804

## PE7

Nombre de valeurs distinctes : 2

Entropie : 0.347216873457617

Entr. conditionnelle : 0.99915984816503

Information gain : 0.000531000448328878

Info. gain ratio : 0.00152930484927511

## PE8

Nombre de valeurs distinctes : 2

Entropie : 0.388850698767908

Entr. conditionnelle : 0.997832888327744

Information gain : 0.00185796027708807

Info. gain ratio : 0.00477808136381162

## PE9

Nombre de valeurs distinctes : 2

Entropie : 0.24868659636798

Entr. conditionnelle : 0.996916102052509

Information gain : 0.0027747465523229

Info. gain ratio : 0.0111576039595521

## PE10

Nombre de valeurs distinctes : 2

Entropie : 0.881239249989242

Entr. conditionnelle : 0.995109162558807

Information gain : 0.00458168604602449

Info. gain ratio : 0.00519913978647731

## PE11

Nombre de valeurs distinctes : 2

Entropie : 0.732447568297124

Entr. conditionnelle : 0.998552730649621

Information gain : 0.00113811795521102

Info. gain ratio : 0.00155385587238284

## PE12

Nombre de valeurs distinctes : 2

Entropie : 0.720743584547644

Entr. conditionnelle : 0.996137139199499

Information gain : 0.00355370940533295

Info. gain ratio : 0.00493061538322723

## PE13

Nombre de valeurs distinctes : 2

Entropie : 0.836825655014524

Entr. conditionnelle : 0.998504182007473

Information gain : 0.00118666659735933

Info. gain ratio : 0.0014180571427853

## PE14

Nombre de valeurs distinctes : 2

Entropie : 0.406836293591733

Entr. conditionnelle : 0.99950136247199

Information gain : 0.000189486132842309

Info. gain ratio : 0.000465755233313727

## PE15

Nombre de valeurs distinctes : 2

Entropie : 0.329086260740365

Entr. conditionnelle : 0.996868784353648

Information gain : 0.00282206425118414

Info. gain ratio : 0.00857545448672084

## IA1

Nombre de valeurs distinctes : 6

Entropie : 0.0458098918294344

Entropie normalisée : 0.0177216852533274

Entr. conditionnelle : 0.998884629597976

Information gain : 0.00080621900685629

Info. gain ratio : 0.0454933599898416

## IA2

Nombre de valeurs distinctes : 2

Entropie : 0.00288368961660139

Entr. conditionnelle : 0.999485824211724

Information gain : 0.000205024393107989

Info. gain ratio : 0.0710979406131867

## IA3

Nombre de valeurs distinctes : 9

Entropie : 1.25247564043422

Entropie normalisée : 0.395112073586707

Entr. conditionnelle : 0.994224226993765

Information gain : 0.00546662161106737

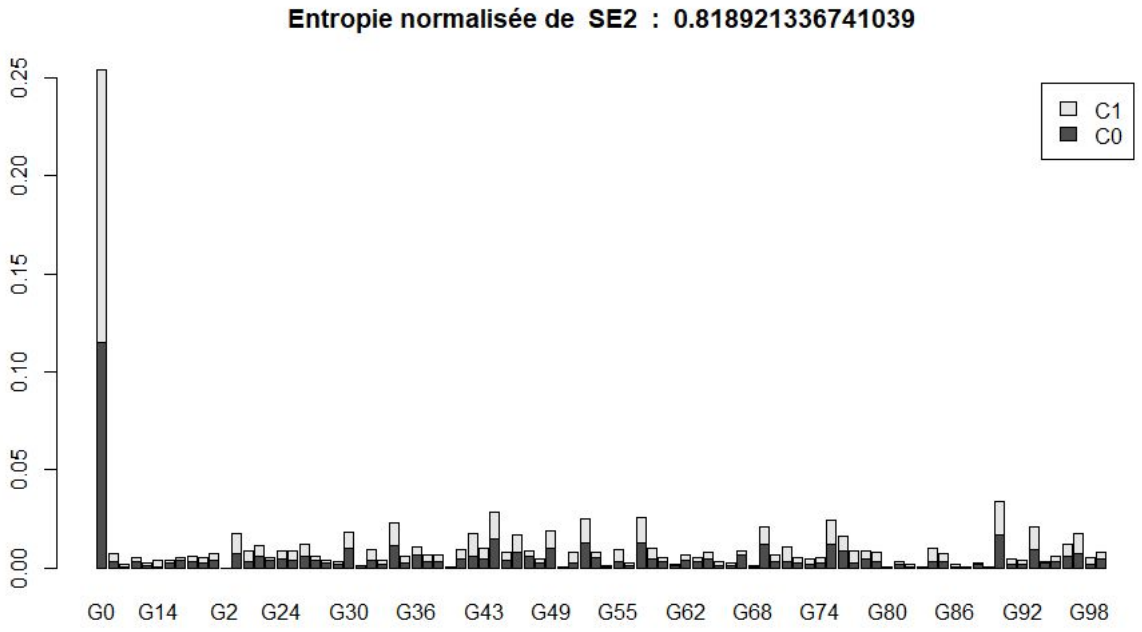
Info. gain ratio : 0.0138356227929029

## Information gain le plus élevé

En regardant l'information gain des variables discrètes, on voit que **SE2** possède celui qui est le plus élevé, ayant comme valeur : 0.021894630918628.

