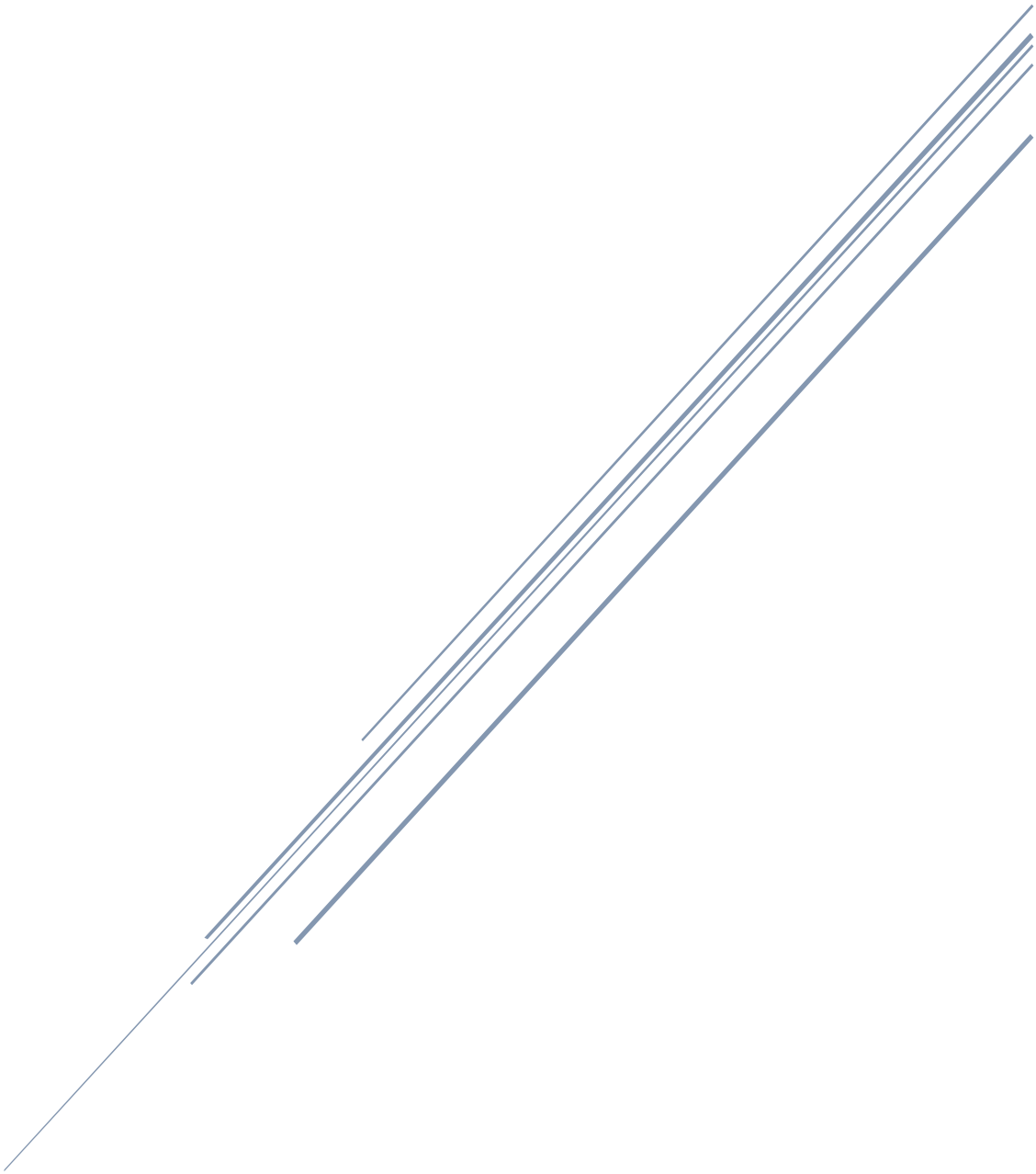
Data Mining Project

Prédiction des factures médicales



Paris 1 Panthéon Sorbonne
Data Mining

Mehdi Padovani

Yasmina Berrada

Shafel Mc Dowall

Table des matières

Introduction

Nous avons pris un ensemble de données de documents médicaux de Singapour pour la période du 1er janvier 2011 au 28 décembre 2015. Notre première étape a consisté à créer un dataset approprié pour commencer notre analyse. Nous avons combiné quatre tables pour créer une table de 13600 lignes et 38 colonnes.

Le dataset a été traité pour remplacer les valeurs manquantes à l'aide de l'algorithme des k-plus proches voisins, tandis que certaines entrées ont été corrigées pour les erreurs de codage ou les fautes de frappe.

Ensuite, nous avons commencé notre analyse exploratoire par une matrice de corrélation et des graphiques simplistes mais puissants. Ces graphiques nous ont indiqué qu'un modèle non linéaire fonctionnerait mieux qu'un modèle linéaire. Cependant, nous avons tout de même testé les deux techniques.

Un ensemble d'algorithmes d'apprentissage supervisé et non supervisé ont été testé sur le dataset et les modèles les plus performants sont présentés dans ce rendu. En fin de compte, ont été gardés les modèles de régression linéaire multiple, et l'arbre de décision.

