[Ecole Centrale Marseille]

[Sujet 3:Analyse de données]

[Projet S8 SISN]

Ailin HE
Chengshuang YIN
Qing ZHU
Rudy NOYELLE

Sommaire

Introduction

Dans le cadre de ce projet, nous avons pour mission d'analyser des données afin de déterminer des corrélations entre les données. Nous avons choisi de travailler sur une base de données médicales concernant des patients atteints de diabète.
La base de données provient du site "UCI Machine learning repository" (http://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008).
Elle est issu de l'article Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014, Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore.

L'ensemble de ces données représentent 10 années (1998-2008) de soin clinique provenant de 130 hôpitaux américains. Ces données incluent 55 caractéristiques représentant des informations du patients et médicales. Les informations ont été extraites de la base de données pour les rencontres qui satisfait les critères suivants :
(1) C'est une rencontre de malade hospitalisé (une admission hospitalière).
(2) C'est une rencontre diabétique, en somme, une rencontre pendant laquelle n'importe quelle sorte de diabète a été entrée au système comme un diagnostic.
(3) La longueur de séjour était au moins 1 jour et au plus 14 jours
(4) Les essais en laboratoire ont été exécutés pendant la rencontre.
(5) Les médicaments ont été administrées pendant la rencontre.
Nous avons analysé ces données avec différentes méthodes : PCA, k-means et matrice de confusion..

Contexte et intêret de l’analyse

Il est de plus en plus reconnu que la gestion de l'hyperglycémie chez le patient hospitalisé a une influence significative sur les résultats, tant en terme de morbidité et mortalité . Cette reconnaissance a conduit au développement de protocoles formalisés dans le cadre de l'unité de soins intensifs (USI) avec des cibles de glucose rigoureuses dans de nombreux établissements. Pourtant, ce n’est pas le cas pour les patients non-USI. Au contraire, des preuves anecdotiques suggèrent que la prise en charge des patients hospitalisés est arbitraire et conduit souvent à l'absence totale de traitement ou à des fluctuations importantes du glucose lorsque des stratégies de prise en charge traditionnelles sont utilisées. Bien que les données soient peu nombreuses, des essais contrôlés récents ont démontré que les stratégies d'hospitalisation basées sur le protocole peuvent être à la fois efficaces et sûres. À ce titre, la mise en place de protocoles dans des règles hospitalières est maintenant recommandée. Cependant, il existe peu d'évaluations nationales des soins du diabète chez les patients hospitalisés qui pourraient servir de référence pour ce changement.

Une vaste base de données cliniques a été entreprise pour examiner les schémas historiques de prise en charge du diabète chez les patients diabétiques admis dans un hôpital américain et pour orienter les orientations futures possibles d'améliorer la sécurité des patients. Les bases de données cliniques contiennent des données précieuses mais hétérogènes et difficiles en termes de valeurs manquantes, d'enregistrements incomplets ou incohérents, et de dimensionnalité élevée comprises non seulement par le nombre de caractéristiques, mais aussi par leur complexité. En outre, l'analyse des données externes est plus difficile que l'analyse des résultats d'une expérience ou d'un essai soigneusement conçu, car elle n'a aucun impact sur la manière et le type d'informations collectées. Néanmoins, il est important d'utiliser ces énormes quantités de données pour trouver de nouvelles informations / connaissances qui ne sont peut-être pas disponibles n'importe où.

Description et pré-traitement de la base de données

Description de la base de données

La base de données contient environ 100 000 rencontres. Chaque rencontre a 55 attributs. (Voir Annexe 2)

Pré-traiment des données

Données manquantes

Le poids a 97 % de valeurs manquantes. Nous décidons donc ne pas le prendre en compte et de le supprimer.
L'attribut "Payer code" a 52% de valeurs manquantes. N'étant pas toujours significatif, nous décidons de ne pas le prendre en compte en donc de le supprimer.
Nous décidons de traiter que la partie de la base de données où les données sont complètes. Il reste 49 735 rencontres.
Pour finir, nous supprimons les attributs qui présentent toujours la même valeur sur ces 49 735 rencontres car ils ne sont porteurs d'aucune information significative pour un traitement. Ceci est le cas pour quelques tests de protéine.
Notre base de données détient finalement 49 735 rencontres et 34 attributs.

Numérisation des données

Pour pouvoir traiter la base de données selon différentes méthodes, nous avons besoin de numériser les données. Pour chaque attribut qualitatif, nous attribuons donc à chaque catégorie un nombre unique.

attributs gender :
Female = -1
male = 1

Attributs Age:

moyenne des âges (ex : (0-10) = 5)

attributs max_glu_serum:
None=0
num=num

attributs metformine,, répaglinide, natéglinide, chlorpropamide, glimépiride, acétohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examen, sitagliptine, insuline, glyburide-metformine, glipizide-metformine, glimépiride-pioglitazone, metformine-rosiglitazone et metformine-pioglitazone :
No=0
up=2
down=-2