SE2 est donc la variable la plus utile pour prédire la valeur de la variable cible InvType. C'est-à-dire qu'avec la connaissance de la localisation géographique d'un client, on aura une meilleure idée de quel produit d'investissement lui proposer.

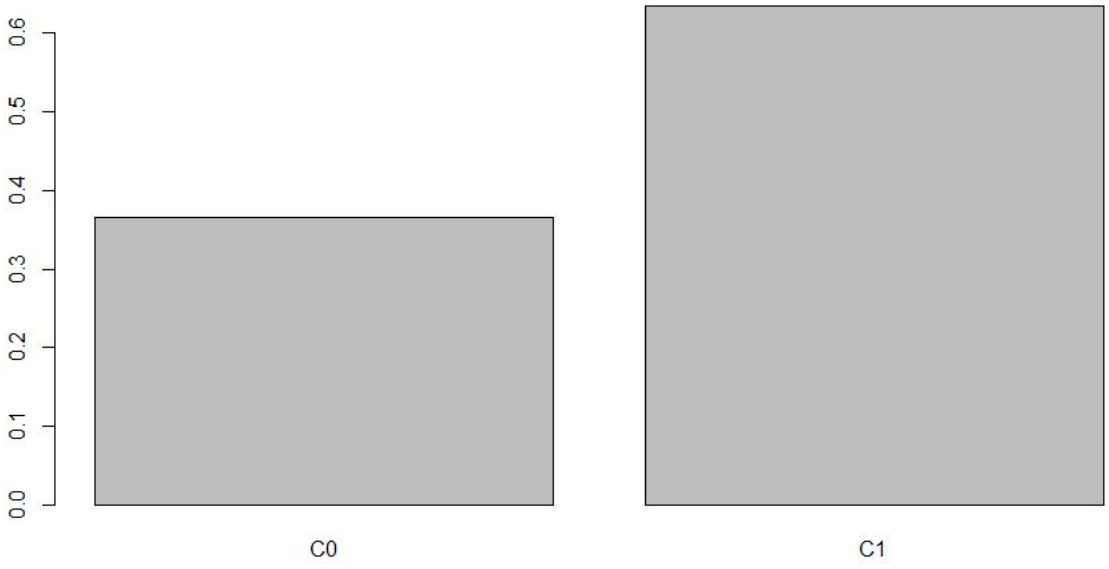
Voici la distribution de probabilité de InvType selon la valeur de SE2 avec un barplot :




On s'aperçoit que l'entropie normalisée de SE2 est relativement haute. Cela signifie que l'incertitude est importante et qu'il sera difficile de prédire la localisation géographique d'un client choisi au hasard. En effet, même si le barplot nous montre que la plupart des clients vivent dans la zone "G0", ces clients ne représentent que 25% et donc les 75% restants sont éparpillés dans les autres zones.

L'attribut le plus prédictive chez les attributs discrets est donc SE2 et cela peut s'expliquer par le fait que dans certaines zones on remarque qu'un des deux produits d'investissement est favorisée, et ce qui nous permettrait donc de savoir plus facilement lequel proposer au client selon sa zone. Un exemple serait la zone "G42", où environ 63% des clients ont choisi le produit d'investissement C1.

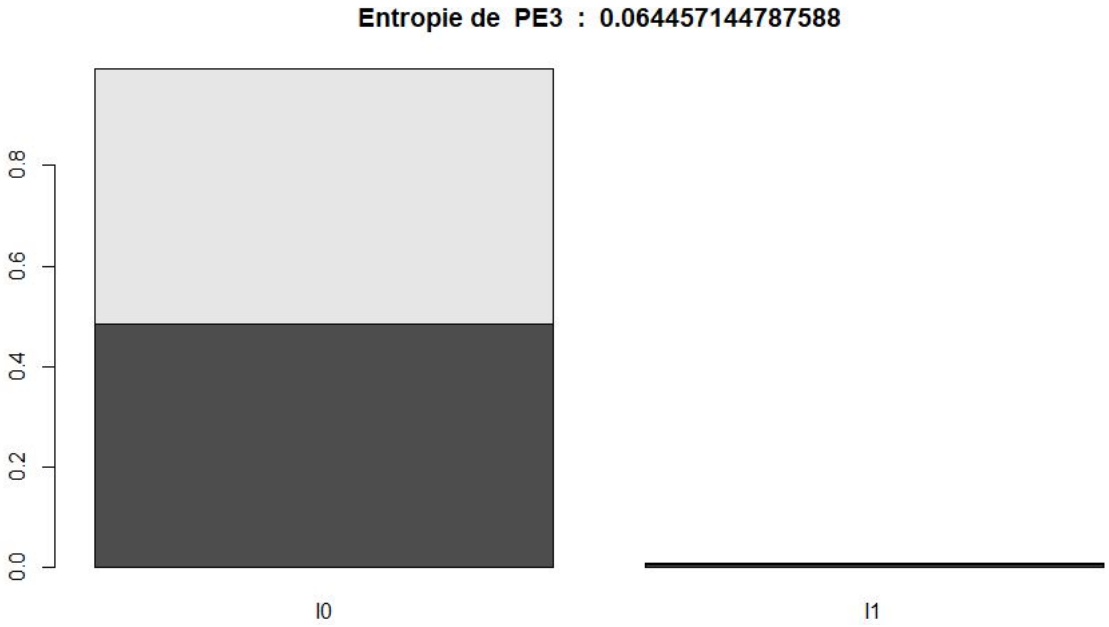
Voici la distribution de probabilité de InvType, sachant SE2 = G42, où l'on voit que C1 est favorisé :