Les variables pour l'apprentissage supervisé ont été choisies par une analyse en composantes principales (ACP) pour la réduction des dimensions. C'était important car nous disposions d'un grand nombre de variables. Sans la réduction dimensionnelle, nous aurions un R^2 artificiellement gonflé, ce qui a été prouvé par l'expérimentation. Nous avons également essayé la transformation logarithmique dans tous les modèles et avons obtenu de moins bons résultats. Finalement, nous nous sommes contentés d'une normalisation des données numériques.

Au final, l'algorithme de l'arbre de décision non linéaire a surpassé la régression linéaire multiple avec une précision de 62%. Ceci est dû au pays testé, à savoir Singapour. Disposer de soins de santé universels implique un contrôle des coûts très réglementé. Par conséquent, un modèle prédictif utilisant les données des patients ne suffirait pas pour prévoir les coûts.

Valeurs manquantes

On cherche d'abord à savoir si la table comporte des valeurs manquantes. Effectivement, deux variables comportent légèrement plus de 5% valeurs manquantes (6,9 et 9%).

Pour les traiter sans donner une valeur fictive ou sans créer des variables dichotomiques qui vont augmenter la dimensionnalité de notre jeu de données, on remplace ces données manquantes à l'aide de l'aglorithme des KNN (k nearest neighboors ou k plus proches voisins).

Cet algorithme projette les observations sur un axe et on lui attribue (on l'estime par) le groupe d'individus qui a les plus proches voisins. C'est un algorithme supervisé car on prévoit une valeur.

One hot encoding

On remplace ensuite des chaînes de caractères « Yes » et « No » par 0 et 1. On transforme les observations d'une variable qualitative en valeur numérique afin de mieux traiter cette variable par la suite, notamment lors de l'ACP (analyse en composante principale).

On continue la transformation de variable afin de rendre les modalités plus compréhensibles et mieux structrées. Et on crée des variables indicatrices pour ces modalités. Cette méthode a l'avantage de transformer des données catégorielles en données numériques mais a le désavantage de créer plus de dimensions. Afin de pallier cette difficulté, on peut tenter l'ACM qui va comme l'ACP, non seulement réduire la dimension et nous permettre d'étudier la corrélation entre les variables mais surtout de transformer de la donnée qualitative en donnée quantitative.

On transforme le format date des tables qu'on va ensuite fusionner au format (jour/mois/année).

Exploration des données

On s'est rendu à l'aide de la fonction « str » (pour structure) que certaines variables dichotomiques étaient restées des variables caractères. On les a logiquement transformées en nombre (« integer »).

Dans cette exploration des données, on ajoute une variable qui correspond à la durée pendant laquelle un patient est resté hospitalisé (« legth_of_stay ») et on la transforme en numérique. Puis on crée une autre variable « current_age » qu'on obtient à partir des dates de naissance et qui correspond à l'âge des individus.

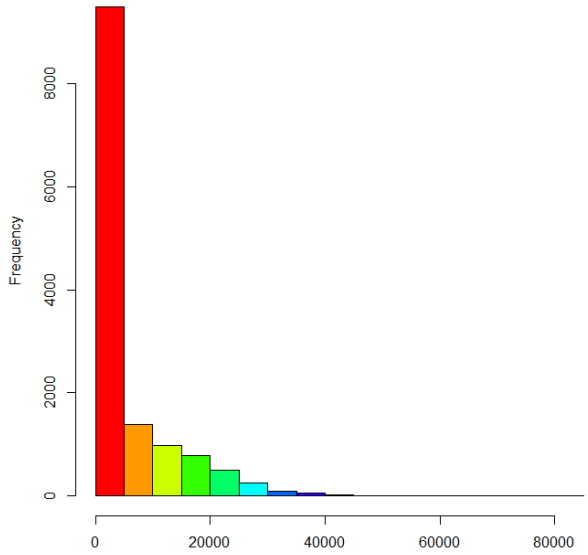
Puis on sépare le jeu de donnée en deux parties, l'une étant numérique et l'autre qualitative. Elles correspondent aux données respectivement numériques et qualitatives de la table qu'on a obtenu après fusion des tables d'origine.

Visualisation des données

En utilisant la fonction « cor », on s'aperçoit que les variables sont très peu corrélées entre elles sauf chez les variables dichotomiques crées précédemment (le fait d'être chinois et d'être malaysien), ce qui porte à croire qu'on peut toutes les garder avant régression linéaire ou autre modèle. Il n'y aurait pas d'information redondante. Mais on sait que cela ne suffit pas et qu'il va falloir faire une ACP ou

une ACM pour s'en assurer. On voit en effet en dressant la matrice des corrélations que certaines variables sont très corrélées entre elles, notamment le sexe et la taille.