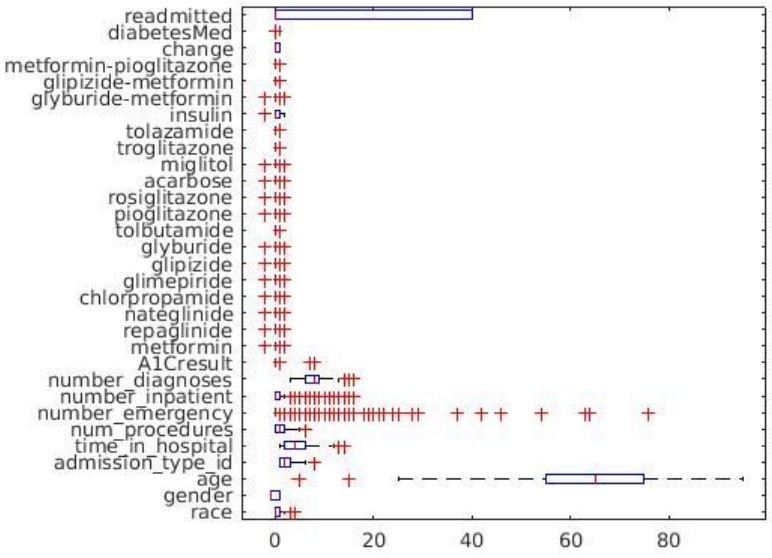
Attribut Race :
Caucasian=0
AfricanAmerican=1
Other=2

Test HbA1c:
None=0

Readmited
No=0
<30=20
>30=40

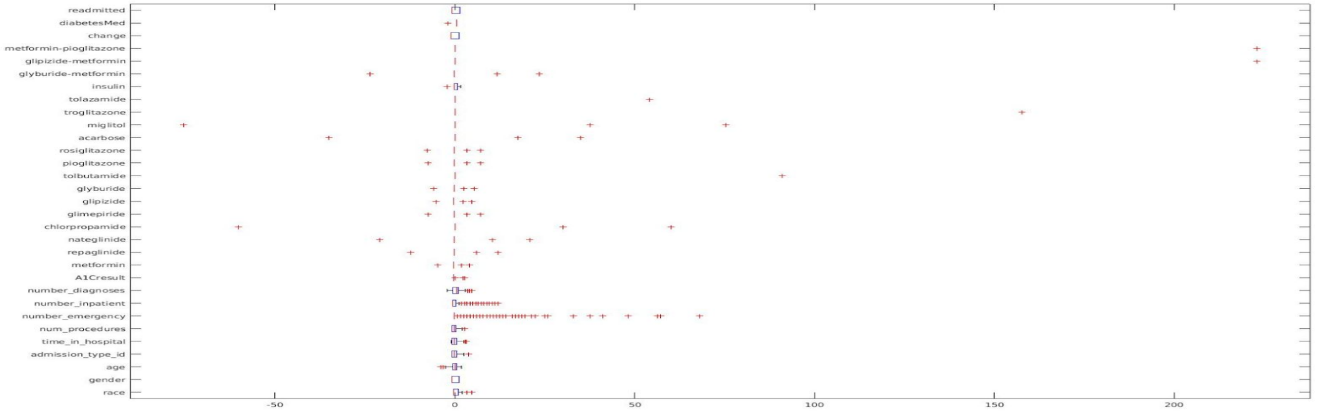
Normalisation des données

Tout d'abord, visualisons les données numérisées :



La numérisation des données a causé des variances arbitraires et inégales des données. Nous décidons de normaliser les données de telle sorte que les attributs soient centrées réduites (moyenne nulle et variance unitaire).

Visualisons les données normalisées



Nous pouvons voir apparaître quelques points extrêmes. Ces points extrêmes peuvent être dûs à un patient “spécial” mais peuvent être aussi dûs au fait que les laboratoires ne réalisent pas de la même manière leur test.