## Information gain le plus bas

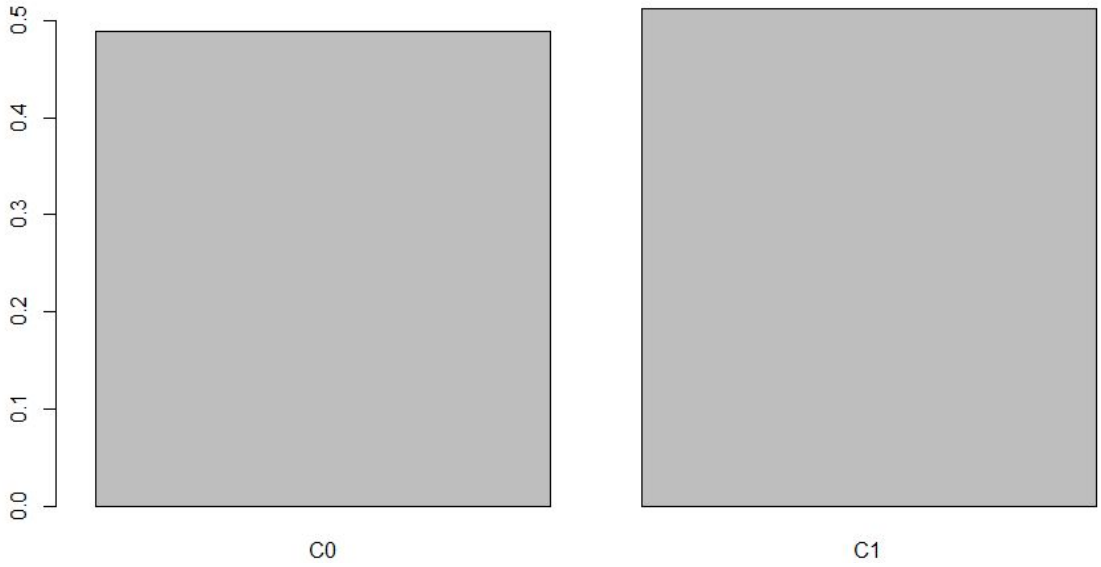
À l'inverse, la variable discrète avec l'information gain le plus bas est **PE3**, ayant comme valeur : 0.0000704662825940838. C'est donc l'attribut le moins utile pour prédire le produit d'investissement d'un client.

Voici la distribution de probabilité de InvType selon la valeur de PE3 avec un barplot :



L’entropie de PE3 est très basse et s’explique par le fait que 99% des clients n’avaient pas le produit d’investissement PE3 l’année dernière. Nous n’avons donc presque pas d’incertitude et il est facile de prédire si un client choisi au hasard avait ou pas le produit PE3.

Voici la distribution de probabilité de InvType pour tous les clients n’ayant pas le PE3 l’année dernière.




Nous pouvons voir qu’environ 49% de ces clients ont acheté le produit d’investissement C0 et 51% le produit C1. La connaissance de l’achat de PE3 ne nous sert donc pas à prédire le produit d’investissement qu’il a acheté.

En conclusion, le fait que 99% des clients n’ont pas acheté PE3 et que la probabilité de C0 / C1 chez ces clients est quasiment pareil, nous permet de justifier que cette variable possède l’information gain le plus bas.

# VARIABLES CONTINUES

Pour les variables continues, nous devons tout d'abord trouver un seuil qui va partager nos instances en deux groupes homogènes. Nous allons calculer l'information gain pour chaque seuil et garder celui pour lequel la valeur de l'information gain est maximale.