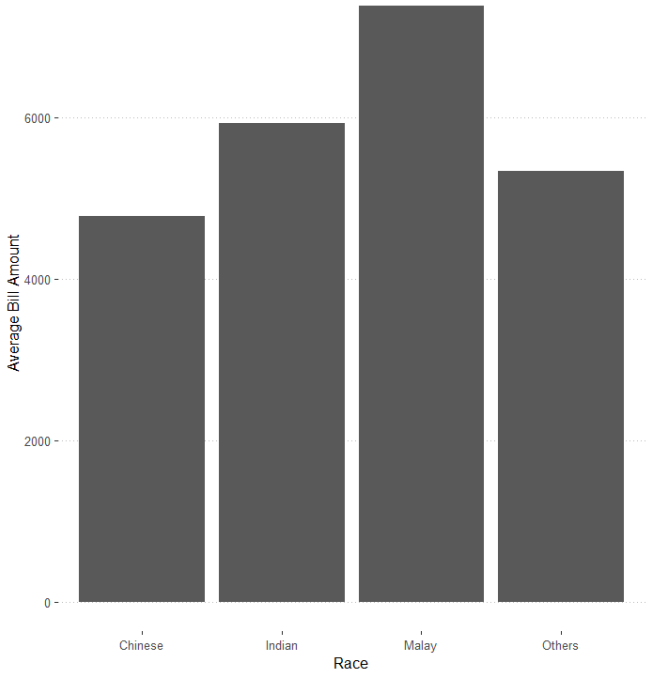
Ensuite, on définit notre cible qui sera le montant (« amount »). Son résumé nous indique que le maximum est de d'environ 81000 et que le minimum est d'environ 80.



La variable « height » ne ressemble pas à une loi normale contrairement aux autres variables numériques. Concernant les variables preop_medication et symptom la valeur 1 est globalement plus importante. On observe le contraire pour la variable medical_history.

On normalise ensuite les données et on procède au V de Cramer afin d'identifier une potentielle corrélation entre les variables qualitatives et les variables numériques. Une faible corrélation s'observe entre resident_status et gender, une relation moyenne entre gender et weight et une relation très forte entre les autres.

Le montant moyen de la facture est plus élevé en Malaysie (7377) et est le moins élevé en Chine (4780). De même respectivement pour les étrangers et les singapourien.



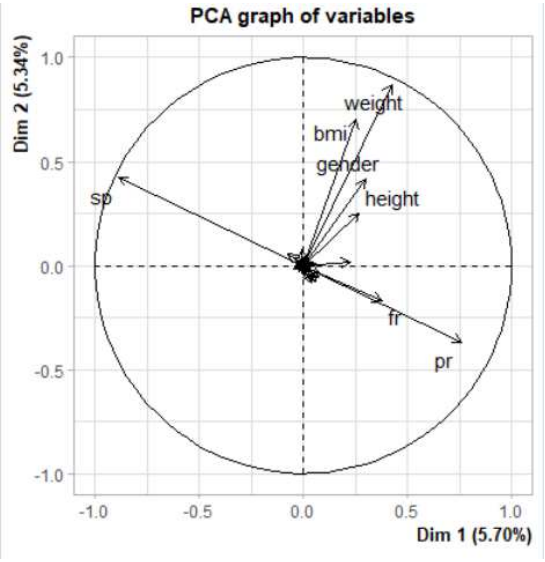
L'indice de masse corporelle en fonction de la nationalité est presque équivalent. En général, les femmes ont un plus grand imc que les hommes sauf les étrangers.

En moyenne les indiennes dont la valeur reste plus longtemps à l'hôpital. Les femmes globalement aussi. Le montant moyen n'est pas significativement différent selon le sexe de l'individu.

Modélisation non supervisée

ACP

Nous avons utilisé l'ACP sur nos données quantitatives.



Plus les flèches sont proches plus corrélations positives et inversement, si elles sont perpendiculaires les variables ne sont pas corrélées

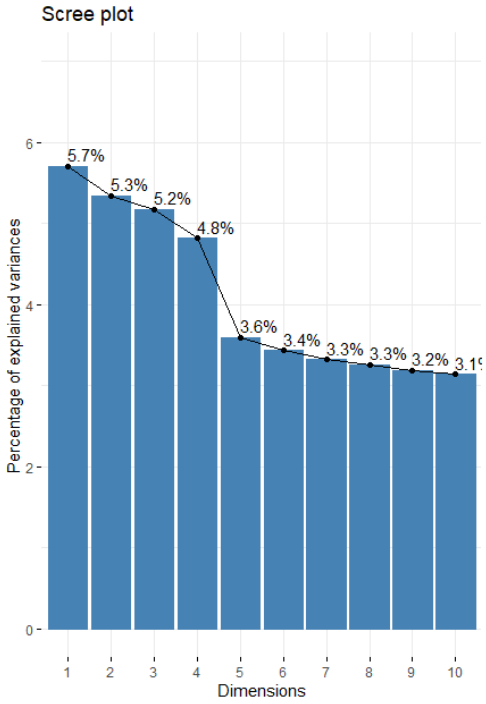
On observe qu'il y a des corrélations positives entre les variables bmi, weight, gender, height, et également entre les variables fr et pr. Sp est correlé négativement avec fr et pr.

