Institut national des sciences appliquées de Rouen

INSA de Rouen

STPI 2, M8

Étude statistique du trouble du spectre autistique grâce à l'application ASD Tests

Auteurs :
Océane RIGA
oecane.riga@insa- rouen.fr
Kévin GATEL
kevin.gatel@insa- rouen.fr

Enseignant :
Stéphane Canu
stephane.canu@insa- rouen.fr

8 Mai 2020

Table des matières

Introduction

Ce projet a été réalisé dans le cadre de notre deuxième année d'étude au sein de l'INSA Rouen Normandie, plus particulièrement dans le but de mettre en œuvre les compétences théoriques acquises dans la matière M8 Introduction à la maîtrise de données.

Lors d'un appel sur Discord en raison du confinement, nous recherchions une base de données sur le site de l'UCI qui pourrait être intéressante à interpréter. Nous sommes alors tombés sur l'étude d'une application qui détermine si un individu est atteint d'un trouble du comportement autistique ou non en fonction de ses réponses à quelques questions bien précises. Ce sujet a retenu notre attention car, en effet, de nos jours l'informatique tend à se développer dans la médecine. Il s'agit ici d'étudier si une application est réellement capable de détecter un handicap mental chez un individu de manière suffisamment fiable. Autrement dit, si l'informatique peut réellement remplacer l'avis d'un médecin spécialiste en la matière. Nous aimerions également mettre en avant les possibles traits communs que présentent les individus présentant cet handicap avéré dans l'échantillon.

En se servant ici uniquement de l'âge, du genre, des origines, du pays de résidence, des réponses à 10 questions comportementales précises ainsi que du fait que l'individu est né ou non avec la jaunisse, nous avons réalisé différentes méthodes statistiques telles qu'une régression linéaire, une ACP ou encore un test dans le but de répondre aux questions suivantes :

- Une application peut-elle détecter un trouble du spectre autistique de manière fiable ?
- Quels sont les points communs entre les individus atteints d'autisme ?

1. Description des variables

Vous trouverez en annexes un aperçu des données que nous avons étudié. Celles-ci représentent les données obtenues par une application qui propose un test visant à détecter un possible trouble du comportement autistique chez un adulte. L'étude est réalisée sur un échantillon de 704 observations et s'intéresse à 20 variables.

Nous étudierons deux types de variables : des variables quantitatives (plus précisément des variables binaires pour la majorité) ainsi que des variables qualitatives.

Variables qualitatives :

Le genre pourra être traduit sous une forme binaire afin de faciliter son étude, ainsi 0 signifiera femme et 1 sera homme. L'éthnique correspond aux origines de l'individu tandis que pays de résidence correspond à son lieu de vie actuel. Enfin "parenté avec l'individu atteint" indique le lien de parenté de la personne qui répond aux questions pour l'individu étudié par l'algorithme de l'application.

Ces variables étant qualitatives, il n'est pas intéressant d'en calculer les estimateurs.

Variables quantitatives :

Les variables A1 jusqu'à A10 correspondent aux réponses aux questions suivantes :

- A1 : Je remarque souvent de petits bruits que d'autres ne remarquent pas
- A2 : En général je me concentre plus sur l'ensemble de l'image plutôt que sur des petits détails
- A3 : Je trouve facile d'entreprendre plusieurs activités à la fois
- A4 : S'il y a une interruption je peux très facilement revenir à ce que je faisais
- A5 : Je trouve facile de lire entre ligne lorsqu'on me parle
- A6 :Je reconnais lorsque quelqu'un qui m'écoute s'ennuie
- A7 : Quand je lis une histoire je trouve difficile de déterminer les intentions des personnages
- A8 : J'aime rassembler des informations sur des catégories d'objets (type de voiture, plantes...)
- A9 : Je trouve facile de deviner ce qu'une personne pense ou comment elle se sent en regardant simplement son visage
- A10 : Je trouve difficile de comprendre les intentions des gens

Né avec la jaunisse est là encore une variable binaire où 1 signifie oui et 0 signifie non. Il en est de même pour la variable binaire autisme : 1 signifie qu'un trouble du spectre autistique a été avéré par des médecins tandis que 0 signifie que ce n'est pas le cas. De la même manière, on définit la variable binaire correspondant au fait que l'individu étudié a oui ou non déjà réalisé un test sur l'application. "Appli détecte autisme" indique si, d'après les différentes réponses aux questions, l'application estime que l'individu étudié est atteint (1) ou non (0) d'un trouble du spectre autistique. L'âge est indiqué ici en années. "Résultat obtenu" correspond à la somme des points obtenus en répondant aux questions A1 à A10. Cette variable peut donc prendre toutes les valeurs entières comprises entre 0 et 10 inclus.

	Variable Âge	Variable Résultat obtenu
Moyenne	29.12358	4.875
Médiane	27	4
Ecart type	9,71029	2,50149
Variance	94,28966	6,25747

FIGURE 1.1 – Tableau représentant les statistiques sur les variables quantitatives étudiées

1.1 Autisme

Lors de notre étude, notre but est d'expliquer cette variable, de savoir s'il est possible de la prévoir. 0 indique que l'individu n'est pas atteint d'un trouble du spectre autistique tandis que 1 indique que l'individu étudié est atteint de cet handicap.

Individu atteint	Individu sain
91	613

FIGURE 1.2 – Tableau de répartition des individus selon l'autisme

On peut remarquer ici que cet handicap est plutôt bien représenté dans un échantillon de taille suffisamment importante puisque 13% de l'échantillon présente un trouble du comportement autistique.

1.2 Âge

L'individu le plus jeune a 17 ans tandis que le plus âgé a 64 ans. Voici un diagramme indiquant la répartition, selon l'âge, des individus atteint d'un trouble du spectre autistique.

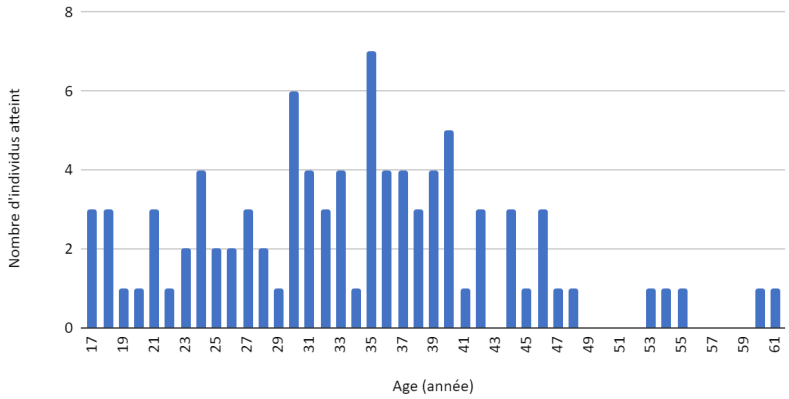

FIGURE 1.3 – Graphique représentant le nombre d'individus atteints en fonction de l'âge

On remarque sur ce diagramme que la tranche d'âge la plus touchée est celle des individus ayant une trentaine d'années ainsi que les plus jeunes. On remarque peu de cas à partir de 50 ans, on peut

donc se demander si la durée de vie est raccourcie pour les personnes atteintes de ce syndrome.