## SE1

Seuil : 53

Information gain : 0.036529359663521

Info. gain ratio : 0.041097320892815

## BA1

Seuil : 24

Information gain : 0.00916062634958559

Info. gain ratio : 0.038015085502612

## BA2

Seuil : 12568.62

Information gain : 0.000750181027136643

Info. gain ratio : 0.00846524300490571

## BA3

Seuil : 18899

Information gain : 0.144456733343609

Info. gain ratio :  0.147370840437726

## BA4

Seuil : 16396

Information gain : 0.048956528883203

Info. gain ratio : 0.0530127180641582

## BA5

Seuil : 16396

Information gain : 0.0559759165004248

Info. gain ratio : 0.0594598867098358

## BA6

Seuil : 48126

Information gain : 0.00147902438003666

Info. gain ratio : 0.00181085262180566

## BA7

Seuil : 16371

Information gain : 0.101643849636697

Info. gain ratio : 0.102630852007384

# TABLEAU RECAPITULATIF (ORDRE SELON INFORMATION GAIN)

| Nom de la variable | Information gain | Information gain ratio |
|---|---|---|
| **BA3** | 0.144456733343609 | 0.147370840437726 |
| **BA7** | 0.101643849636697 | 0.102630852007384 |
| **BA5** | 0.0559759165004248 | 0.0594598867098358 |
| **BA4** | 0.048956528883203 | 0.0530127180641582 |
| **SE1** | 0.036529359663521 | 0.041097320892815 |
| **SE2** | 0.021894630918628 | 0.0267359390167551 |
| **BA1** | 0.00916062634958559 | 0.038015085502612 |
| **IA3** | 0.00546662161106737 | 0.0138356227929029 |
| **PE10** | 0.00458168604602449 | 0.00519913978647731 |
| **PE12** | 0.00355370940533295 | 0.00493061538322723 |
| **PE15** | 0.00282206425118414 | 0.00857545448672084 |
| **PE9** | 0.0027747465523229 | 0.0111576039595521 |
| **PE5** | 0.00272165584023987 | 0.00306028284423913 |
| **PE8** | 0.00185796027708807 | 0.00477808136381162 |
| **PE4** | 0.00173944285833216 | 0.0323068436467063 |
| **PE6** | 0.00152701093601826 | 0.00278429646781804 |
| **BA6** | 0.00147902438003666 | 0.00181085262180566 |
| **PE13** | 0.00118666659735933 | 0.0014180571427853 |
| **PE11** | 0.00113811795521102 | 0.00155385587238284 |
| **IA1** | 0.00080621900685629 | 0.0454933599898416 |
| **BA2** | 0.000750181027136643 | 0.00846524300490571 |
| **PE3** | 0.000704662825940838 | 0.0109322686920586 |
| **PE7** | 0.000531000448328878 | 0.00152930484927511 |
| **IA2** | 0.000205024393107989 | 0.0710979406131867 |
| **PE14** | 0.000189486132842309 | 0.000465755233313727 |

| | | |
|---|---|---|
| **PE2** | 0.0000743293904470477 | 0.00240787188228477 |
| **PE1** | 0.0000351130414905088 | 0.000123650868335805 |
| **PE3** | 0.0000704662825940838 | 0.0109322686920586 |

# ATTRIBUTS LES PLUS IMPORTANTS

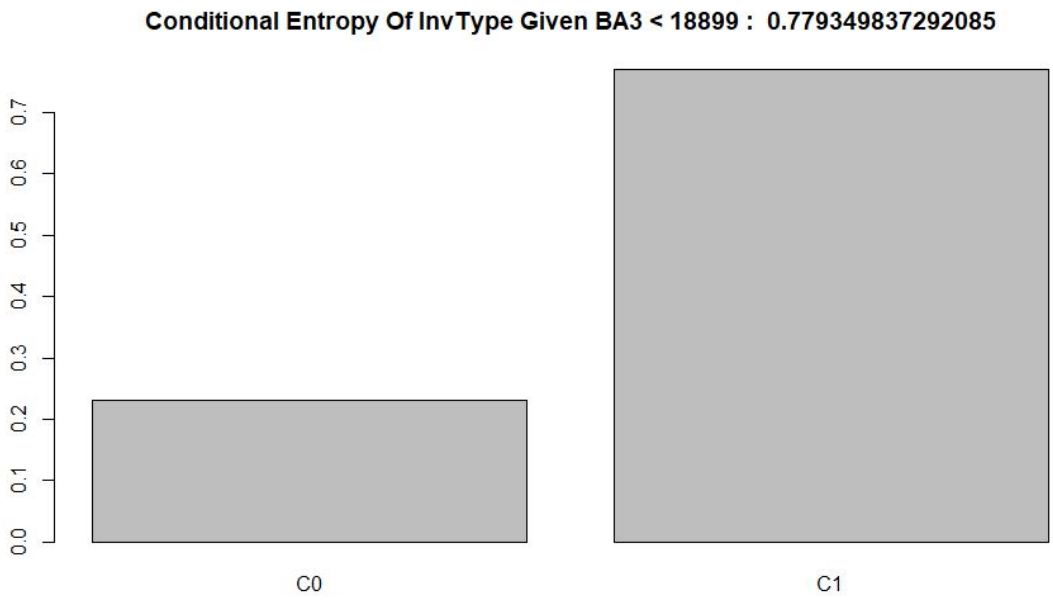
Les 3 variables suivantes sont celles qui ont eu les information gain les plus élevés, et nous allons maintenant regarder de plus près comment chaque valeur de ces variables impactent le choix du produit d'investissement (C0, C1) des clients.

## BA3

BA3 possède l'information gain le plus important parmis toutes les variables, avec une valeur de : 0.144456733343609.