Méthode 1 : eigen values

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	1.996727e+00	5.704934e+00	5.704934
Dim.2	1.870613e+00	5.344607e+00	11.049541
Dim.3	1.810417e+00	5.172621e+00	16.222162
Dim.4	1.685626e+00	4.816075e+00	21.038237
Dim.5	1.256232e+00	3.589235e+00	24.627472
Dim.6	1.200756e+00	3.430732e+00	28.058204
Dim.7	1.161982e+00	3.319950e+00	31.378154
Dim.8	1.139914e+00	3.256896e+00	34.635050
Dim.9	1.114829e+00	3.185225e+00	37.820275
Dim.10	1.098076e+00	3.137361e+00	40.957635
Dim.11	1.077724e+00	3.079210e+00	44.036845
Dim.12	1.067817e+00	3.050906e+00	47.087751
Dim.13	1.053047e+00	3.008704e+00	50.096456
Dim.14	1.034798e+00	2.956566e+00	53.053022
Dim.15	1.026267e+00	2.932193e+00	55.985215
Dim.16	1.010649e+00	2.887569e+00	58.872784
Dim.17	1.001752e+00	2.862150e+00	61.734934
Dim.18	9.992145e-01	2.854898e+00	64.589832
Dim.19	9.833601e-01	2.809600e+00	67.399433
Dim.20	9.811774e-01	2.803364e+00	70.202797
Dim.21	9.692669e-01	2.769334e+00	72.972131
Dim.22	9.611525e-01	2.746150e+00	75.718281
Dim.23	9.424918e-01	2.692834e+00	78.411115
Dim.24	9.358207e-01	2.673773e+00	81.084888
Dim.25	9.266569e-01	2.647591e+00	83.732479
Dim.26	9.100673e-01	2.600192e+00	86.332672
Dim.27	9.031772e-01	2.580506e+00	88.913178
Dim.28	8.860497e-01	2.531571e+00	91.444749
Dim.29	8.829630e-01	2.522752e+00	93.967500
Dim.30	8.600835e-01	2.457381e+00	96.424882
Dim.31	7.120457e-01	2.034416e+00	98.459298
Dim.32	4.340959e-01	1.240274e+00	99.699572
Dim.33	1.023141e-01	2.923261e-01	99.991898
Dim.34	2.835725e-03	8.102071e-03	100.000000
Dim.35	4.718439e-26	1.348126e-25	100.000000

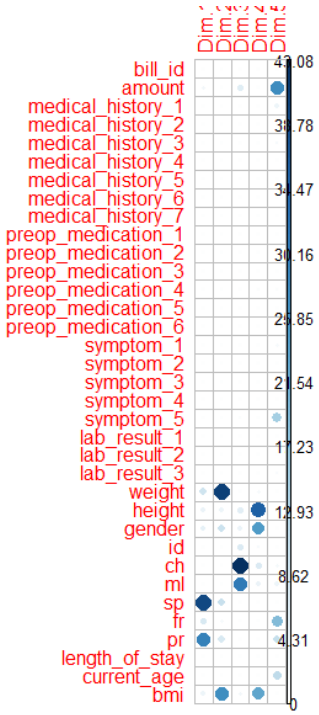
La méthode des eigen values consiste à sélectionner les dimensions dont les valeurs propres sont supérieures à 1. Ceci revient à sélectionner les 17 premières dimensions. Elles ne représentent que 62% de l'information. Nous allons maintenant explorer une deuxième méthode appelée la méthode du coude pout tenter de récupérer un plus gros pourcentage.

Méthode 2 : méthode du coude



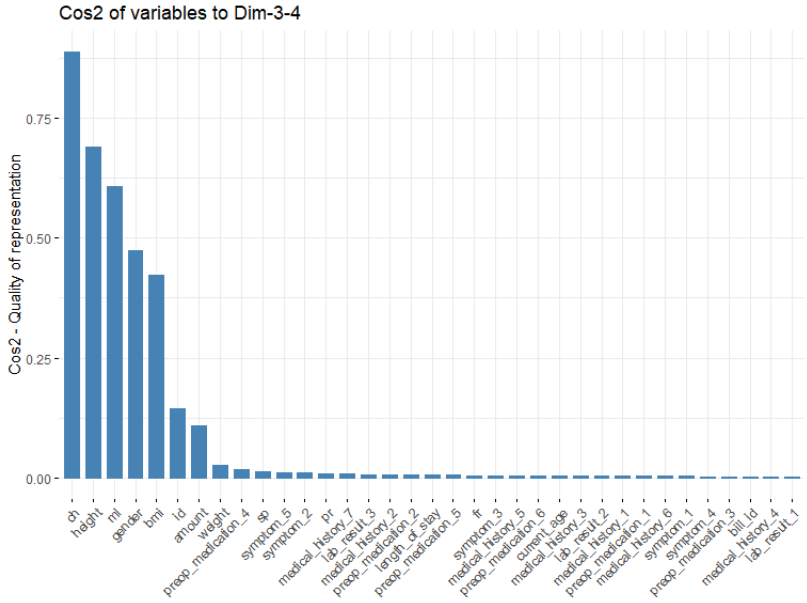
On voit qu'il y a une chute brutale de l'information à partir de la dimension 4, cette méthode indique qu'il faudrait garder les 4 premières dimensions