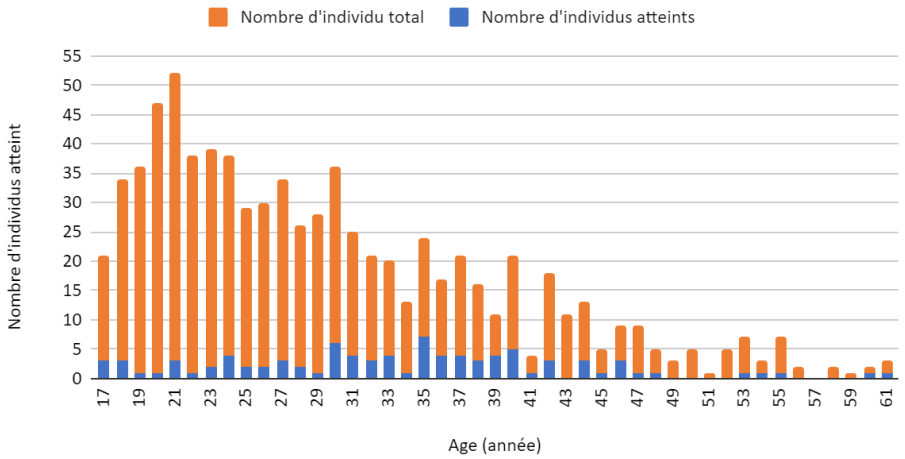

FIGURE 1.4 – Graphique représentant le nombre d'individus atteints et total selon leur age

Toutefois, ce diagramme indique que la tranche d'âge à partir de 48 ans est tout simplement très peu représentée dans l'échantillon d'où le manque de données sur ces individus. On notera que deux individus n'ont pas indiqué leur âge, cependant ce ne sont pas des individus atteints cela ne pose donc pas énormément de problème.

Ci-dessous, nous avons représenté la boite à moustache de la variable âge afin de visualiser la médiane ainsi que la distance interquartile. Les valeurs supérieure s à 56 ans sont hors épure, ces points sont ainsi suspects (très peu de personnes âgées parmis nos individus).

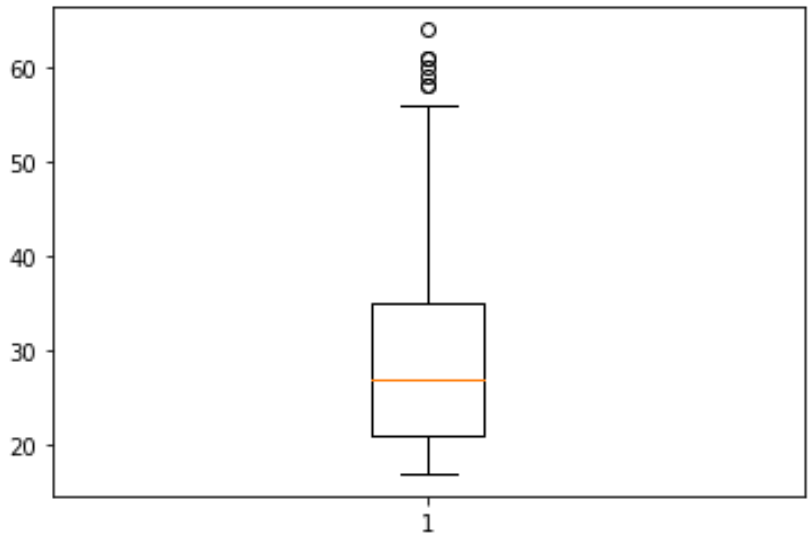

FIGURE 1.5 – Boîte à moustache de la varaible Âge

La répartition de notre échantillon est en grande majorité sur des personnes ayant entre 20 et 40 ans, ce que nous confirme ce schéma.

1.3 Genre

Le meilleur moyen de présenter cette variable est là encore un tableau.

	Femmes	Hommes
Atteint(e)s	54	37
Total(e)s dans l'échantillon	337	367

FIGURE 1.6 – Tableau du nombre d'individus atteints et total selon le sexe

Les femmes semblent plus touchées par ce syndrome puisqu'on relève : 16% des femmes sont atteintes contre 10% chez les hommes. Dans cet échantillon, parmi les 91 individus présentant ce trouble, presque un tiers sont des femmes.

En effet, on peut évaluer le rapport entre les individus atteints et non atteints pour chacun des deux genres :

— Femmes : $\frac{54}{283} = 0,19081$

— Hommes : $\frac{37}{337} = 0,10979$

Le rapport pour les femmes étant bien supérieur, la gente féminine semble donc plus vulnérable face à ce syndrome.

1.4 Né avec la jaunisse

Sur l'ensemble des 704 individus, seulement 69 d'entre eux sont nés avec la jaunisse, soit un peu moins de 10%. Or dans l'échantillon des 91 individus atteints, on relève 20 personnes nées avec la jaunisse soit 21,92% contre 49 cas sur 564 parmis les individus sains soit environ 8%.

	Individus atteints	Individus non atteints	Total
Né avec la jaunisse	20	49	69
Né sans la jaunisse	71	564	635

FIGURE 1.7 – Tableau de contingence entre Né avec la jaunisse et Autisme

Même si parmi les individus atteints la proportion de Né avec la jaunisse est plus importante que parmi les individus non atteints, il semblerait à priori que cette variable n'est pas réellement décisive sur le fait que l'individu présente le trouble ou non.

1.5 A1_score à A10_score

Tout d'abord, calculons la moyenne des scores obtenus pour chacune de ces questions. Nous rappelons ici que la valeur 0 indique que l'individu est d'accord ('oui') avec la proposition A ainsi 1 traduit son désaccord ('non').

	A1 score	A2 score	A3 score	A4 score	A5 score	A6 score	A7 score	A8 score	A9 score	A10 score
Moyenne obtenue par les individus atteints	0,8351	0,5495	0,6044	0,7473	0,6154	0,4176	0,4066	0,6923	0,5275	0,7253
Moyenne obtenu par les individus sains	0,4389	0,4356	0,4584	0,4812	0,2643	0,4192	0,6427	0,2936	0,5514	0,4388
Moyenne obtenue dans tout l'échantillon	0,7216	0,4531	0,4574	0,4957	0,4986	0,2841	0,4176	0,6491	0,3239	0,5739

FIGURE 1.8 – Moyenne des scores obtenus sur les 10 questions des individus atteints et non atteints

Ici, un diagramme sera plus représentatif afin de comparer les moyennes obtenues par les individus sains et celles obtenues par les individus atteints du trouble.

