APPRENTISSAGE STATISTIQUE

Sujet 6

Construction d'un modèle de prédiction du cancer du sein à partir de biomarqueurs

BOULAHIA AHMED KHALIL

TABLE DES MATIERES

INTRODUCTION

L'étude réalisée dans le cadre de ce projet consiste à la construction d'un modèle de prédiction du cancer du sein à partir de biomarqueurs. Pouvoir prédire pour chaque femme la présence ou pas du cancer du sein permettrait de faire un dépistage rapide et ainsi faire une prise en charge rapide. Celui-ci peut être diagnostiqué par plusieurs tests différents à savoir : l'IRM, biopsie, échographie mammaire …

ENVIRONNEMENT D'ETUDE

Les patientes observées sont plus ou moins âgées, et plusieurs facteurs sont pris en compte dans le jeu de données pour la prédiction du cancer. Le taux de glucose, le taux d'insuline, l'indice de masse corporelle (BMI), HOMA, Leptin, Adiponectin, Resistin et MCP.1 sont ainsi enregistrés et nous serviront de variables explicatives pour la classification

OBJECTIFS

Le but de cette étude est de pouvoir construire un modèle de classification afin de prédire la présence ou non d'un cancer. Pour cela nous avons les résultats de 116 patientes indiquant la présence ou non du cancer. Plus précisément notre objectif est de pouvoir classer les patientes afin de pouvoir faire une prédiction pour cela on va utiliser 3 méthodes à savoir : l'arbre de décision, Knn et SVM.

I. ETUDES PREALABLES

1. Préparation de la base de données

La base de données comporte initialement 116 observations de 10 variables explicatives, la dernière variable explicative étant la variable de sortie (présence ou pas du cancer du sein), donc on est dans un problème de classification supervisée.

a- Étude des correlations

Nous avons voulu étudier les différentes corrélations entre les variables explicatives pour diminuer l'espace de travail.
Nous avons représenté dans la figure 1 la matrice des corrélations pour l'ensemble des variables, (on utilise la couleur bleue pour une forte corrélation et la couleur rouge pour une faible corrélation).

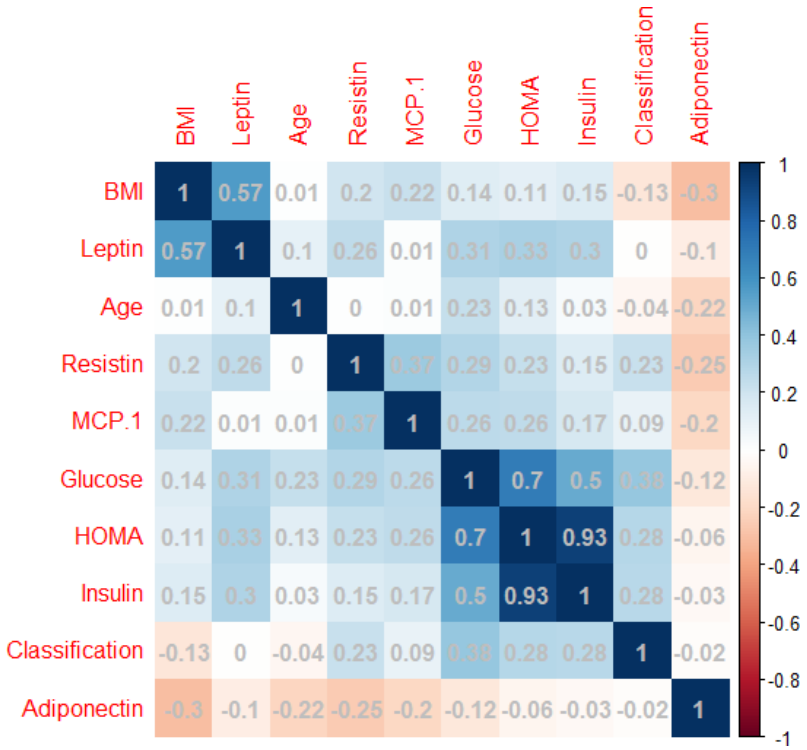


Figure 1 – Représentation graphique des corrélations

On peut constater que les variables Insuline et HOMA, Glucose et HOMA sont corrélées entre elles, Cependant, on peut voir que la variable Classification représentant le fait d'avoir un cancer du sein ou non n'a aucune corrélation avec les autres variables.

b- Selection des variables

Afin de réduire la complexité du modèle recherché, on va commencer par éliminer les variables qui ne sont pas significatives. Pour cela on va utiliser l'algorithme boruta, qui est basé sur le principe de sélection de variables des arbres de décision afin de mesurer l'impact de chaque variable sur la capacité de classification. L'algorithme attribue ainsi à chaque variable explicative un score d'importance.