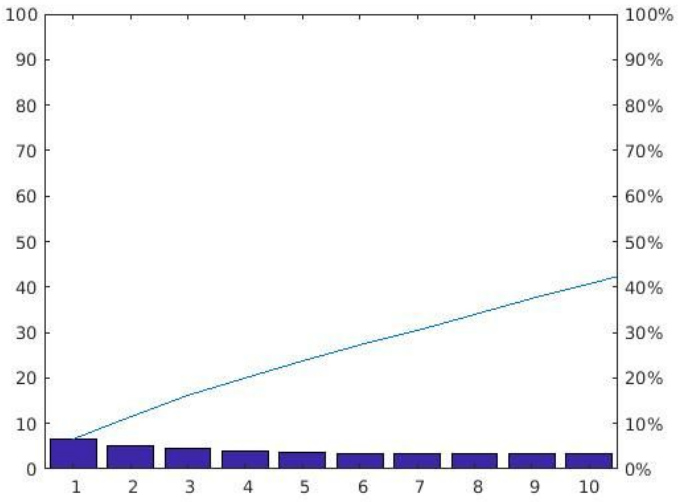
Analyse de PCA

Nous effectuons alors le PCA sur ces données en utilisant la pondération par rapport à la variance.
Voici les composantes des 10 premières variables :

```
c3 =

   -0.0049   -0.1381   -0.2046   -0.1696   -0.1768   -0.0004    0.0387   -0.0481    0.0076    0.1418
    0.0138   -0.0658    0.0592   -0.2505    0.2914   -0.0814    0.0157    0.0860   -0.0024   -0.1490
   -0.1745    4.9276    6.8730    6.4860   -1.4895    0.2771    1.1939   -0.4176    0.3627    0.2472
   -0.0717   -0.0939    0.2580   -0.1030    0.9440    0.1169   -0.0984    0.1219    0.0413    0.0821
    0.4986    1.1598    0.6092   -1.0560   -0.5573   -0.4359    0.1316   -0.0252   -0.1819    0.1412
    0.0153    0.1913    0.5078   -0.9458    0.5019   -0.1299   -0.0231    0.0063   -0.0979   -0.1776
    0.0886    0.2572   -0.5463    0.0239    0.2348   -0.0149   -0.0219   -0.0336   -0.0552   -0.0757
    0.0732    0.4994   -0.6194    0.0308    0.1995   -0.0984    0.0480   -0.0025   -0.0596   -0.0549
    0.2114    1.0156    0.3076   -0.1093   -0.2719   -0.0893   -0.0298    0.0242   -0.0183    0.0459
    0.4158   -0.3365   -0.2489   -1.1244   -0.7779    0.0377   -0.3708    0.2505    0.0787    0.2621
    0.1494   -0.1140    0.0031    0.0597    0.0202   -0.0553   -0.0270    0.0429   -0.0484    0.0315
    0.0181    0.0172    0.0070    0.0008   -0.0223    0.0183   -0.0231    0.0192    0.0269    0.0509
    0.0076    0.0022    0.0013    0.0064   -0.0076    0.0196   -0.0139    0.0041   -0.0149    0.0006
   -0.0001    0.0007    0.0009    0.0002    0.0013   -0.0005    0.0013    0.0003    0.0001   -0.0012
    0.0399    0.0018    0.0104    0.0135   -0.0232    0.0878   -0.1137    0.0220   -0.0272   -0.1505
    0.0725   -0.0271    0.0225   -0.0015    0.0235    0.0964    0.2684    0.0297    0.1088    0.0420
    0.0548   -0.0353    0.0332    0.0698    0.0388   -0.2051   -0.0500   -0.0454   -0.1055    0.0713
   -0.0000    0.0001    0.0003    0.0003   -0.0005   -0.0005    0.0001   -0.0003   -0.0004    0.0001
    0.0506   -0.0134    0.0217    0.0026    0.0457    0.1196    0.0005   -0.1121   -0.0976    0.0471
    0.0573   -0.0243    0.0117    0.0341    0.0168   -0.0604   -0.0288    0.0986    0.1059   -0.0481
    0.0028   -0.0001    0.0015    0.0017    0.0023    0.0106   -0.0003    0.0050    0.0120   -0.0021
    0.0006    0.0002    0.0007    0.0004   -0.0012    0.0059   -0.0044    0.0035    0.0011   -0.0013
    0.0000    0.0000    0.0000   -0.0000    0.0002   -0.0001   -0.0004    0.0008    0.0007    0.0004
   -0.0000   -0.0005    0.0003    0.0007    0.0013    0.0002   -0.0000    0.0005    0.0006   -0.0020
    0.0714   -0.0498   -0.0215   -0.0141   -0.1528   -0.1936    0.4110   -0.2421   -0.1142   -0.6483
    0.0035   -0.0002    0.0004   -0.0025    0.0042   -0.0005   -0.0185   -0.0430    0.0403   -0.0005
    0.0000    0.0000    0.0001   -0.0000    0.0001   -0.0005   -0.0007   -0.0022    0.0018   -0.0002
   -0.0000   -0.0000    0.0000   -0.0000    0.0000    0.0005    0.0001   -0.0007   -0.0004   -0.0000
    0.2837   -0.0150    0.0044   -0.0006    0.0173    0.0166   -0.0251    0.0019    0.0074    0.0446
    0.2260   -0.0221   -0.0060    0.0143    0.0022   -0.0091    0.0468   -0.0346   -0.0062   -0.0403
    1.8424    5.5450   -3.9180    1.7406    3.9577    1.2448   -0.1498    1.4300    1.6234    1.4736
```

La première variable du PCA (première colonne) a de grand coefficient ligne 5, 10 et 31. Elle décrit donc principalement les variables n°5,10 et 31 soit : l'âge, le test d'Hb1ac et le taux de réadmission.
La deuxième variable du PCA (deuxième colonne) a de grand coefficient ligne 3, 5, 9 et 31. Elle décrit donc principalement les variables n°3, 5, 9 et 31 soit : l'âge, le temps à l'hôpital, le nombre de diagnostic et et le taux de réadmission.

On peut afficher le pareto du PCA pour connaître la quantité d'information par variance du PCA. On peut voir quelques variables du PCA ne suffisent pas à décrire les données. Par exemple, les 10 premières variables du PCA ne représentent qu'environ 40% des données.