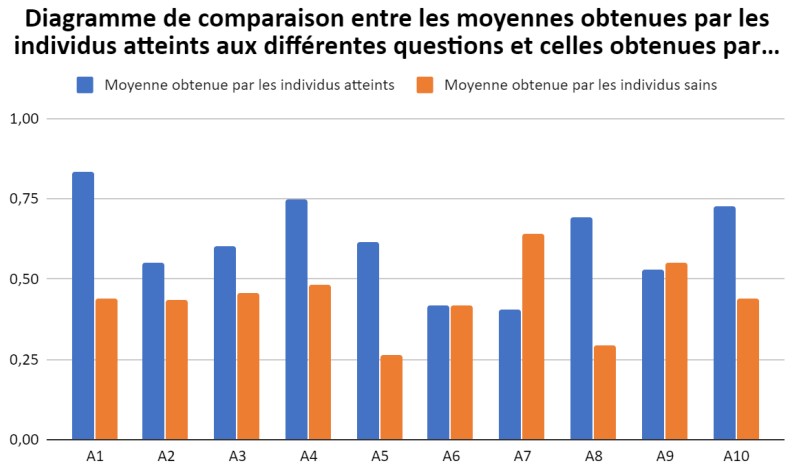

FIGURE 1.9 – Diagramme comparant les moyennes des scores des individus atteints et non atteints

Grâce à ce diagramme, on remarque que certaines réponses aux questions semblent être décisives quant à la présence du trouble chez un individu ou non. En effet, un individu atteint du trouble du spectre autistique :

- remarquera moins souvent de petits bruits que d'autres ne remarquent pas
- se concentrera davantage sur les détails que sur l'image
- ne trouvera pas facile d'entreprendre deux choses à la fois
- aura du mal à revenir à ce qu'il faisait en cas d'interruption
- aura des difficultés à lire entre les lignes quand on lui parle
- rencontrera de plus grandes difficultés pour déterminer les intentions des personnages d'une histoire
- ne rassemblera pas particulièrement d'informations sur des catégories d'objets

— ne trouvera pas difficile de comprendre les intentions des gens

1.6 Score obtenu

En moyenne, un individu atteint de l'autisme obtient une note d'environ 6,12 tandis qu'un individu sain obtiendra plutôt un 4,68 c'est à dire une majorité de "oui" aux questions A1 à A10.

1.7 Appli détecte autisme

L'application a détecté, en tout, 189 cas d'autisme parmi les 704 individus de l'échantillon. Parmi eux on relève 146 individus sains et donc seulement 43 individus réellement atteints du trouble du spectre autistique. parmi les 91 individus réellement atteints, l'application n'en a détecté que 47,3% d'entre eux, soit moins de la moitié. L'application semble ainsi à premier abord très peu fiable.

1.8 Déjà utilisé l'application

Parmi l'ensemble des individus, seulement 12 avaient déjà utilisé l'application, et seulement 2 sont atteints parmi ces derniers. Cette variable est inutile pour prévoir si oui ou non un individu est atteint ou pas.

1.9 Pays de résidence

Grâce à des calculs sur Colab Notebook, nous avons pu obtenir les pourcentages de répartition des individus par pays de résidence. Nous avons ainsi pu remarquer que 16,1% vivent aux Etats Unis, 11,6% aux Emirats Arabes, 11,5% en Inde ou encore en Nouvelle Zélande, 10,9% en Grande Bretagne et seulement 1,6% en France.

```
Paysvals = np.zeros((len(uPays),1))
for i in range(len(Pays)):
    for j in range(len(uPays)):
        if(Pays[i]==uPays[j]):
            Paysvals[j]=Paysvals[j]+1
```

```
PaysVals = 100*Paysvals/np.sum(Paysvals)
```

```
a=1                                    #a le nombre de chiffre après la virgule
print("Parmi les individus")
for i in range(len(uPays)):
    print(round(float(PaysVals[i])*10**a)/10**a,"% résident en ",uPays[i])
```

```
Parmi les individus
1.8 % résident en  Afghanistan
0.3 % résident en  AmericanSamoa
0.1 % résident en  Angola
0.3 % résident en  Argentina
0.3 % résident en  Armenia
0.1 % résident en  Aruba
3.8 % résident en  Australia
0.6 % résident en  Austria
0.1 % résident en  Azerbaijan
0.3 % résident en  Bahamas
0.4 % résident en  Bangladesh
0.4 % résident en  Belgium
0.1 % résident en  Bolivia
1.3 % résident en  Brazil
0.1 % résident en  Burundi
```

FIGURE 1.10 – Capture de *Colab* désignant la répartition des individus selon leur pays de résidence

Nous avons ensuite trouvé intéressant de calculer le pourcentage de répartition des individus par pays de résidence mais cette fois en sachant qu'ils sont atteints du trouble du spectre autistique, c'est à dire que nous avons classé les 91 individus atteints en fonction de leur lieu de vie. Ainsi, parmi les 91 cas, 25,3% vivent aux Etats Unis, 18,7% en Grande Bretagne, 3,3% aux Emirats Arabes, 2,2% en Inde et enfin 3,3% en France.

Ces chiffres n'étant pas très parlant, nous allons donc calculer les rapports entre ceux atteints et ceux qui résident au sein des Pays afin que ces chiffres nous soient plus significatifs :

— Etats-Unis : $\frac{0.253*91}{0,161*704} = 0,203$
— Grande Bretagne : $\frac{0.187*91}{0.109*704} = 0,222$
— Emirats Arabes : $\frac{0.033*91}{0,116*704} = 0,037$
— France : $\frac{0.033*91}{0,016*704} = 0,267$

Il semblerait ici que certains pays soient plus touchés que d'autres par cet handicap, par exemple les Emirats Arabes sont six fois moins touchés que la Grande Bretagne alors qu'ils sont plus représentés dans l'échantillon. En effet, si on calcule les effectifs pour chaque pays, 3,7% des individus de l'échantillon vivants aux Emirats Arabes sont atteints contre 22,1% en Grande Bretagne. Cette variable ne semble pas très significative de la détection d'un cas d'autisme.

1.10 Ethnique

Nous nous sommes ensuite intéressés à la variable qualitative Ethnique. De même que pour la variable Pays de résidence nous avons dressé la répartition des individus en fonction de leur Ethnique.

Parmi les individus :
— 17.5 % ont une ethnique Asian
— 6.1 % ont une ethnique Black
— 1.8 % ont une ethnique Hispanic
— 2.8 % ont une ethnique Latino
— 13.1 % ont une ethnique Middle Eastern '
— 4.3 % ont une ethnique Others
— 1.7 % ont une ethnique Pasifika
— 5.1 % ont une ethnique South Asian'
— 0.9 % ont une ethnique Turkish
— 33.1 % ont une ethnique White-European
— 13.5 % ont une ethnique inconnue
— 0.1 % ont une ethnique others

Nous remarquons donc une grande majorité d'Européens dans notre échantillon et très peu de Turcs, Hispanics ou encore de Latinos.
Ensuite nous avons voulu savoir qu'elle était alors, en terme d'individus atteints du trouble d'autisme, la nouvelle répartition.

Parmi les individus atteints :
— 5.5 % ont une ethnique Asian
— 5.5 % ont une ethnique Black

— 1.1 % ont une ethnique Hispanic
— 8.8 % ont une ethnique Latino
— 9.9 % ont une ethnique Middle Eastern ’
— 2.2 % ont une ethnique Others
— 2.2 % ont une ethnique Pasifika
— 2.2 % ont une ethnique South Asian’
— 1.1 % ont une ethnique Turkish
— 54.9 % ont une ethnique White-European
— 6.6 % ont une ethnique inconnue

On remarque ici une grosse concentration sur l’Ethnique White-European parmi ceux atteints, à se demander si le fait de faire partie de cette Ethnique influe sur le fait d’être atteint.