L'algorithme boruta procède comme suit :

1. Pour chaque variable explicative :

   - Tout d'abord, il duplique le jeu de données et mélange aléatoirement les valeurs de chaque colonne.
   - Ensuite il forme un classificateur, tel qu'un classificateur de foret aléatoire, sur le jeu de données pour prédire la variable d'intérêt.
   - L'importance de la variable est calculée en mesurant la perte moyenne de précision (*Mean Accuracy Loss*) sur l'ensemble des arbres utilisés par rapport au modèle original.
   -

2. Pour chaque variable copiée, on calcule le *Z*-score qui est la *Mean Accuracy Loss* divisée par son écart-type.

3. La variable copiée ayant le *Z*-score le plus élevé est utilisée comme référence : toute variable qui a une importance significativement plus élevée est considérée comme décisive pour la prédiction de la variable d'intérêt. Si l'importance d'une variable est plus faible, elle est considérée comme non décisive. Sinon, son sort reste à déterminer.

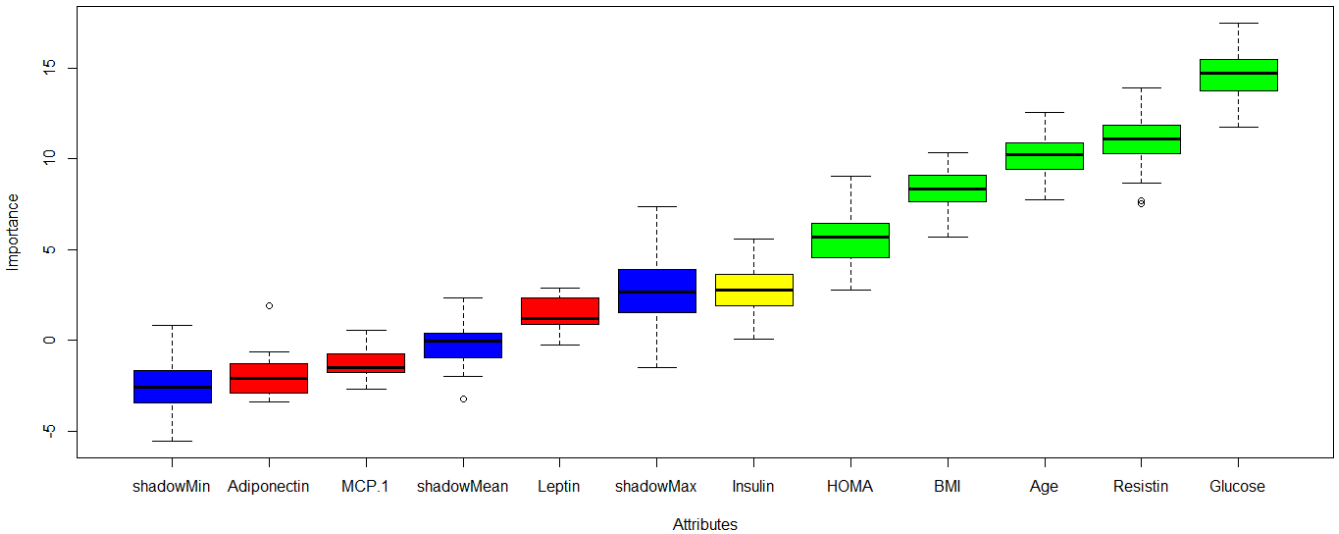


Figure 2 – Importance des variables lors de l'analyse avec Boruta

Résultats obtenus par R :

```
Boruta performed 99 iterations in 4.167812 secs.
 5 attributes confirmed important: Age, BMI, Glucose, HOMA, Resistin;
3 attributes confirmed unimportant: Adiponectin, Leptin, MCP.1;
1 tentative attributes left: Insulin;
```

La figure 2 nous indique qu'un certain nombre de variables ont des scores d'importance élevés, elles sont représentées en vert : (HOMA, BMI, Age, Resistin et Glucose)

Les variables représentées en rouge ont une importance peu élevée : (Adiponectin, MCP.1 et Leptin)

La variable représentée en jaune a une importance indéterminée (Insulin), nous allons la maintenir dans l'échantillon.