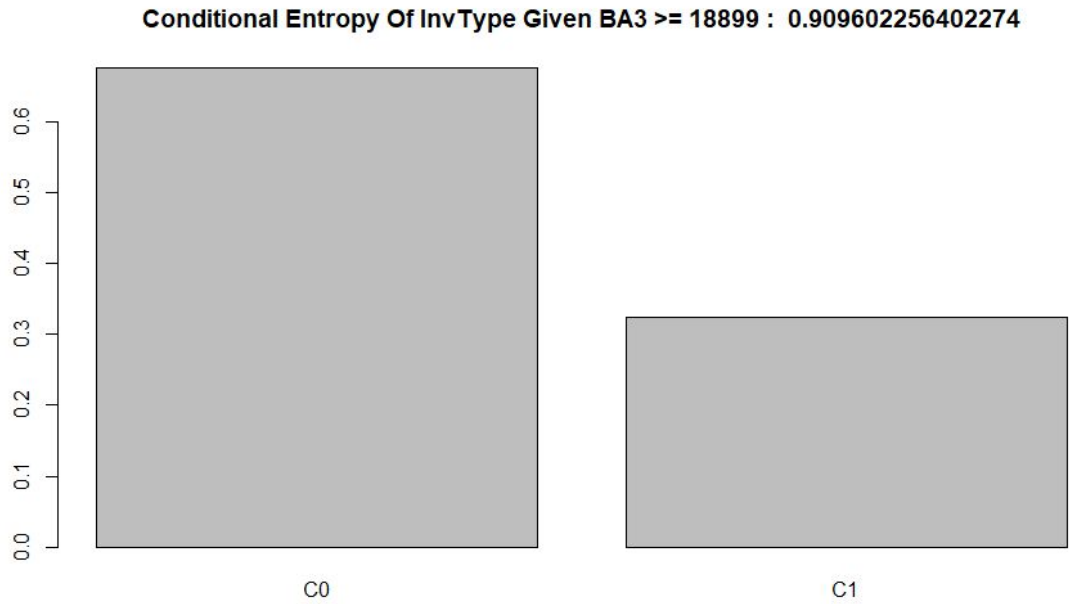
Cet information gain a été obtenu en prenant comme seuil : 18899. Ce seuil nous a permis de départager les instances selon la valeur de leur BA3, et de rendre la variable sous forme binaire.

Voici la distribution de la probabilité de invType pour BA3 < 18899 :



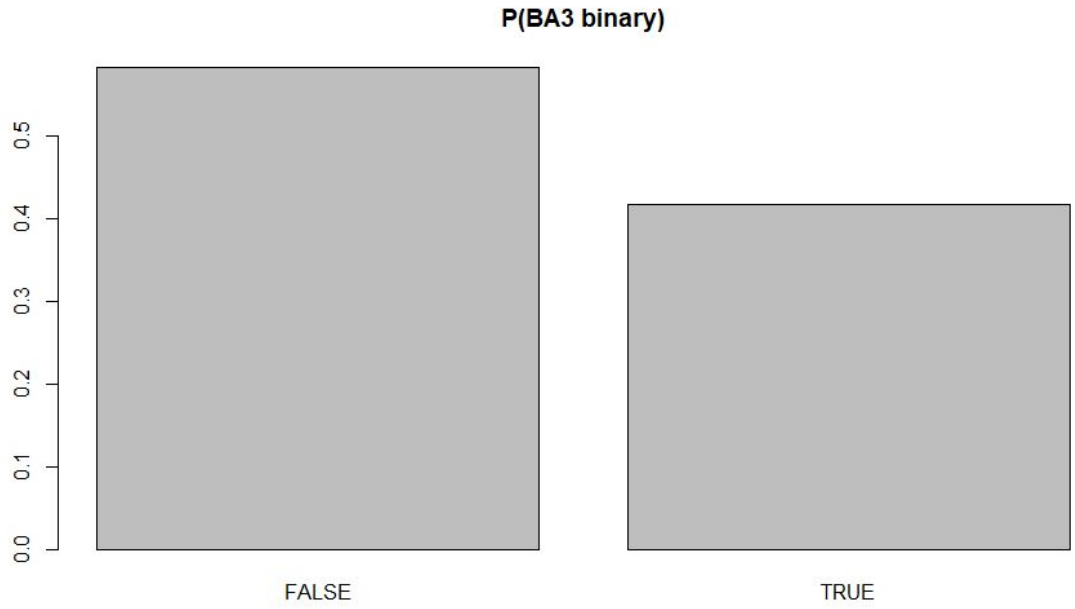

Selon le barplot ci-dessus, nous voyons bien que le produit d'investissement C1 est largement favorisé pour les clients ayant une valeur de BA3 inférieure à 18899. En effet, environ 76% ont choisi le produit d'investissement C1.

Voici la distribution de la probabilité de invType pour BA3 >= 18899 :




Tout comme pour BA3 < 18899, le barplot ci-dessus nous montre qu'un des produits d'investissement est favorisée. Dans le cas de BA3 >= 18899, c'est C0 qui a une part plus importante d'environ 67%.