Pour avoir une idée plus précise, nous nous sommes intéressés à chaque ethnique pour ainsi mieux mettre en avant les plus touchées. Grâce à Colab nous obtenons les proportions suivantes :

- 4.1 % ayant une ethnique Asian sont atteints
- 11.6 % ayant une ethnique Black sont atteints
- 7.7 % ayant une ethnique Hispanic sont atteints
- 40.0 % ayant une ethnique Latino sont atteints
- 9.8 % ayant une ethnique Middle Eastern ’ sont atteints
- 6.7 % ayant une ethnique Others sont atteints
- 16.7 % ayant une ethnique Pasifika sont atteints
- 5.6 % ayant une ethnique South Asian’ sont atteints
- 16.7 % ayant une ethnique Turkish sont atteints
- 21.5 % ayant une ethnique White-European sont atteints
- 6.3 % ayant une ethnique inconnu sont atteints

Avec cette étude, les Latinos semblent être les plus vulnérables face à cet handicap contrairement à l’Asie du Sud, même si l’échantillon de Latinos est assez faible pour conclure. Les Européens restent cependant les deuxièmes plus touchés avec 21,5%, nous verrons par la suite si ce facteur est vraiment significatif.

1.11 Parenté avec l’individu

Enfin, nous avons analysé de la même façon la variable “Parenté avec l’individu réalisant le test” et avons remarqué que :

Parmi les individus :

- 13.5 % ont une Parenté inconnue
- 0.6 % ont une Parenté Health care professional’ (=professionnel de la santé)
- 0.7 % ont une Parenté Others
- 7.1 % ont une Parenté Parent
- 4.0 % ont une Parenté Relative (=Ami)
- 74.1 % ont une Parenté Self (=Soi même)

De plus,

— 6.3 % ayant une Parenté Inconnue sont atteints

— 25.0 % ayant une Parenté Health care professional sont atteints

— 20.0 % ayant une Parenté Others sont atteints

— 26.0 % ayant une Parente Parent sont atteints

— 21.4 % ayant une Parenté Relative sont atteints

— 12.3 % ayant une Parenté Self sont atteints

La répartition des cas est plus ou moins répartie de la même manière selon les différents liens de parenté. On remarque cependant un taux un peu plus faible pour les individus atteints lorsqu'on possède une parenté Self. La variable Parenté avec l'individu semblerait alors très peu significative pour reconnaître un cas d'autisme ou non.

2. Analyse en Composantes Principales (ACP)

Pour comprendre les liens entre les différentes variables et les individus nous avons réalisé une ACP. Les variables étudiées ici sont uniquement les variables quantitatives.

2.1 Centrer et réduire

Nous avons commencé par centrer et réduire dans le but de pouvoir comparer des choses comparables sans effet de "taille" notamment pour l'âge puisque le reste des variables sont principalement des variables binaires. On obtient alors la matrice symétrique de Covariance-Variance suivante :

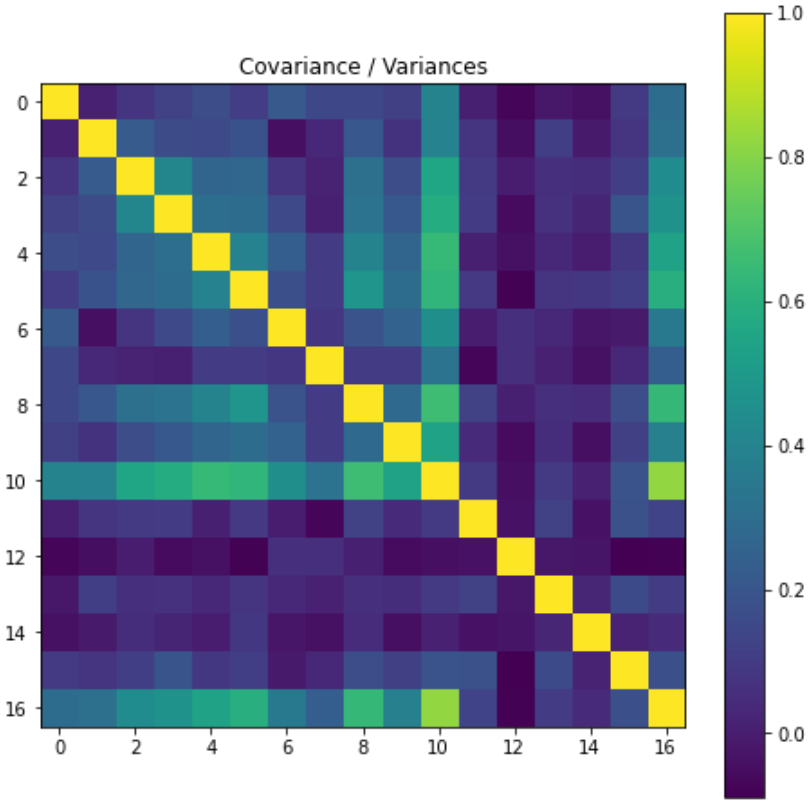

FIGURE 2.1 – Matrice de covariance-variance des variables

On retrouve bien que la 11ème variable (n°10 sur le schéma) ,qui représente la note totale obtenue aux réponses aux 10 premières questions, dépend bien des 10 premières variables qui sont elles-mêmes ces réponses. En effet cette 11ème variable est la somme des 10 premières variables binaires. De plus, on

peut noter que la 17ème variable, qui n'est autre que la variable qui indique si l'application considère l'individu présente ou non ce trouble, présente une forte dépendance de la 11ème variable c'est à dire la note totale. En effet, si cette somme est supérieur à 7, l'algorithme conclut que l'individu est atteint de cet handicap. Implicitement, la variable indiquant la réponse de l'application dépend ainsi uniquement des 10 premières variables qui sont les réponses aux questions A1 à A10.

2.2 Valeurs propres

Nous avons ensuite calculé les valeurs propres et vecteurs propres associés afin de déduire les axes principaux les plus importants. Nous avons obtenu la matrice de vecteurs propres suivante :

```
array([3.29074436e+03, 8.40993117e-14, 1.00947280e+03, 1.70449668e+02,
       8.14486089e+02, 3.52512015e+02, 3.93828347e+02, 4.18765637e+02,
       4.57960186e+02, 7.48644291e+02, 5.13241308e+02, 5.23469839e+02,
       5.90962302e+02, 6.51436504e+02, 6.11404637e+02, 7.12729385e+02,
       7.07892631e+02])
```

FIGURE 2.2 – Matrice des vecteurs propres originale

En réarrangeant les valeurs de la matrice afin d'avoir les plus significatives en premières on a obtenu la matrice suivante :

```
array([3290.74436138, 1009.47279887,  814.48608903,  352.51201525,
        393.82834659,  418.76563703,  457.96018642,  748.64429127,
        513.24130836,  523.46983898,  590.96230229,  651.43650359,
        611.40463731,  712.72938525,  707.89263071,  170.44966768,
          0.        ])
```

FIGURE 2.3 – Matrice des vecteurs propres arrangée

On obtient alors :

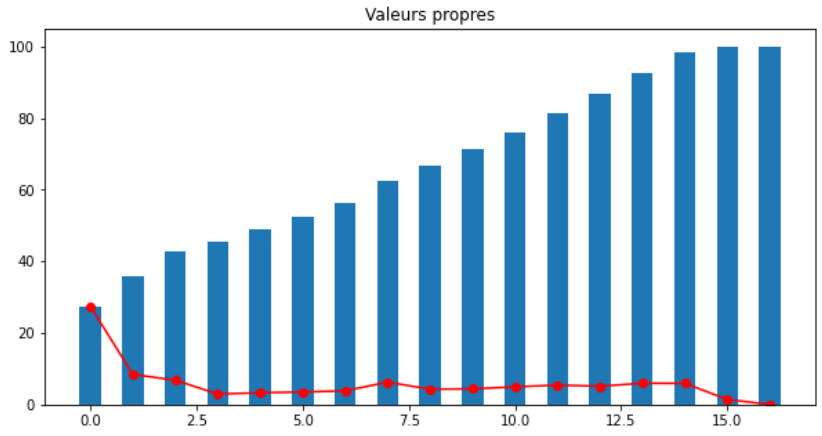

FIGURE 2.4 – Diagramme bâton des vecteurs propres

Les trois premières valeurs représentent ainsi 43% de l'information.