Nous allons donc effectuer l'étude sans les variables explicatives d'importances limitée selon cet algorithme.

Après cette sélection de variables, la matrice de corrélations des variables restants est représentée sur la figure 3.

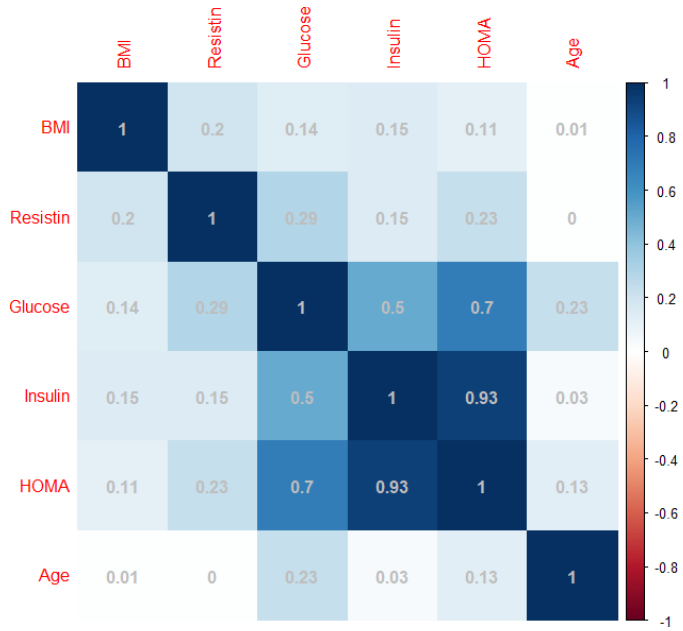


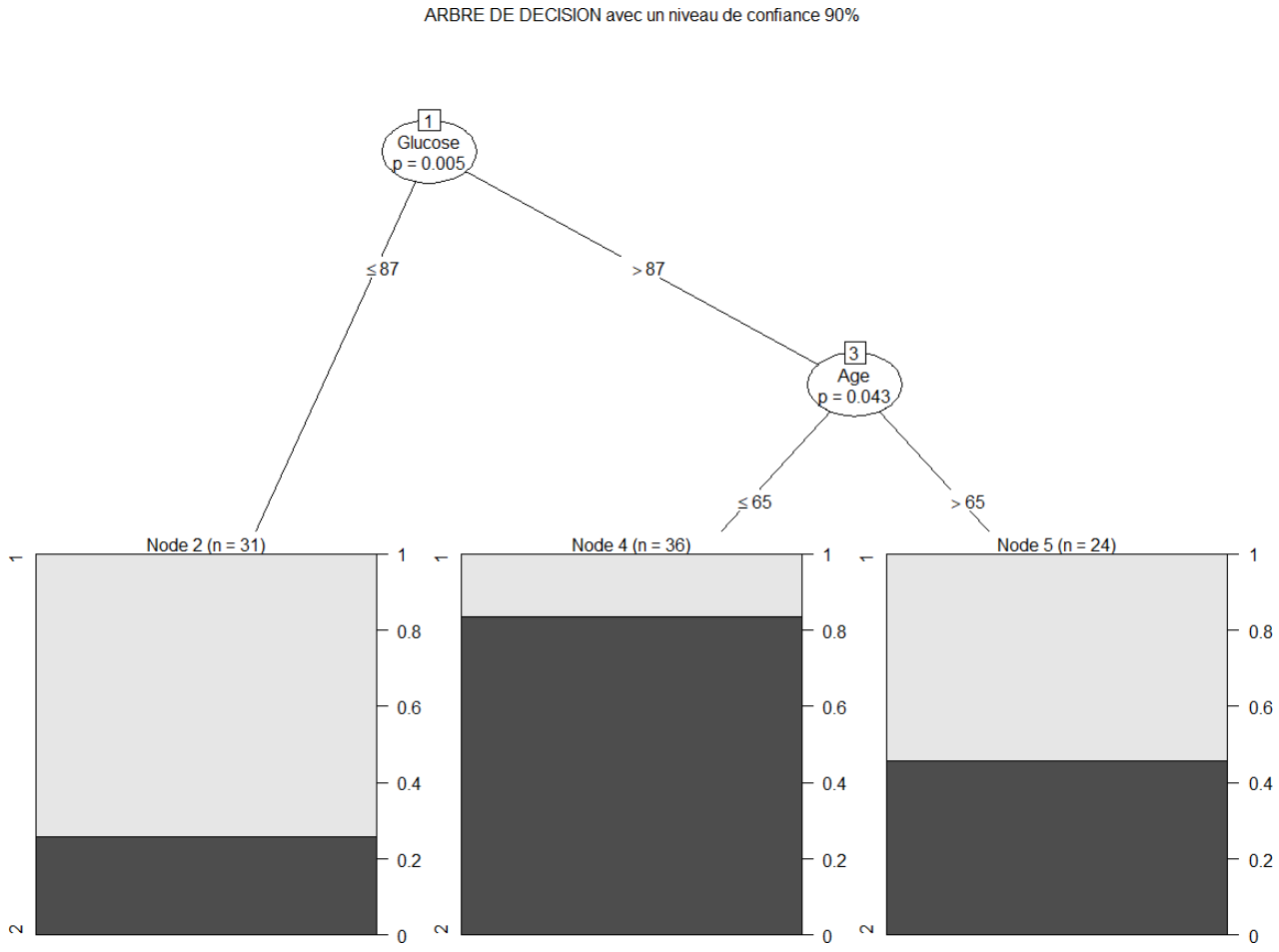
Figure 3 – Représentation graphique des corrélations entre les variables sélectionnées