Voici la distribution de la probabilité de la variable BA3 sous forme binaire :



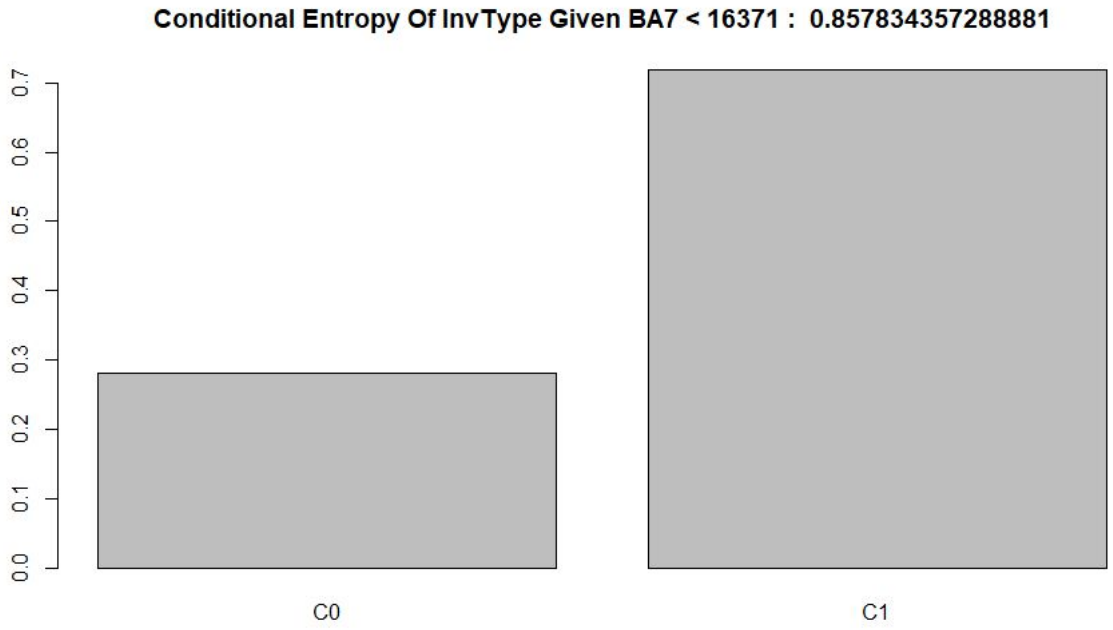

Comme on peut le voir, on a plus de clients qui ont plus de 18899 pour l'activité bancaire BA3. En effet, cette part représente environ 58% des clients.

## BA7

BA7 possède le 2ème information gain le plus important parmis toutes les variables, avec une valeur de : 0.101643849636697.

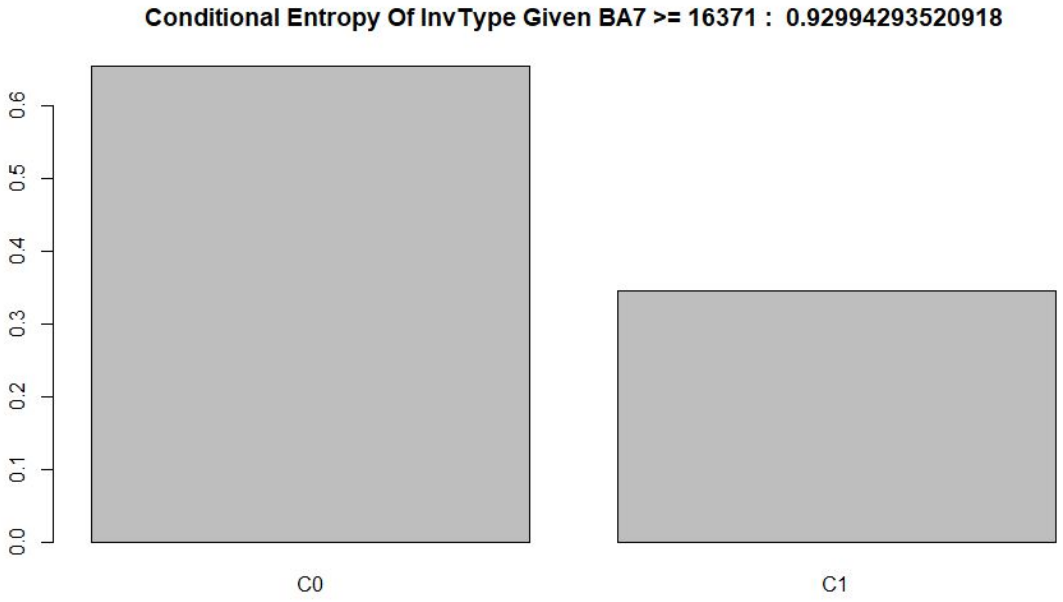
Cet information gain a été obtenu en prenant comme seuil : 16371. Ce seuil nous a permis de départager les instances selon la valeur de leur BA7, et de rendre la variable sous forme binaire.

Voici la distribution de la probabilité de invType pour BA7 < 16371:



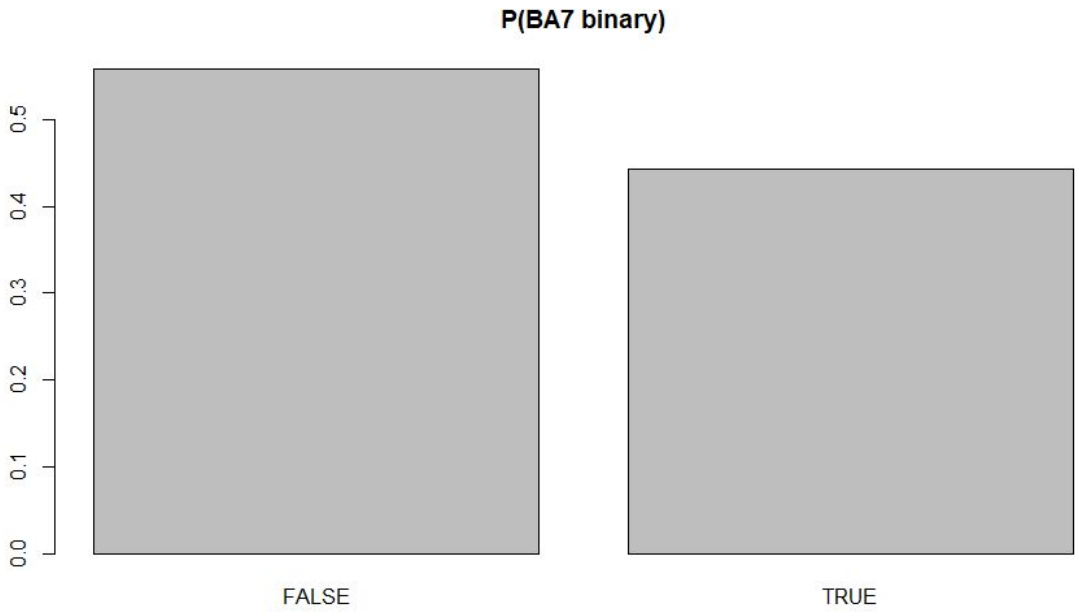

Selon le barplot ci-dessus, nous voyons bien que le produit d'investissement C1 est largement favorisé pour les clients ayant une valeur de BA7 inférieure à 16371. En effet, environ 72% ont choisi le produit d'investissement C1.

Voici la distribution de la probabilité de invType pour BA7 >= 16371 :




Tout comme pour BA7 < 16371, le barplot ci-dessus nous montre qu'un des produits d'investissement est favorisée. Dans le cas de BA7 >= 16371, c'est C0 qui a une part plus importante d'environ 65%.

Voici la distribution de la probabilité de la variable BA7 sous forme binaire :



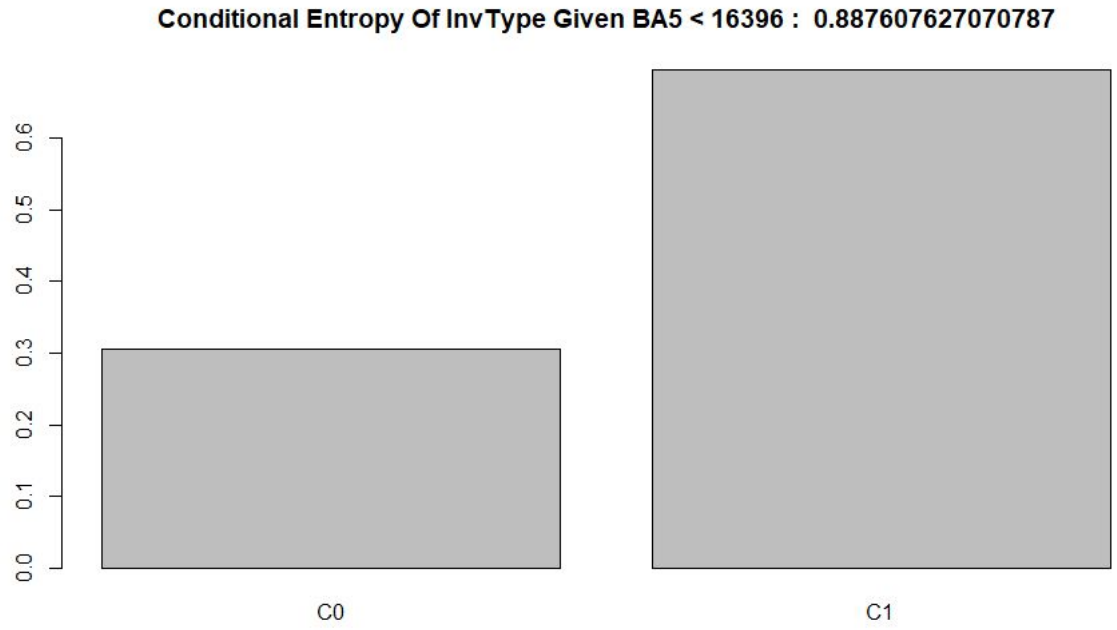

Comme on peut le voir, on a plus de clients qui ont plus de 16371 pour l'activité bancaire BA7. En effet, cette part représente environ 55% des clients.

## BA5

Finalement, BA5 possède le 3ème information gain le plus important parmis toutes les variables, avec une valeur de : 0.0559759165004248.