Nous pouvons visualiser la répartition des données selon les premières variables du PCA. Nous obtenous les figure suivantes :

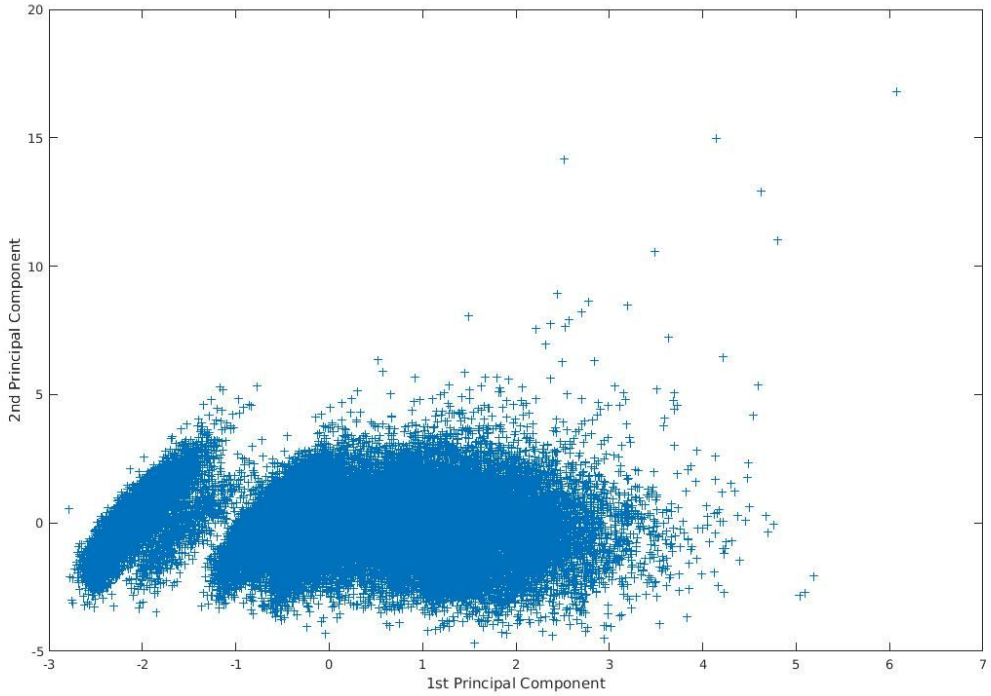


figure 1 : 1ère et 2ème composante du PCA

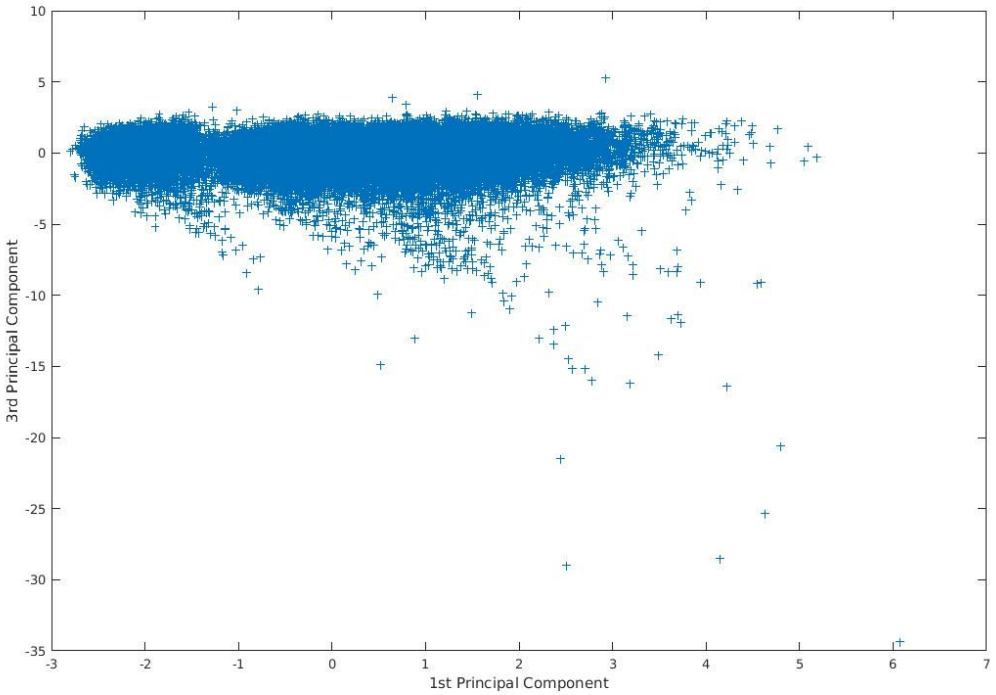


figure 2 : 1ère et 3ème composante du PCA

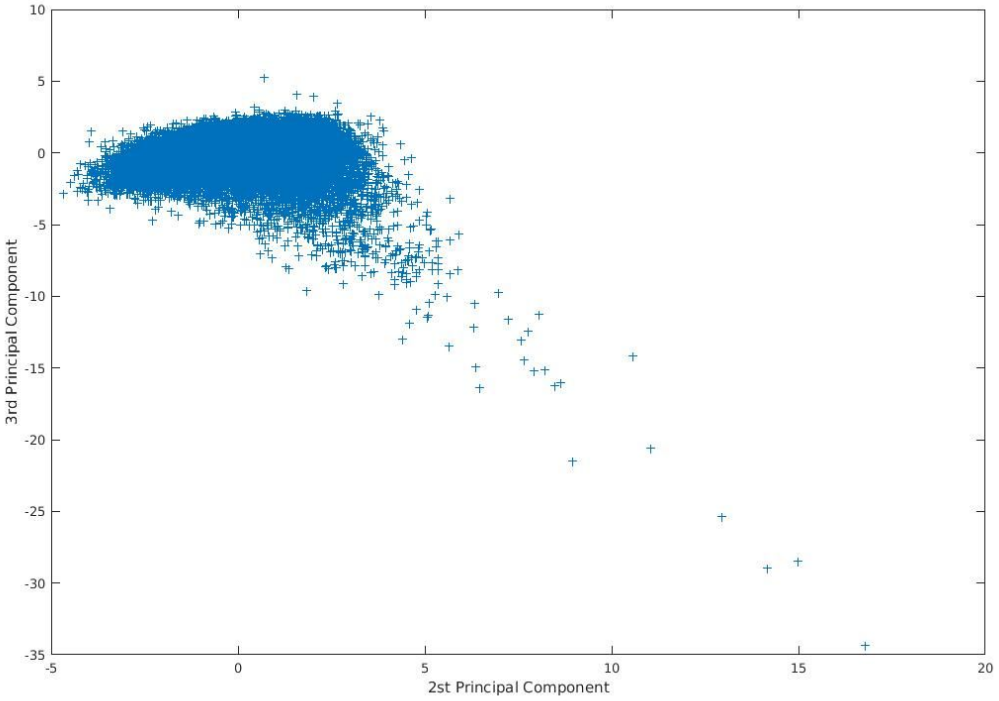


figure 3 : 2ème et 3ème composantes du PCA

Bien que 2 variables ne représentent pas l'ensemble des informations, il semblerait que les données se scindent en 3 groupes :

- un groupe significative que l'on peut voir en bas au milieu de la figure 1
- un groupe que l'on peut voir en bas à gauche de la figure 1
- des valeurs extrêmes

Recherche des variables qui peuvent être corrélées avec la réadmission

Tout d'abord, nous effectuons la normalisation sur chacun des variables, de sorte que leur moyenne est 0 et sa variance et 1. Puis nous utilisons PCA pour obtenir des composants principaux, où nous allons travailler dessus. Et nous avons utilisé 28 composants principaux qui représentent 90% de l'information pour construire l'espace.
Ensuite, dans la dimension du premier et deuxième composants de PCA, nous appliquons Kmeans sur tous les 34 variables avec k=3 car il existe 3 groupes de réadmission. Après avoir affiché la répartition de tous les exemples （*figure 4-plan de Kmeans*) attribués à 3 groupes et représentés avec 3 différentes couleurs, nous voulons visualiser le classement de réadmission (*figure 4- répartition de réadmission*) de tous les exemples. Pour cela, nous appliquons Kmeans seulement sur la réadmission, mais nous les visualisons toujours sur la figure 4.
Afin de chercher quelles variables sont corrélées avec la réadmission,de la même manière, nous affichons la répartition de k-means de chaque variable sur le plan de Kmeans en les divisant en 3 trois groupes(*figure 5*). Puis nous comparons la répartition de chaque variable et celle de la réadmission, afin de choisir lesquelles sont plus ressemblantes. Plus la répartition de k-means d'un variable ressemble à celle de réadmission , plus la variable est corrélée avec la réadmission.