2.3 Projection selon plusieurs axes

Nous avons donc décidé de projeter dans un premier temps le nuage de points représentant les individus sur les axes des deux principaux facteurs. On obtient alors le nuage de points suivant :

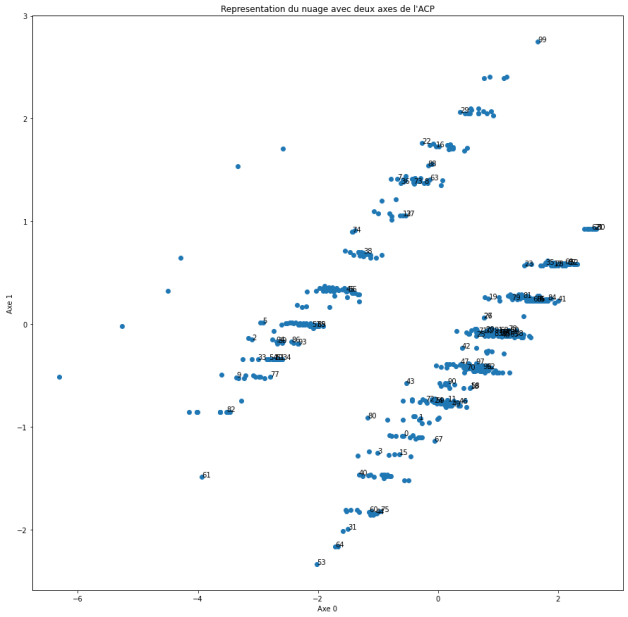

FIGURE 2.5 – Projection des individus sur les deux axes principaux

Enfin, en projetant les variables dans le plan des deux principaux facteurs,on a obtenu la représentation suivante :

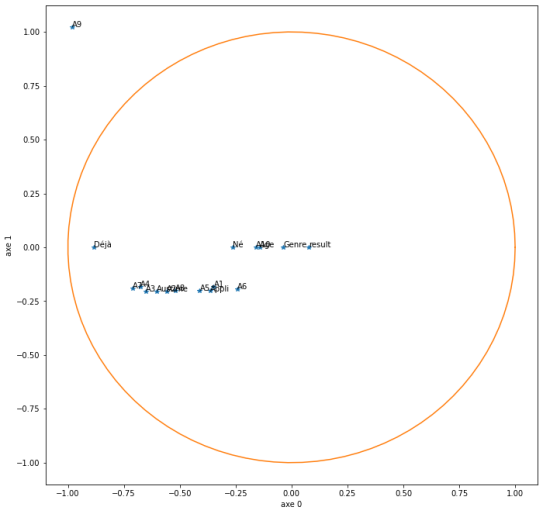

FIGURE 2.6 – Projection des variables sur les deux axes principaux

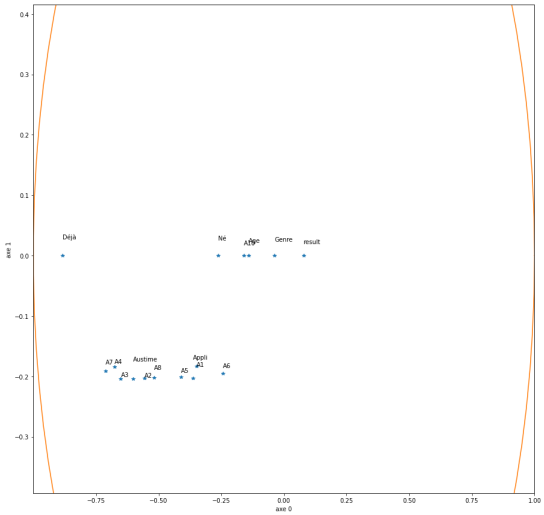

FIGURE 2.7 – Projection zoomée des variables sur les deux axes principaux

On en a donc déduit que la variable A9 étant hors du cercle de corrélation, n'était pas réellement significative tandis que les variables allant de A1 à A8 ainsi que la réponse de l'appli, formant un nuage de points avec la variable autisme présente un lien fort entre elles.

Afin de confirmer cela, nous avons réalisé la projection des variables dans le plan définit par le premier et le troisième principal facteur. Nous avons alors obtenu la répartition du nuage de points suivant :

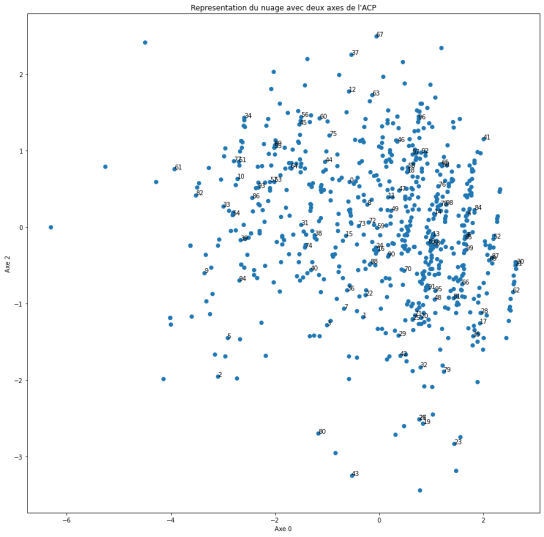

FIGURE 2.8 – Projection des individus sur le premier et troisième axe

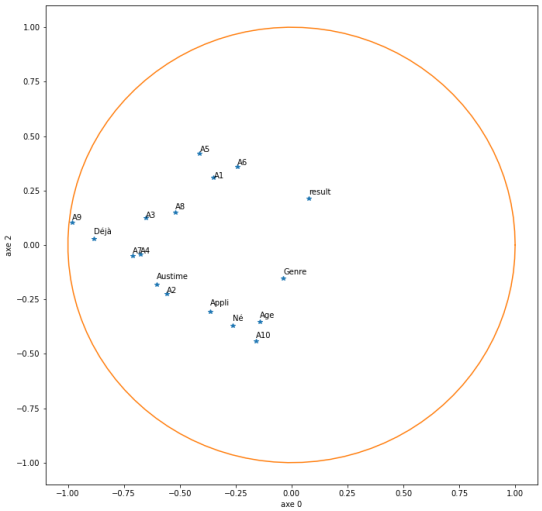

FIGURE 2.9 – Projection des variables sur le premier et troisième axe

Là encore, on remarque l'éloignement de la variable A9 par rapport à celle autisme. Or les variables A2, A4 et A7 sont très proches de la variable autisme et semblent donc avoir une importance plus déterminante que les autres.