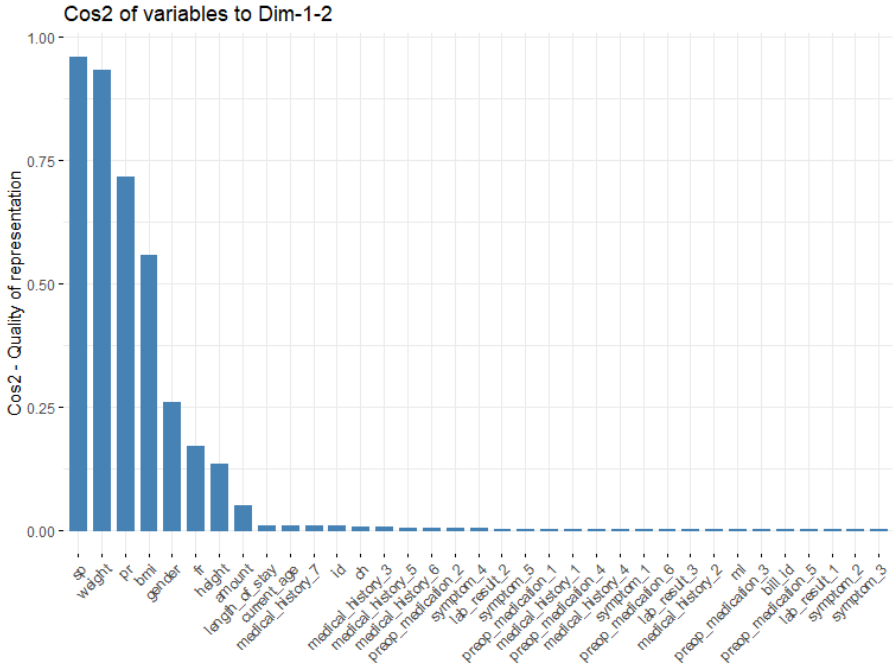
Contribution des variables

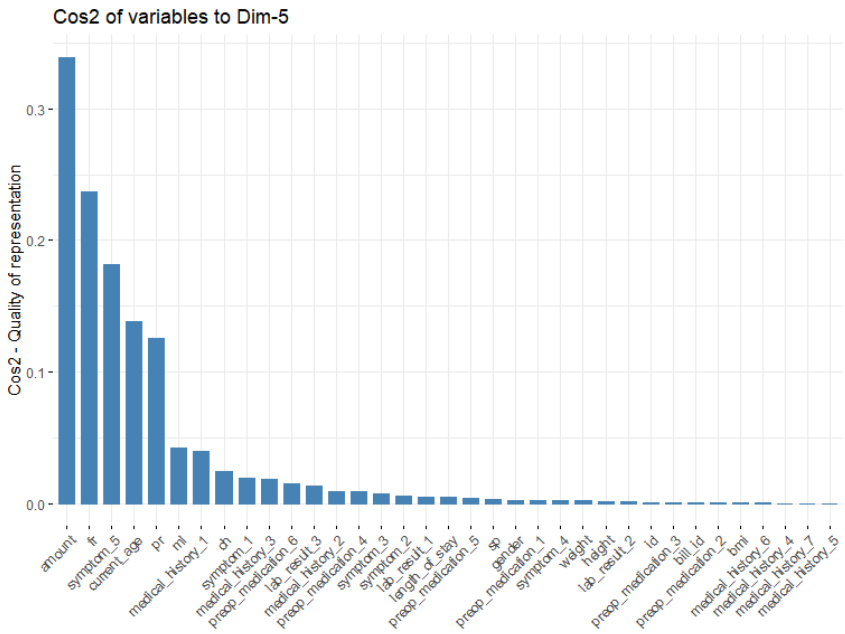


D'après le graph ci-dessous les variables qui contribuent le plus aux dimensions 1 sont pr et sp. Celle qui contribuent le plus à la dimension 2 sont bmi et weight. Les variables « ch » et « ml » représentent le plus la dimension 3. La dimension 4 est representée majoritairement par les variables « height » et « gender» et « bmi ». Les variables « symptom_5 », « fr », «amount » et « current_age » sont celles qui contribuent le plus à la dimension 5.

Cos2 : qualité de représentation







On note que :

Un cos2 élevé indique une bonne représentation de la variable sur les axes principaux en considération. Dans ce cas, la variable est positionnée à proximité de la circonférence du cercle de corrélation.

Un faible cos2 indique que la variable n'est pas parfaitement représentée par les axes principaux. Dans ce cas, la variable est proche du centre du cercle.

Pour une variable donnée, la somme des cos2 sur toutes les composantes principales est égale à 1.

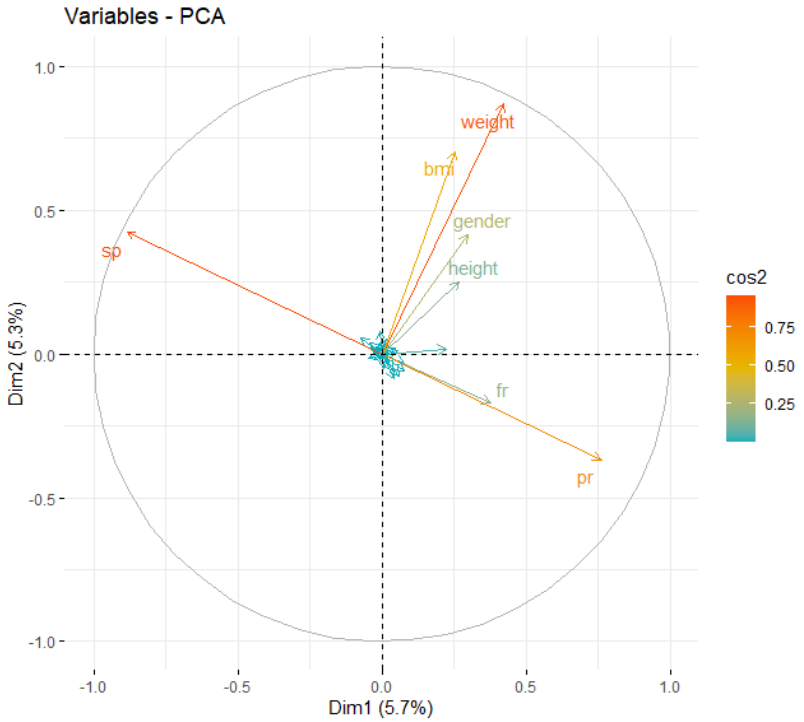
Pour nos variables, plus de 2 axes sont nécessaires pour représenter parfaitement les données. Dans ce cas, les variables sont positionnées à l'intérieur du cercle de corrélation.