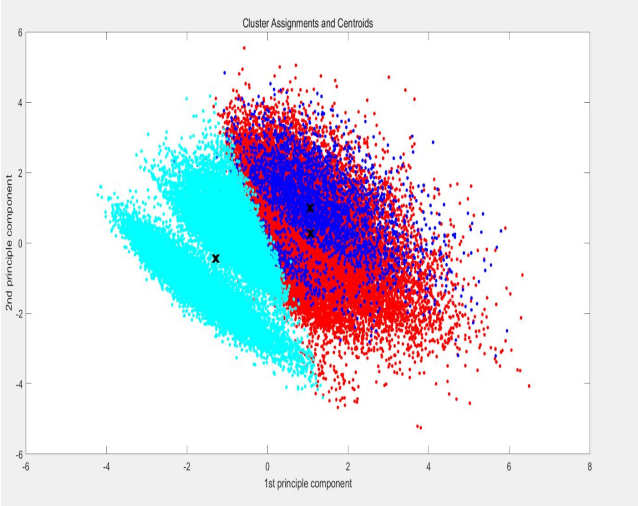


figure 4-plan de k-means

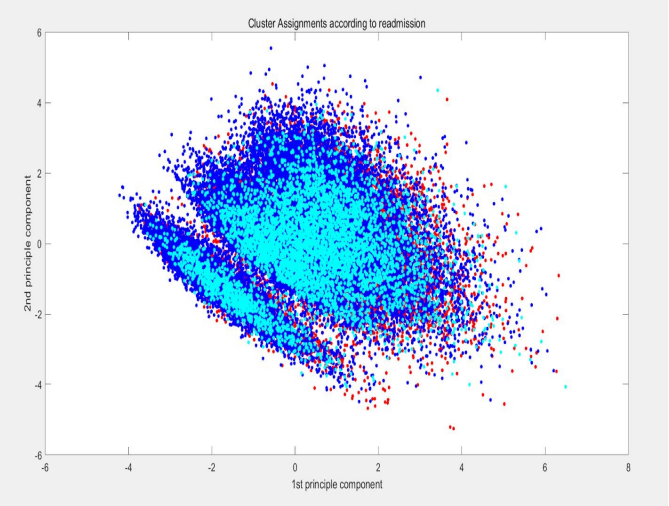


figure 4- répartition de réadmission

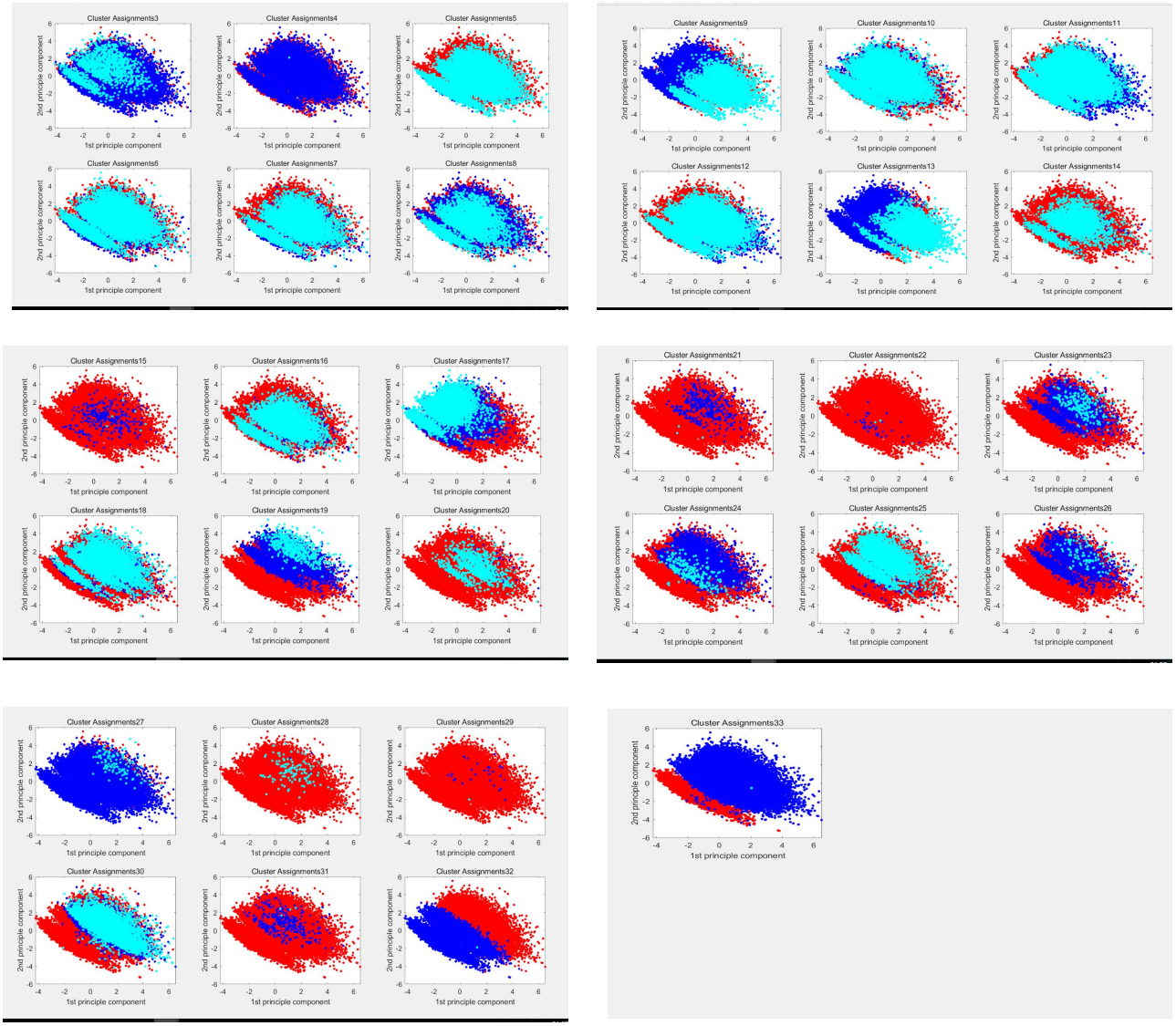


figure 5

Ici, on trouve que les répartitions des variables de colonnes *3,*6,7,9,10,11,12,13,16 sont assez similaires à celle de réadmission.