3. Régression linéaire multiple

Afin d'analyser la possibilité de prédire l'autisme d'un individu ou non en fonction des réponses aux 10 questions A1 à A10, nous avons utilisé la régression linéaire. En effet, l'application annonce que l'individu est autiste uniquement si le résultat obtenu à ces questions est supérieur ou égal à 7. Si cela est possible par hypothèse, nous devrions alors obtenir un indice de qualité de régression (R2) proche de 1 signifiant ainsi que le modèle est bon. En choisissant $y = 1$ si l'individu est autiste, 0 s'il ne l'est pas, et $x1 = A1, x2 = A2 \ldots, x10 = A10$. Soit $X = [x1, x2, ..., x10]$, nous obtenons $Y = aX$ avec a la matrice de régression :

```
a= [ 0.12196816,  0.06821539,  0.07106671,  0.1554585 ,  0.06627364,
       0.06019186,  0.0170575 ,  0.08809841,  0.13248486,  0.10949802,
      -0.06837607,  0.00475199, -0.04283624,  0.1423172 ,  0.03160695,
      -0.1170402 ]
```

La représentation des résidus de la régression $e = y - Xa$ est la suivante :

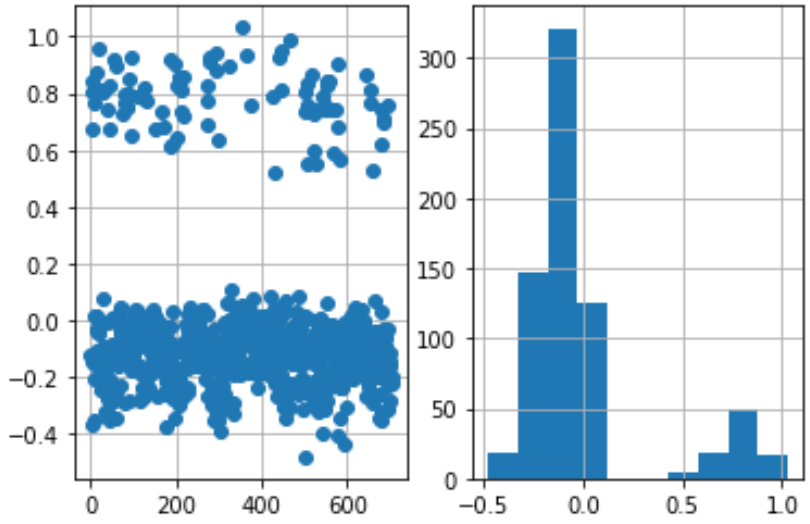

FIGURE 3.1 – Nuage de point et diagramme des résidus

Ainsi on obtient un R2=0,10538 et le schéma de la régression suivant où les points rouges représentent y et les points verts sont $z = Xa$ c'est à dire la prédiction de la variable autisme grâce à cette régression linéaire :

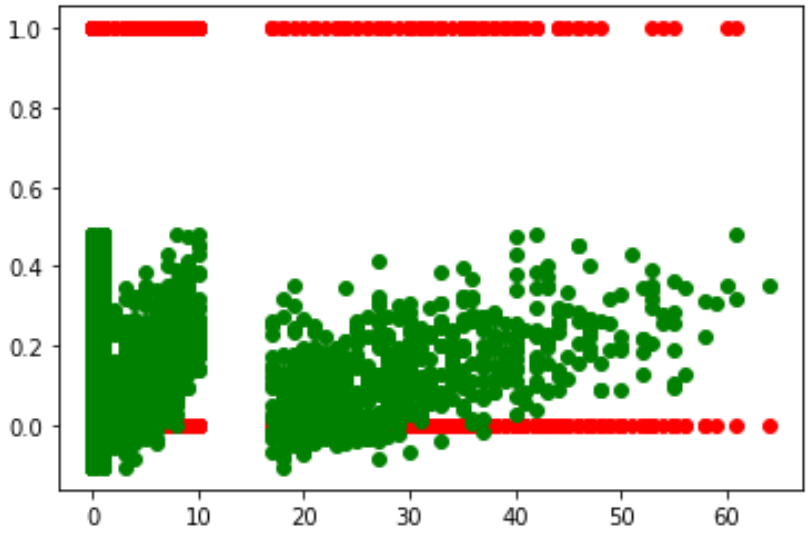

FIGURE 3.2 – Résultat de la régression linéaire multiple

On remarque donc que le R2 est proche de 0 c'est à dire que le modèle de régression est mauvais donc que la variable autisme ne peut pas être prédit à l'aide uniquement des réponses aux questions A1 à A10. Là encore, on en déduit que l'algorithme de détermination d'un cas d'autisme sur lequel repose l'application ASD Tests n'est pas fiable.

4. Régression logistique

Notre régression linéaire n'ayant pas été satisfaisante, nous avons tenté de réaliser une régression logistique avec l'aide de notre chargé de TD, Cyprien Ruffino. Ce modèle semble plus adapté à nos données. En effet, en posant p la probabilité que notre variable autisme (ici y) vaut 1, on pose alors la régression linéaire comme :

$$\ln \frac{p}{1-p} = \beta + \sum_{i=1}^{p} \alpha_p x_p$$

Ainsi il existe une relation entre X (la matrice des variables $x0 \dots xn$) et le logit de y telle que :

$$y = \frac{1}{1 + \exp(-z)}$$

$$z = \beta + \sum_{i=1}^{p} \alpha_p x_p$$

La représentation graphique obtenue est la suivante :

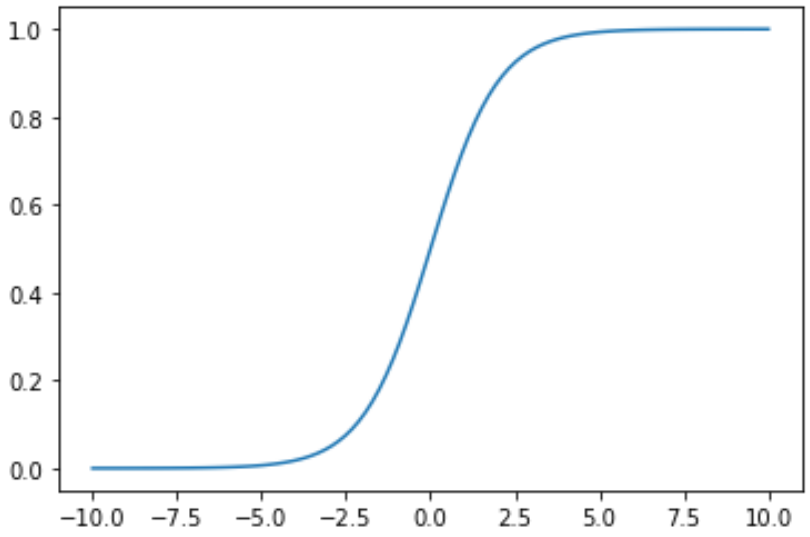

FIGURE 4.1 – Logit de la régression logistique

On notera que cette représentation correspond bien à une représentation de régression logistique. Nous avons ensuite procédé à une sélection de variables grâce à la fonction SelectFromModel. Cette dernière a affirmé que les six variables avec les plus grands coefficients sont les variables Âge, Né avec la jaunisse, le Résultat obtenu, le Genre ainsi que les réponses aux questions A4 et A7. Lorsque l'on évalue le modèle, on obtient une précision de 0,87, ce qui est proche de 1. Ainsi cette valeur permet d'admettre raisonnablement que le modèle est bon.