En résumé, Les valeurs de cos2 sont utilisées pour estimer la qualité de la représentation.

Plus une variable est proche du cercle de corrélation, meilleure est sa représentation sur la carte de l'ACP (et elle est plus importante pour interpréter les composantes principales en considération)

Les variables qui sont proches du centre du graphique sont moins importantes pour les premières composantes.

Cercle de corrélation en fonction du cos2



On voit bien que les flèches le plus petites (en bleu) sont celles qui ont un cos2 le plus faible, et représentent alors des variables qui n'apportent pas beaucoup en termes d'information.

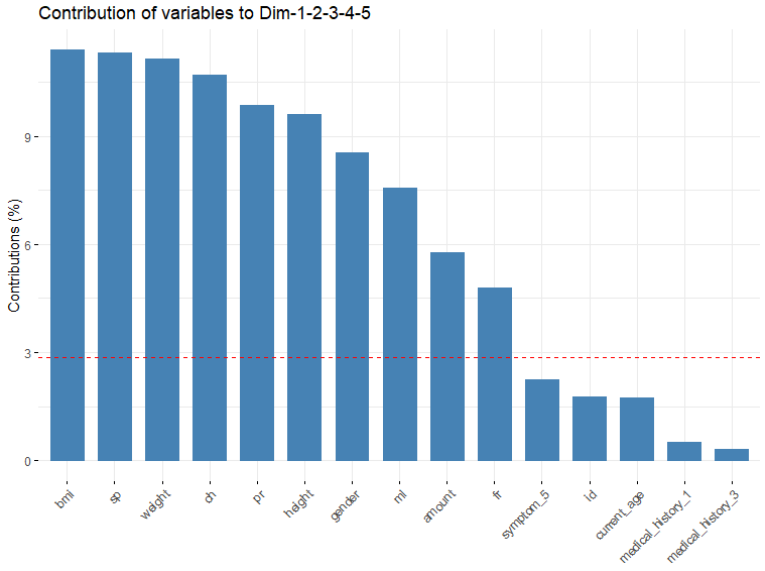
Les variables au cos2 le plus élevé sont : « weight », « sp » , « pr » et « bmi »

Contributions des variables aux axes principaux(1-2-3-4-5)

Le bar plot ci-dessus représente la contribution des 15 variables qui contribuent le plus aux composantes principales.

La ligne en pointillé rouge, sur le graphique ci-dessus, indique la contribution moyenne attendue.

Pour une composante donnée, une variable avec une contribution supérieure à ce seuil pourrait être considérée comme importante pour contribuer à la composante.



Conclusion :

On peut voir que les variables ***« bmi »***, ***« sp »***, ***« weight »***, ***« ch »***, ***« pr »***, ***« height »***, ***« gendre »***, ***« ml »***, ***« amount »***, et ***« fr »*** sont celles qui contribuent le plus aux 5 composantes principales.

Modélisation supervisée

1. Arbres de décision

L'arbre de décision est un type d'algorithme d'apprentissage supervisé qui peut être utilisé à la fois dans les problèmes de régression et de classification. Il fonctionne pour les variables d'entrée et de sortie catégoriques et continues

L'arbre de décision est facile à interpréter, fonctionne même s'il existe des relations non linéaires entre les variables, il ne nécessite pas d'hypothèse de linéarité et n'est pas sensible aux valeurs aberrantes. Cependant L'arbre de décision overfit généralement, il ne fonctionne pas bien sur l'échantillon de validation, c'est pour cela qu'on utilise une méthode connue sous le nom de **l'élagage**.

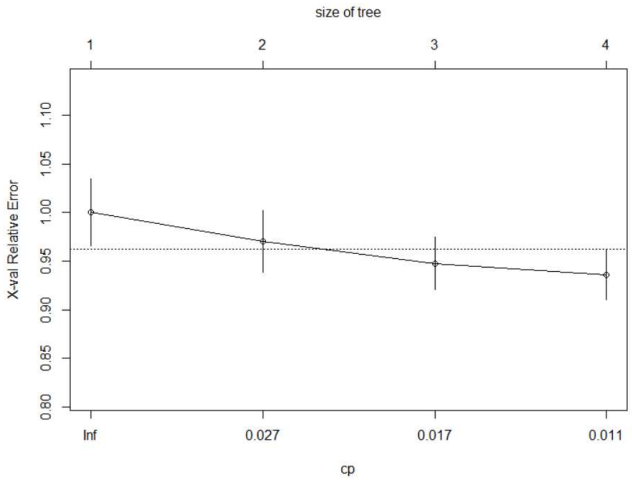
La méthode complexity parameter :

La méthode fait référence au fait de construire un arbre complet et de le « tailler » par la suite.