Nous effectuons le même traitement en dimension du deuxième et troisième composante de PCA et celle du premier composant et troisième composant.(figures voir annexe 1)

Résultat

En analysant les trois cas ci-dessus, on trouve que les répartitions des variables qui correspondent aux colonnes 6,7,9,10,11,12,13 sont assez similaires à celle de répartition. Ils correspondent aux variables admission_type_id, discharge_disposition_id, time_in_hospital, medical_specialty, num_lab_procedures, num_procedures, num_medications

Pourtant, Il apparaît que le résultat de k-means ne correspond pas au résultat de PCA, qui montre l'âge et Hb1Ac sont éléments de composants principaux. Le choix de faire 3 groupes ne semble pas applicable dans tous les cas. De plus, en ce qui concerne la thèse originale de cette base de données, elle trouve aussi que 'l'âge' et 'Hb1Ac test result' influencent la réadmission. Nous décidons de faire l'hypothèse sur ces deux variables et refaire le traitement.

Hypothèse

hypothèse 1

âge: selon le résultat de la thèse, l'âge comporte 3 intervalles [0,30], [30,60] et [60,100], qui ont 3 probabilités de réadmission distinctes. Nous posons l'hypothèse que plus âgé, plus un patient est potentiel d'avoir une réadmission.

hypothèse 2

Hb1Ac: L'action de prendre mesure de Hb1Ac diminue la possibilité de réadmission, cela est plus significative pour les patients patients d'avoir diabète dans première diagnosis. Cependant la valeur de Hb1Ac n'est pas corrélé avec la réadmission.
Remarque: Ici une réadmission ne comporte que celle dans moins de 30 jours.

Matrice de confusion

La matrice de confusion est un outil servant à mesurer la qualité d'un système de classification. Dans la méthode supervisée, la matrice de confusion permet de nous indiquer à quel niveau les prédictions sont correspondantes avec les classes réelles, autrement dit la robustesse de notre algorithme.
Mais ici, on s'intéresse pas aux valeurs diagonales, on utilise la matrice de confusion autrement. On divise tous les exemples en deux ou trois ou quatre groupes en utilisant une seule variable à la fois pour faire le K-means. K-means nous retourne un vecteur avec la classification de chaque exemple. Et puis on construit la matrice de confusion selon ces vecteurs. Chaque ligne de la matrice représente le nombre d'occurrences d'une classe

déterminée selon la réadmission, tandis que chaque colonne représente le nombre d'occurrences d'une classe déterminée selon un autre variable telle que l'âge, le Hb1Ac, etc.

Retraitement

retraitement 1

On peut s'intéresser aux taux de réadmission selon les tranches d'âge. Ainsi on forme un tableau de proportion ci-dessous.

Les lignes correspondent à l'age : [0-10[, [10-20[, …, [90-100[
Les colonnes correspond à la réadmission: No, <30, >30.

```
MatriceConfusion10 =

         124           3          25
         335          31         187
         592         150         369
        1454         293         845
        3508         719        2270
        6282        1132        4038
        7457        1622        5225
        8676        2061        6310
        5235        1276        3721
         964         167         464
```

```
MatriceConfusion10pourcentage =

   81.5789    1.9737   16.4474
   60.5787    5.6058   33.8156
   53.2853   13.5014   33.2133
   56.0957   11.3040   32.6003
   53.9942   11.0666   34.9392
   54.8550    9.8847   35.2602
   52.1323   11.3395   36.5282
   50.8946   12.0901   37.0153
   51.1630   12.4707   36.3663
   60.4389   10.4702   29.0909
```

Puisque les populations selon l'âge n'est pas uniforme, on s'intéresse plutôt à la répartition de la réadmission en pourcentage pour chaque intervalle d'âge.

On peut alors remarquer la tendance que les jeunes sont moins souvent réadmis.
Ainsi si peut former 3 groupes d'âge : [0-30[, [30-60[et [60-100[, on obtient cette fois-ci :

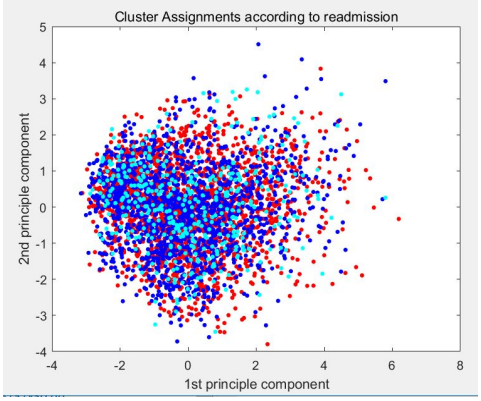
```
MatriceConfusion3 =

        1051         184         581
       11244        2144        7153
       22332        5126       15720


MatriceConfusion3pourcentage =

   57.8744   10.1322   31.9934
   54.7393   10.4377   34.8230
   51.7208   11.8718   36.4074
```

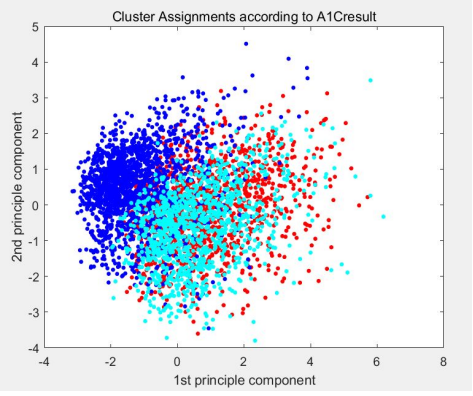
On peut alors voir que taux de réadmission est significativement influé par l'âge. En effet, être jeune favorise la non-réadmission alors qu'être agé favorise une réadmission fréquente (plus de 30 fois).