II. METHODES DE CLASSIFICATION

1. l'Arbre de décision

A partir du notre base de données, nous créons un échantillon d'apprentissage correspondant à 70% de l'échantillon initial (91 observations) et un échantillon de test qui contient les 30% restants (25 observations).

L'application de l'arbre de décision sur l'échantillon d'apprentissage à un intervalle de confiance à 90% et à une limitation minimale de 10 individus par nœuds nous donne une classification en 3 groupes qui se présente comme suit :



On constate que la probabilité d'être en bonne santé pour un individu ayant un taux de glucose inférieure à 87 est d'environ 75%, par contre la probabilité d'être en bonne santé pour un individus présentant à la fois un taux de glucose supérieur à 87 et un âge inférieur à 65 est environ 17%.

La matrice de confusion d'une prédiction :

	Predicted 1	Predicted 2
True 1	36	19
True 2	6	30

Idéalement, seules les valeurs de la diagonale devraient être non nulles.
Dans cet exemple, 66 observations ont été correctement prédites. 19 patientes ont été prédites comme ayant le cancer alors qu'elles ne l'ont pas, et 6 patientes atteintes n'ont pas été détectées. 30 patientes atteintes ont été correctement classifiées.

2. K-PLUS PROCHES VOISINS

La taille de l'échantillon étant petite, nous allons donc essayer la méthode des k-plus proches voisins qui fait partir des méthodes locales

1. Description de la méthode et du principe utilisé

a) Méthode Naïve
Ici on effectue des analyses sans accorder une importance particulière a une variable d'intérêt en utilisant la fonction d'erreur si après

$$R(h) = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \mathbb{1}_{\{Y_i \neq h(X_i)\}}$$

Les résultats de cette analyse ne sont pas très concluants car nous obtenons un accuracy de 57,80%

Alors nous avons décidés de passer a un modèle un peu plus rigoureux qui accorde une importance particulière a la bonne classification de patients malade

b) Méthode Améliorée

L'étude étant faite sur des patients seins et des patients malades, il nous a été important de mettre l'accent sur la bonne classification
lorsque le patient est malade et donc nous sommes parvenus a ce cette fonction d'erreur :

$$R(h) = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left(\alpha \mathbb{1}_{\{Y_i=1,\ h(X_i)=0\}} + \gamma \mathbb{1}_{\{Y_i=0,\ h(X_i)=1\}} \right)$$