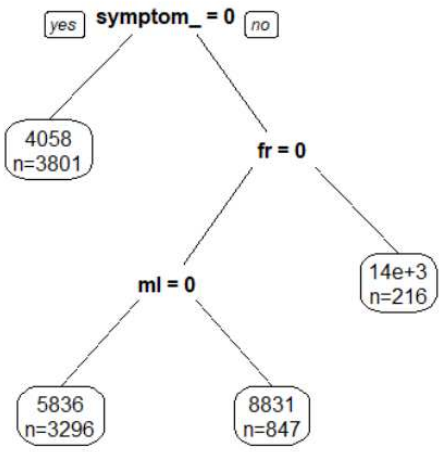
La complexité est mesurée par les paramètres nombres de feuilles dans l'arbre (taille de l'arbre) et le taux d'erreur de l'arbre (taux d'erreur de classification ou somme de l'erreur quadratique)

Nous cherchons à avoir la valeur cp du plus petit arbre qui a la plus petite erreur

En d'autres termes, on fait référence à un compromis entre la taille d'un arbre et le taux d'erreur pour aider à éviter l'overfitting. Ainsi les grands arbres avec un faible taux d'erreur sont pénalisés au profit des arbres plus petits.



Dans ce cas nous choisissons l'arbre ayant CP = 0,01 car il a le moins d'erreur de validation croisée (X-val Relative Error).



Un arbre de décision comprend trois composants principaux:

- *Nœud racine*: le nœud le plus haut est appelé nœud racine. Cela implique le meilleur prédicteur (variable indépendante) ici dans notre cas c'est « symptom_ »
- *Nœud interne:* Les nœuds dans lesquels les variables explicatives sont testés et chaque branche représente un résultat du test. Ici dans notre cas « fr » et « ml »
- *Nœud terminal / terminal*: il détient une étiquette de classe (catégorie) - Oui ou Non (résultat final de la classification).

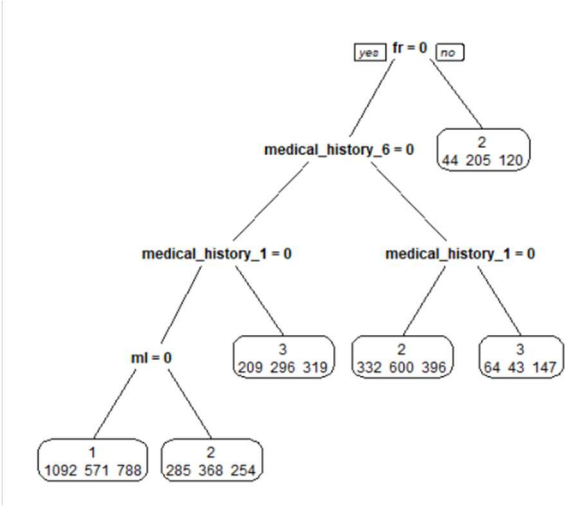
L'algorithme prend en compte tous les individus et recherche les variables qui séparent le mieux les individus. Pour faire ce choix l'algorithme utilise une métrique (l'entropie)

Dans notre cas c'est la variable « symptom_ » qui obtient la meilleure entropie, l'arbre de décision sépare donc la population en fonction de ceux qui ont ce symptôme et ceux qui ne l'ont pas.

Si les individus ne l'ont pas l'arbre est fractionné ensuite en fonction de la variable « fr ». si les individus sont bien « fr » alors l'arbre est fractionné cette fois ci en fonction de la variable « ml »

Elagage de l'arbre avec le cp optimal et représentation graphique de l'arbre optimal

On utilise cette technique pour corriger les problèmes d'overfitting. Elle réduit la taille des arbres de décision en supprimant les sections de l'arborescence qui fournissent peu de puissance pour classer les instances



Matrice de confusion

```
Confusion Matrix and Statistics


     1   2   3
1 257 166 126
2 128 344 112
3   8 108 424
```

Les valeurs présentes sur la diagonale représentent les valeurs bien prédites (true positif), et les valeurs en dehors de la diagonale représentent les valeurs mal prédites.

Statistiques

```
Overall Statistics

               Accuracy : 0.623
                 95% CI : (0.5992, 0.6463)
    No Information Rate : 0.4167
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4335

 Mcnemar's Test P-Value : < 2.2e-16
```

Nous sommes à une **précision** de 61% . Ce pourcentage représente le rapport du nombre de valeurs bien prédites sur le nombre de valeur totale ou aussi connue sous le nom de true positifs. Cela veut dire que lorsque nous classifions nos observations nous avons raison 61% du temps toutes choses égales par ailleurs.

On observe que la p-value est très faible ce qui indique que notre modèle est significatif.

Statistiques par class

```
Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.6539   0.5566   0.6405
Specificity            0.7719   0.7725   0.8853
Pos Pred Value         0.4681   0.5890   0.7852
Neg Pred Value         0.8790   0.7484   0.7899
Prevalence             0.2349   0.3694   0.3957
Detection Rate         0.1536   0.2056   0.2534
Detection Prevalence   0.3282   0.3491   0.3228
Balanced Accuracy      0.7129   0.6646   0.7629
```

- **La sensibilité** d'un classificateur est le rapport entre la quantité correctement identifiée comme positive et la quantité réellement positive.

Sensitivity = TP / FN+TP

- **La spécificité** d'un classificateur est le rapport entre la quantité correctement classée comme négative et la quantité réellement négative.

Specificity = TN/FP+TN

Notre modèle a relativement une bonne spécificité.