Retraitement 2

1. Patients diabètiques

En analysant les sous groupes de la mesure de Hb1Ac, nous trouvons qu'il existe parfaitement 3 niveaux différents. Donc nous avons éliminé des exemples sans mesure de Hb1Ac, qui compte presque la moitié. Puis nous recommençons le traitement.





Premièrement, la répartition de k-means et celle de valeur de Hb1Ac sont décorrélées, il montre qualitativement que les 3 trois niveaux de Hb1Ac ne signifient pas la possibilité de réadmission

Deuxièmement, nous utilisons la matrice de confusion pour avoir une examination quantitative. Nous avon extrait des patients diabétiques dans le premier diagnostic , nous utilisons la matrice de confusion pour avoir une examination quantitative. La matrice obtenue est suivante:

5127	418	1815	260
874	55	170	38
0	0	0	0
0	0	0	0

l'information d'Excel qui montre le nombre de patients de chaque groupe de réadmission est suivante: Dans première figure, 0 signifie pas de réadmission, 1 signifie une réadmission dans moins de 30 jours. Dans deuxième figure, 0 signifie pas de mesure de Hb1Ac, 5 signifie un niveau de Hb1Ac normal, 7 signifie un niveau de Hb1Ac entre 7% et 8%, 8 signifie un niveau de Hb1Ac dépassant 8%.

En combinant la matrice et l'information de nombre, nous pouvons déterminer que dans la matrice de confusion, la première ligne représente pas de réadmission, la deuxième ligne représente une réadmission dans moins de 30 jours. De même, la première colonne représente pas de mesure de Hb1Ac, la deuxième colonne représente un niveau de Hb1Ac normal, la troisième colonne représente un niveau de Hb1Ac dépassant 8%, la quatrième colonne représente un niveau de Hb1Ac entre 7% et 8%.

Alors, la possibilité de réadmission pour un patient avec un niveau de Hb1Ac normal est:55/(418+55)=11.63%. Également, pour un patient avec un niveau de Hb1Ac entre 7% et 8% cela égale 38/(260+38)=12.75%, pour un patient avec un niveau de Hb1Ac,cela égale 170/(1815+170)=8.56%. D'ici on peut voir que le taux de réadmission n'est pas corrélé avec la valeur de Hb1Ac.

De la même manière, pour des patients qui n'ont pas eu de mesure de Hb1Ac, le taux de réadmission égale 874/(5127+874)=14.56%, pour des patients d'avoir eu la mesure, le taux égale (55+170+38)/(418+55+1815+170+260+38)=9,54%. Donc cela vérifie l'hypothèse que l'action de prendre mesure ou pas de Hb1Ac suscite une différence significative de réadmission.

2. Patients malade de l'appareil circulatoire

Afin de comparer le cas d'autres malades, nous montrons ici les patients malade circulatoire, puisqu'ils sont plus nombreux.

La matrice de confusion:

2901	277	165	142
22233	2152	1396	1171
0	0	0	0
0	0	0	0

l'information dans Excel:

les résultats du taux de réadmission est suivant:
Hb1Ac normal : 10.57% Hb1Ac entre 7% et 8%: 10.8% Hb1Ac dépassant 8%:11.4%
ayant mesuré l'Hb1Ac: 11.01% pas de mesure de Hb1Ac: 11.54%

Alors que le taux reste élevé, il montre que l'influence de mesure de Hb1Ac n'est pas importante ici.

Problème et réflexion

1. Il y a deux types de variable dans notre base de données, les variables quantitatives et les variables qualitatives. Le PCA et le K-means ne s'adaptent que pour des variables quantitatives. Quand on fait le K-means avec ces deux types de variables en même temps, trouver une distance qui satisfasse les deux types de variable est complexe. Pour les variables qualitatives, il faut utiliser la distance de Hamming qui est en fait pour des données binaires, mais pour les données données quantitatives, il faut utiliser la distance euclidienne. Ceci peut expliquer pourquoi on n'a pas obtenu des résultats satisfaisants avec le K-means.
2. Notre base de données comporte un grand nombre d'individus et de nombreux attributs complexes ce qui complique l'analyse et l'exploitation des résultats. De plus, nos données comportent des valeurs extrêmes. Elles peuvent nuire à l'efficacité des méthodes comme la PCA ou le k-means d'autant plus que nous avons utilisé la distance euclidienne.