Ou alpha et gamma représentent l'importance accordée à chaque type d'erreur. Dans notre cas, une importance bien plus grande est accordée à la bonne classification des patients malade

Dans la matrice de confusion, cela revient à minimiser le nombre en bas à gauche et maximiser celui en haut à droite

c) Implémentation de l'algorithme des k-plus proches voisins

Pour arriver au résultat plus performant, nous avons passé les étapes ci- après

Vu la taille de l'échantillon petite, nous avons pris 70% de l'échantillon comme étant notre base en effectuant un tirage aléatoire et sans remise ainsi 30% servent de base pour le test

Avec cette méthode nous sommes arrivés a un accuracy de 83 ,33% ce qui était normal a cause de la disproportionnalité entre les 2 variables de sorties

En effectuant les prédictions, on observe une précision faible et un recall grand ce qui signifie beaucoup de retour ne seront pas correct d'où il a fallu rétablir un certain équilibre entre les variables d'intérêts

Apres ce changement nous avons eu les sorties suivantes :

	precision	recall	f1-score	support
1	0.67	1.00	0.80	6
2	1.00	0.75	0.86	12
micro avg	0.83	0.83	0.83	18
macro avg	0.83	0.88	0.83	18
weighted avg	0.89	0.83	0.84	18

Le recall mesure la complétude du model alors ici p our le label 2, seul 75% était vrai

D'après le graphe ci-dessous qui trace l'erreur en fonction du k, on voit que le k qui minimise l'erreur est 5 ou 6

donc le meilleur score est obtenu pour 5 ou 6 voisins :

```
Text(0, 0.5, 'Mean Error')
```

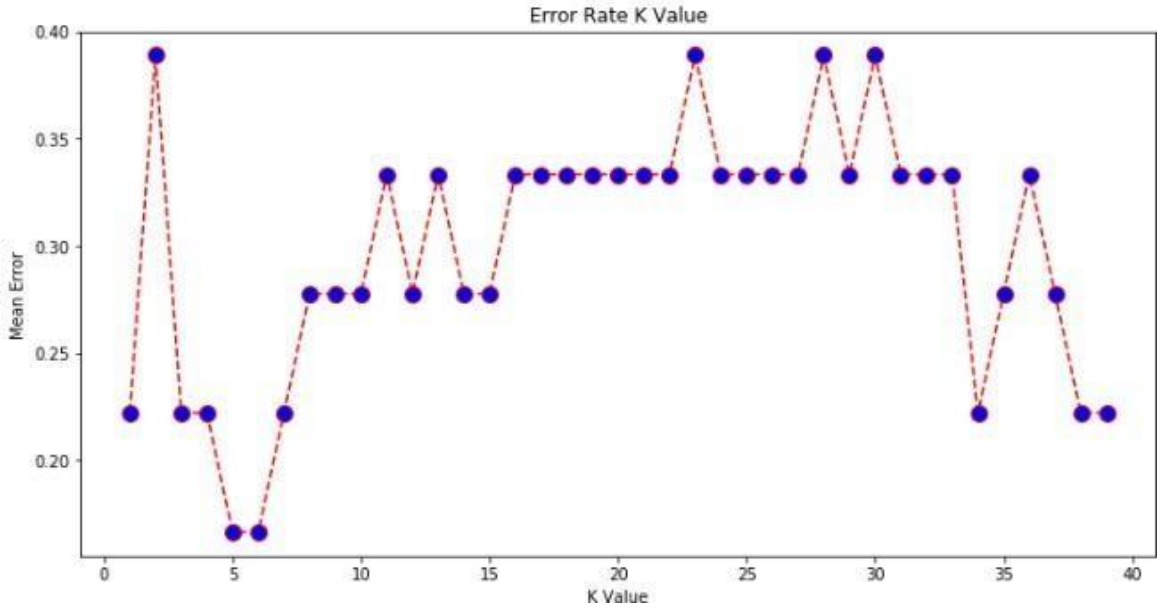


Figure 4 : Erreurs des K.ppv en fonction des k

On constate que cette méthode d'évaluation est meilleure pour prédire les patients seins car ceux si sont majoritaire dans l'échantillon. Mais beaucoup de patients malades sont prédits seins d'où notre migration vers la méthode des SVM qui est un peu plus performante que les k plus proches voisins

3. LA METHODE DE SVM

1- Description de la méthode

L'objectif de la méthode est de trouver un hyper plan séparateur permettant de séparer les données.