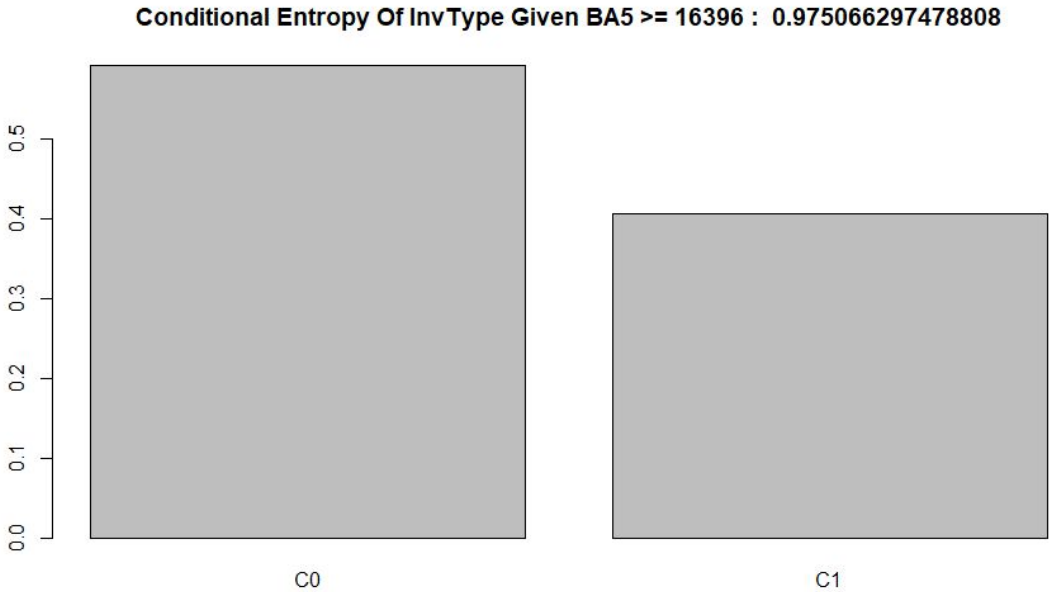
Cet information gain a été obtenu en prenant comme seuil : 16396. Ce seuil nous a permis de départager les instances selon la valeur de leur BA5, et de rendre la variable sous forme binaire.

Voici la distribution de la probabilité de invType pour BA5 < 16396 :



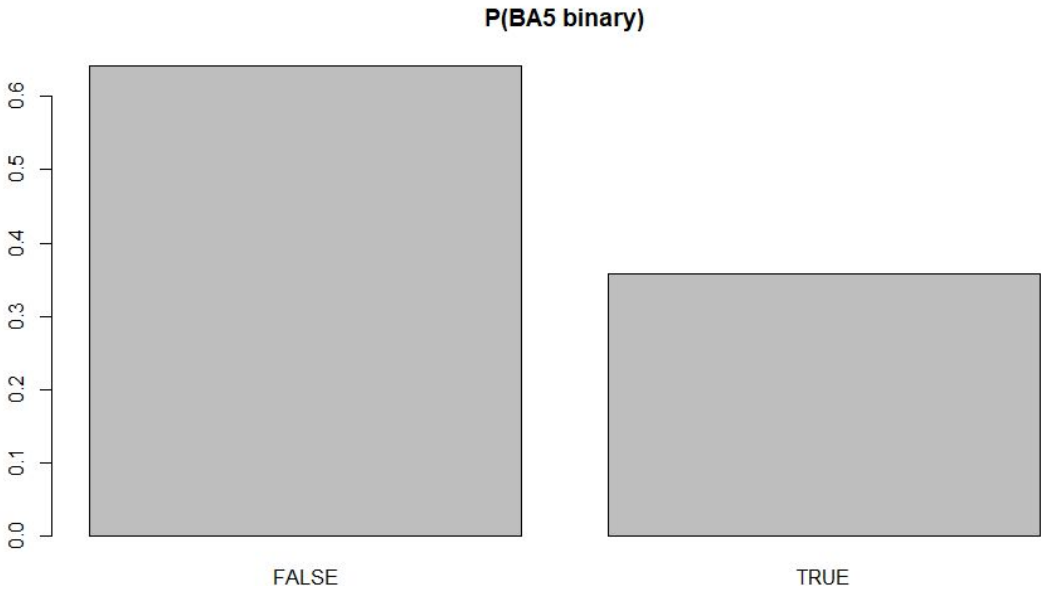

Selon le barplot ci-dessus, nous voyons bien que le produit d'investissement C1 est largement favorisé pour les clients ayant une valeur de BA5 inférieure à 16369. En effet, environ 69% ont choisi le produit d'investissement C1.

Voici la distribution de la probabilité de invType pour BA5 >= 16396 :




Tout comme pour BA5 < 16396, le barplot ci-dessus nous montre qu'un des produits d'investissement est favorisée. Dans le cas de BA5 >= 16396, c'est C0 qui a une part plus importante d'environ 59%.

Voici la distribution de la probabilité de la variable BA5 sous forme binaire :




Comme on peut le voir, on a plus de clients qui ont plus de 16396 pour l'activité bancaire BA5. En effet, cette part représente environ 64% des clients.

# CONCLUSION

Dans le premier TP, nous avons trouvé que les variables les plus informatives étaient : SE1 (l'âge) et IA3. Cependant, avec l'aide de  l'information gain des variables, voici celles qu'on a trouvé : BA3, BA7 et BA5. Nous remarquons que les variables choisies au premier TP ne sont pas les mêmes que celles sélectionnées à l'aide des nouveaux outils qu'on a utilisé.

Les meilleures variables sont donc continues et la raison pour laquelle nous n'avions pas choisi celles-ci c'est parce qu'il est difficile de voir seulement avec des histogrammes, surtout à quand les distributions de probabilités conditionnelles convergent et qu'il y a beaucoup de valeurs distinctes. Ce TP nous a donc permis de voir à quel point le feature selection est important afin de sélectionner les meilleures variables.