La classe avec la plus grande sensibilité est la classe 1. Ce qui veut on se trompe le moins en prédisant la class 1.

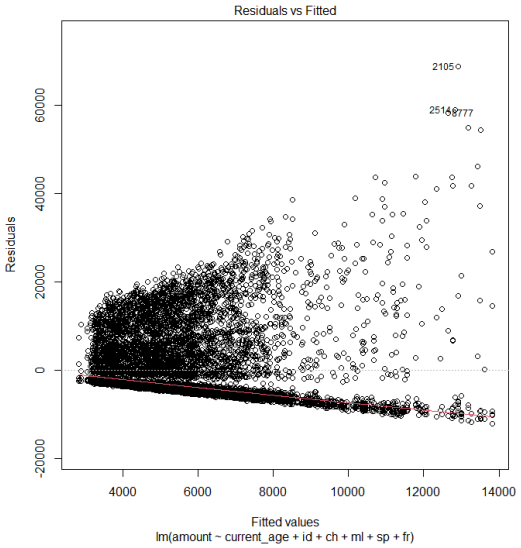
2. Régression linéaire

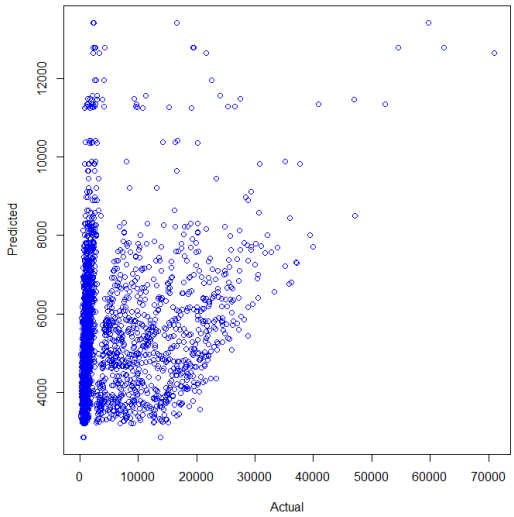
Un nouveau dataframe a été créé à partir de l'ensemble initial de variables afin d'inclure les variables mises en évidence par le processus de sélection ACP. Les données ont été divisées en un ensemble de données de formation et de test en utilisant un rapport 80:20. Par un processus d'élimination, nous avons le modèle linéaire final.

Modèle linéaire de formation :

Amount= β_0 + β_1 current age + β_4 Indian+ β_5 Chinese+ β_6 Malay+ β_7 Singaporean+ β_8 Foreigner + ε_0

Résultats:






```
Residuals:
   Min     1Q Median     3Q    Max
-12191  -4452  -3182   2264  68934

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2750.386    469.154   5.862 4.69e-09 ***
current_age   52.957      4.971  10.652  < 2e-16 ***
id           794.948    384.971   2.065   0.0390 *
ch          -534.432    322.523  -1.657   0.0975 .
ml          1927.963    347.632   5.546 2.99e-08 ***
sp          -995.492    202.694  -4.911 9.18e-07 ***
fr          4459.469    378.340  11.787  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7548 on 10873 degrees of freedom
Multiple R-squared:  0.05113,   Adjusted R-squared:  0.0506
F-statistic: 97.64 on 6 and 10873 DF,  p-value: < 2.2e-16
```

Avec un R^2 de 5% et une RMSE de 7351, on peut voir que le modèle a un faible pouvoir de prédiction. Bien qu'il faille noter que les variables proposées par l'ACP étaient en fait significatives, et que leurs signes étaient conformes aux attentes. Cependant, les faibles propriétés prédictives d'un modèle linéaire étaient attendues car la matrice de corrélation indiquait que les seules relations significatives étaient celles entre le sexe, la taille et le poids.

Conclusion

Dans le cadre de nos recherches, nous nous sommes efforcés de prévoir les coûts médicaux des patients à Singapour sur la base de leurs dossiers médicaux, dans cette entreprise, nous avons obtenu un succès modéré. Nous avons tenté de prédire le montant de la facture en utilisant une régression linéaire multiple et un modèle non linéaire, en particulier un arbre de décision.

En utilisant l'algorithme des k plus proches voisins, nous avons pu compléter les données manquantes pour créer un ensemble de données complet. Nous avons ensuite utilisé l'analyse en composantes principales pour identifier les variables les plus significatives. Ces variables étaient le poids, le Singapourien, le résident permanent et l'IMC, un indicateur indirect de l'obésité. Elles ont ensuite été

ajoutées au modèle linéaire et étaient en effet statistiquement significatives. Cependant, avec un R^2 de 5 %, nous savions que la puissance de prédiction du modèle était faible, et donc qu'il ne correspondait pas aux données.

L'algorithme de l'arbre de décision non supervisé a eu beaucoup plus de succès, avec une précision de la puissance de prédiction de 62%. Bien que nous souhaitions une précision plus élevée, ce n'est pas réaliste. Singapour dispose d'un système de santé universel, ce qui signifie que les coûts médicaux sont très réglementés. Cela signifie que le facteur déterminant le plus important en matière de coûts serait le statut de résident, qui a été mis en évidence comme étant significatif dans les modèles linéaires et non linéaires. Par conséquent, notre modèle montre que les soins de santé universels fournis à Singapour ne font pas de discrimination en fonction du sexe ou du poids, etc.