En effet, les SVM offrent une accuracy meilleure que celles des autres méthodes de classifications.

L'objectif principal est de séparer le jeu de données donné de la meilleure façon possible. La distance entre les points les plus proches est appelée la marge. L'objectif est de sélectionner un hyperplan avec la marge maximale possible entre les vecteurs de support dans l'ensemble de données donné. SVM recherche l'hyperplan marginal maximal dans les étapes suivantes

2- Générez des hyperplans qui séparent les classes de la meilleure façon. Figure de gauche représentant trois hyperplans noirs, bleu et orange. Ici, le bleu et l'orange. Bleu et orange. Ici, le bleu et l'orange ont une erreur declassement plus élevée, mais le noir sépare correctement les deux classes
3- Sélectionner l'hyperplan ayant la séparation maximale des points parmi les points les plus proches

Il existe plusieurs noyaux utilisables dans les SVM mais dans notre cas, le noyau utilisé est le noyau linéaire. Ceci est expliqué par son accuracy élevé et équilibré pour chacune des classes

Le noyau linéaire peut être utilisé comme produit scalaire normal sous deux observations données.

4- Implémentation de l'algorithme des k-plus proches voisins

Ici tout comme dans les autres méthodes, on tire aléatoirement 70 % datasheet pour faire l'entrainement et 30 % restant servent de d'échantillons de test

On effectue donc l'algorithme des svm sur les données d'entrainement. Apres cette entrainement, nous obtenons un accuracy de 83 % qui est sensiblement égal a celui trouvé dans les kppv ceci s'explique par le fait que la dimension est réduite donc les svm se comporteront presque comme les kppv

Mais un avantage des svm est que nous n'avons pas besoin comme dans les kppv d'accorder plus d'importance a une variable

	precision	recall	f1-score	support
1	0.89	0.80	0.84	10
2	0.78	0.88	0.82	8
micro avg	0.83	0.83	0.83	18
macro avg	0.83	0.84	0.83	18
weighted avg	0.84	0.83	0.83	18

Les résultats obtenus sont donnés par le tableau ci-dessus.

Il montre un une bonne précision pour chacune des variables ainsi que de bons recalls. Cela signifie que 80% des prédictions sur les malades sont justes et 88 % des prédictions sur les non malades sont justes

Ceci est justifiés par le fait que nous avons une proportion plus grande d'individus non malades.

Le svm offre l'avantage d'avoir un meilleur accuracy que les autres mais il n'est pas recommandé en très grande dimension car il est très lent

CONCLUSION

Les données soumises a notre etude etaient des données etiquetées alors il s'agissait d'effectuer une classification supervisée. Pour arriver a nos objectifs qui etaient de creer une modele de classification de données, nous avons utilsés 3 methodes(les arbres de decision, la methode des k plus proches voisins et la methode des SVM)

Le choix de ces methodes a éte volontaire et justifié par la taille reduite des données a etudier

La methode locale qui est celle des kppv nous a donnée un bon accuracy pour la prediction des patients seins mais il arrive parfois que des patients malades soit predits sains ceci est justifié par la dominination des patients seins dans l'etude alors nous avons amelioré l'algorithme en accordant plus d'importance aux patients malades en essayant d'equilibrer l'echantillon d'apprentissage.cette procedure nous a permis d'ameliorer la qualité des resultats afin d'avoir une meilleure classification.

A cause de ces petits problemes de jeu de données nous avons donc pensés a mettre sur pied une methode de classification par les SVM

Ce qui nous a donné les resultats plus concluants ayant un equilibre dans la prediction. Enfin nous

avons poussé la recherche avec une classification par les arbres de decision.

Au terme de notre etude, nous avons observé que la methode des SVM etait meilleure que les autres ceci se justifie par le fait que la methode trouve un hyperplan separateur

pour classifier le jeu de donnés en 2 et la selection est faite avec l'hyperplan optimal

mais cette methode peut poser un probleme pour une taille de donner plus grande car elle est lente en grande dimension.

Reference :

Base de données fournie par : archive.ics.uci.edu
Lien : http://archive.ics.uci.edu/ml/machine-learning-databases/00451/