5. Tableaux de contingence et test du chi2

Dans le but de juger de la dépendance entre les variables (que la régression logistique a jugé importante) et la variable autisme, nous avons réalisé des tableaux de contingence, évalué les distances du chi2 obtenues pour chacune des variables associées à la variable autisme, et avons calculé la p-valeur associée à chacune d'entre elles. On a également étudié ces estimateurs pour la variable A2 que notre ACP avait considéré comme proche de la variable autisme.

5.1 Effectifs normaux

Voici les tableaux de contingence des observations obtenues :

	Femme	Homme	Jaunisse	Pas de jaunisse	A4 oui	A4 non	A7 oui	A7 non	A2 oui	A2 non
Autiste	54	37	20	71	68	23	37	54	41	50
Non autiste	283	330	49	564	281	332	257	356	344	269

FIGURE 5.1 – Tableau de contingence entre plusieurs variables

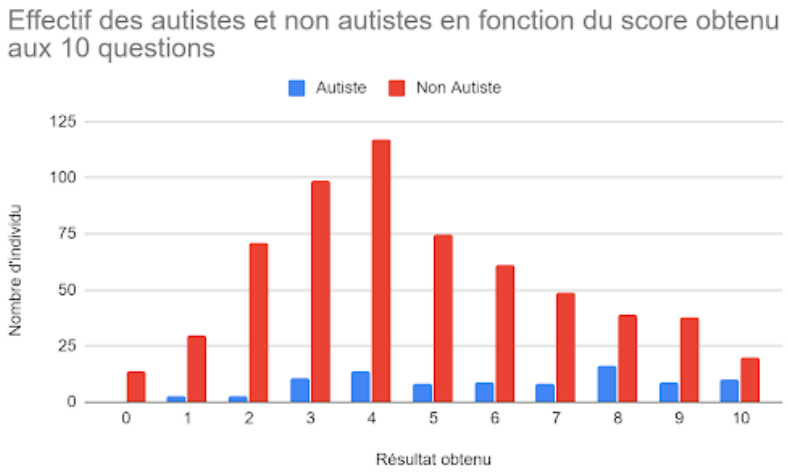

FIGURE 5.2 – Graphique représentant la répartition des individus selon leurs réponses aux questions

	16<Age<20	19<Âge<30	29<Âge<40	39<Age<50	49<Age<60	59<Age
Autisme	7	21	40	18	3	2
Non autisme	77	321	124	62	27	2

FIGURE 5.3 – Tableau de contingence entre Âge et Autisme

5.2 Effectifs théoriques

Voici ensuite les tableaux de contingence des effectifs théoriques :

	Femme	Homme	Jaunisse	Pas de jaunisse	A4 oui	A4 non	A7 oui	A7 non	A2 oui	A2 non
Autiste	43,56	47,44	8,92	82,08	45,11	45,89	38,00	53,00	49,77	41,23
Non autiste	293,44	319,56	60,08	552,92	303,89	309,11	256,0	357,0	335,23	277,77

FIGURE 5.4 – Tableau de contingence des effectifs théoriques entre plusieurs variables

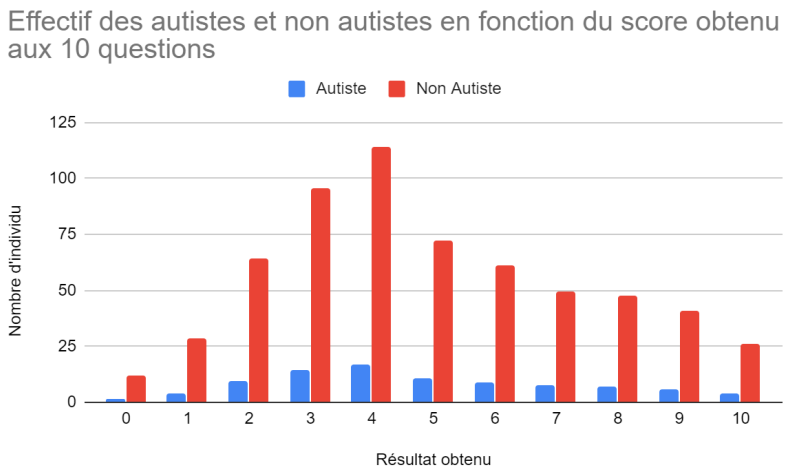

FIGURE 5.5 – Graphique représentant la répartition théorique des individus selon leurs réponses aux questions

	16<Age<20	19<Age<30	29<Age<40	39<Age<50	49<Age<60	59<Age
Autisme	10,86	44,21	21,20	10,34	3,88	0,52
Non autisme	73,14	297,79	142,80	69,66	26,12	3,48

FIGURE 5.6 – Tableau de contingence des effectifs théoriques entre Âge et Autisme

5.3 χ^2et p-valeurs

On obtient alors les valeurs de chi2 et les p-valeurs entre la variable et les différentes variables suivantes :

Variable	Genre	Âge	Résultat	Jaunisse	A2	A4	A7
chi 2	5,51	46,34	35,45	17,53	3,91	26,45	0,05
p-valeur	0,02	7,73e-9	1,4e-4	2,83e-5	0,05	2,71e-7	0,82

FIGURE 5.7 – Tableau avec le χ^2 et la p-valeur des variables suivantes

En étudiant les p-valeurs et en fixant l'erreur de première espèce acceptable à 0,05, on peut raisonnablement affirmer que la variable autisme dépend de sa réponse aux questions A2 et A7.

Conclusion

En étudiant l'algorithme de décision de l'application ASD Tests qui permet d'identifier un trouble du spectre autistique chez un individu, ainsi que l'ACP obtenue lors de l'étude de la base de données de l'application, nous avons pu justifier raisonnablement que celle-ci n'était pas fiable. En effet, elle présente un nombre conséquent d'erreurs et l'ACP a justifié cela en mettant en avant notamment le fait que la variable Autiste était peu dépendante du score obtenu, alors que l'application affirme que l'individu est atteint si ce score est supérieur à 7.

Toutefois, grâce aux données que l'application enregistre sur l'individu étudié, le programme peut être amélioré en modifiant l'importance de certaines variables. Ceci a été prouvé par la régression logistique. La régression a estimé les variables Genre, Âge, Né avec la jaunisse, Résultat ainsi que les réponses aux questions A4 (S'il y a une interruption je peux très facilement revenir à ce que je faisais) et A7 (Quand je lis une histoire je trouve difficile de déterminer les intentions des personnages) comme plus importantes. Le test du chi2 et l'étude des p-valeurs nous ont permis d'étudier la dépendance entre la variable Autiste et chacune de ces variables définies comme caractéristiques d'un autiste par la régression. L'individu type semble rencontrer des difficultés à déterminer l'intention des personnages quand il lit une histoire puisque cette hypothèse a été validée par l'ACP, la régression logistique ainsi que le test du chi2 ce qui est très révélateur. En ce qui concerne les autres variables, les différents estimateurs ne permettent pas de conclure suite à des contradictions.

Du fait de la liberté du choix du thème de notre projet, nous avons pu travailler sur un sujet qui nous plaît réellement, ce qui nous a facilité l'envie d'approfondir certaines notions. Restituer les connaissances vu en cours afin de les adapter de manière concrète dans notre projet fut très bénéfique dans l'apprentissage et la compréhension de la M8. Ce projet nous a ainsi permis d'étudier une base de données afin de la comprendre et de l'expliquer scientifiquement.

Table des figures

Annexes

Voici un aperçu de nos données :

A1_Scor	A2_Score	A3_Score	A4_Score	A5_Score	A6_Score	A7_Score	A8_Score	A9_Score	A10_Score	Age	Genre	Ethnique
1	1	1	1	0	0	1	1	0	0	26	f	'White-European
1	1	0	1	0	0	0	1	0	1	24	m	Latino
1	1	0	1	1	0	1	1	1	1	27	m	Latino
1	1	0	1	0	0	1	1	0	1	35	f	'White-European
1	0	0	0	0	0	0	1	0	0	40	f	?
1	1	1	1	1	0	1	1	1	1	36	m	Others
0	1	0	0	0	0	0	1	0	0	17	f	Black
1	1	1	1	0	0	0	0	1	0	64	m	'White-European
1	1	0	0	1	0	0	1	1	1	29	m	'White-European
1	1	1	1	0	1	1	1	1	0	17	m	Asian
1	1	1	1	1	1	1	1	1	1	33	m	'White-European
0	1	0	1	1	1	1	0	0	1	18	f	'Middle Eastern '
0	1	1	1	1	1	0	0	1	0	17	f	?
1	0	0	0	0	0	1	1	0	1	17	m	?
1	0	0	0	0	0	1	1	0	1	17	f	?
1	1	0	1	1	0	0	1	0	1	18	m	'Middle Eastern '
1	0	0	0	0	0	1	1	1	1	31	m	'Middle Eastern '
0	0	0	0	0	0	0	1	0	1	30	m	'White-European
0	0	1	0	1	1	0	0	0	0	35	f	'Middle Eastern '
0	0	0	0	0	0	1	1	0	1	34	m	?
0	1	1	1	0	0	0	0	0	0	38	m	?
0	0	0	0	0	0	0	0	0	0	27	f	Black
0	0	0	1	0	0	1	1	1	1	27	m	'Middle Eastern '
0	0	0	0	0	0	0	1	0	1	42	m	'Middle Eastern '
1	1	1	1	0	0	0	1	0	0	43	m	?
0	1	1	0	0	0	0	1	0	0	24	f	?
0	0	0	0	0	0	0	1	0	0	40	m	Pasifika
0	0	0	0	0	0	0	1	0	0	40	m	'Middle Eastern '
0	0	0	0	0	0	0	1	0	0	48	m	Black
0	1	1	0	0	0	0	0	1	1	31	m	'Middle Eastern '
0	0	0	0	0	0	0	0	0	0	18	m	'White-European
1	0	0	1	1	1	1	1	0	1	37	f	'White-European
1	1	0	0	0	0	1	0	0	1	55	f	Others
1	1	1	1	1	1	1	1	1	1	18	f	'White-European
1	1	1	1	1	1	1	1	1	1	18	f	'White-European
0	0	1	0	0	0	0	0	0	0	55	m	'White-European
0	1	1	0	1	0	0	1	1	1	50	m	'Middle Eastern '
1	0	1	1	1	1	0	0	1	0	34	f	'White-European
1	0	0	1	1	1	1	0	1	1	53	f	'White-European
1	0	1	1	0	1	1	1	1	1	35	f	'White-European
1	0	1	1	1	0	1	1	0	1	20	f	Latino
0	0	0	0	1	1	0	0	0	0	28	f	Asian
0	0	1	1	0	0	0	0	0	1	34	f	'Middle Eastern '
0	1	1	1	0	0	0	0	0	1	36	f	'White-European
1	1	1	1	1	1	0	1	0	1	27	f	'White-European
1	0	1	1	1	1	0	1	1	0	53	f	'White-European
1	1	1	1	0	1	0	0	0	0	24	f	Pasifika
0	0	1	1	1	0	1	0	0	0	24	m	Pasifika
0	1	1	0	0	1	0	0	0	0	55	m	'White-European

Extrait des données de notre échantillon [1]

1. Données obtenues sur le site de l'UCI

Né avec la jauni	Austime	Pays de résidenc	Déjà utilise l'application a	result numeric	Parenté avec l'individu at	Appli detecte autisme
no	no	'United States'	no	6	Self	NO
no	yes	Brazil	no	5	Self	NO
yes	yes	Spain	no	8	Parent	YES
no	yes	'United States'	no	6	Self	NO
no	no	Egypt	no	2	?	NO
yes	no	'United States'	no	9	Self	YES
no	no	'United States'	no	2	Self	NO
no	no	'New Zealand'	no	5	Parent	NO
no	no	'United States'	no	6	Self	NO
yes	yes	Bahamas	no	8	'Health care professional'	YES
no	no	'United States'	no	10	Relative	YES
no	no	Burundi	no	6	Parent	NO
no	no	Bahamas	no	6	?	NO
no	no	Austria	no	4	?	NO
no	no	Argentina	no	4	?	NO
no	yes	'New Zealand'	no	6	Parent	NO
no	no	Jordan	no	5	Self	NO
no	no	Ireland	no	2	Self	NO
no	yes	'United Arab Emirate	no	3	Self	NO
yes	no	'United Arab Emirate	no	3	?	NO
no	no	'United Arab Emirate	no	3	?	NO
no	no	'United Arab Emirate	no	0	Self	NO
no	no	Afghanistan	no	5	Self	NO
yes	no	'United Arab Emirate	no	2	Relative	NO
no	no	Lebanon	no	5	?	NO
yes	no	Afghanistan	no	3	?	NO
yes	yes	'United Arab Emirate	no	1	Self	NO
yes	yes	Afghanistan	no	1	Parent	NO
no	no	'New Zealand'	no	1	Self	NO
no	no	'United Kingdom'	no	4	Self	NO
no	no	'United Kingdom'	no	0	Self	NO
no	yes	'United States'	no	7	Self	YES
no	no	'New Zealand'	no	4	Self	NO
yes	no	'South Africa'	no	10	Self	YES
no	no	'South Africa'	no	10	Self	YES
no	no	'New Zealand'	no	1	Self	NO
no	no	'United Arab Emirate	no	6	Self	NO
no	no	'New Zealand'	no	6	Self	NO
no	no	'New Zealand'	no	7	Self	YES
no	yes	'United States'	no	8	Self	YES
yes	no	Italy	no	7	Self	YES
no	no	Pakistan	no	2	Self	NO
no	yes	Egypt	no	3	Self	NO
yes	yes	'United States'	no	4	Self	NO
no	no	'New Zealand'	no	8	Self	YES
no	no	'New Zealand'	no	7	Relative	YES
no	no	'New Zealand'	no	5	Relative	NO
no	no	'New Zealand'	no	4	Relative	NO
no	no	'New Zealand'	no	3	Relative	NO

Extrait des données de notre échantillon [1]

1. Données obtenues sur le site de l'UCI