Proyecto análisis estadística de datos

Autores:
Dayana Valentina González Vargas
María Fernanda Rodríguez Conde
Nelson Santiago Guayazan Palacios

Docente:
Edwin Santiago Alférez Baquero

Asignatura:
Análisis estadística de datos

Universidad del rosario
Matemáticas aplicadas y ciencias de la computación

24 de mayo de 2021

Índice

1. Resumen ejecutivo

La enfermedad del Alzheimer es un trastorno que afecta principalmente al sistema neurológico, donde el cerebro empieza a presentar atrofia provocando la muerte de neuras. Esta enfermedad afecta a toda la población, pero es más presente en personas mayores de 60 años, los cuales pueden presentar perdida de memoria y una deficiencia en sus habilidades cognitivas.

Mediante el análisis de una base de datos, tomada del proyecto *The Open Access Series of Imaging Studies*(OASIS), el cual es reconocido por brindarle a la comunidad científica datos de resonancias magnéticas de una población en especifico con el fin de que se realicen sus propios análisis. Utilizando esta base de datos miraremos una población de adultos mayores de entre 60 a 90 años de edad, los cuales se realizaron una resonancia magnética y pueden estar padeciendo la enfermedad del Alzheimer.

Nuestros objetivos a demostrar con respecto a esta base de datos es ver si el Alzheimer puede presentarse en un rango especifico de edades, también si existen algunas interacciones entre variables independientes y dependientes o viceversa, observando si entre estas hay efectos significativos. Por otro lado, podemos realizar un análisis en el cual determinemos si la clasificación realizada por el proyecto *The Open Access Series of Imaging Studies*(OASIS), es una buena clasificación o se encuentran falsos positivos y determinar cual seria el número de clustering mínimos, en donde se puedan clasificar toda las variables numéricas simultáneamente.

Para mirar esto se realizaron pruebas vistas en la asignatura de análisis estadístico de datos como: ANOVA, MANOVA, TWO WAY MANOVA, componetes principales, análisis de factores y clustering con los cuales vamos a tener soluciones a los problemas más precisas.

Con estas pruebas, podemos concluir del análisis de datos realizado que si existe una relación entre ambos examenes, CDR y MMSE, y que estos influyen en el grupo de cada persona, el cual esta determinado por su demencia. También identificamos que las variables como la edad, años de educación, tamaño del cerebro, volumen del cerebro, sexo y estrato, logran afectar el estudio, ya sea para ayudar a la clasificación de grupo de las personas, o para otorgarnos mas información al respecto con datos que pueden servir a futuro en diversos campos.

2. Descripción de la problemática

Actualmente, una de las enfermedades más comunes y degenerativas en los adultos mayores es el Alzheimer, el cual es un trastorno neurológico progresivo que genera una perdida de la memoria debido a un encogimiento del cerebro, provocando así, que las neuronas empiecen a morir. Esta enfermedad es una de las principales causas de la demencia y el deterioro continuo en el pensamiento, comportamiento y habilidades sociales[1].

Uno de los proyectos encargados de brindar a la comunidad científica datos de neuroimagen para su libre investigación es *The Open Access Series of Imaging Studies*(OASIS), el cual utilizaremos una de sus base de datos conformada por una población de 139 personas, los cuales tiene edades entre los 60 a 90 años que se realizaron una imagen de resonancia magnética (MRI) para determinar su salud mental.

Se quiere determinar un análisis de esta población para saber la probabilidad de obtener Alzheimer dependiendo de su edad, años de estudio, estrato socio económico y diferentes exámenes que miden la perdida de la memoria, el estado de demencia y la perdida en habilidades sociales.

2.1. Descripción de los datos

Esta base de datos fue tomada del proyecto *The Open Access Series of Imaging Studies*(OASIS)(2), donde tendremos en cuenta los siguientes datos:

- **Grupo:** Clasifica a la persona según sus resultados como demente, no demente y convirtiéndose.
- **Género:** Representa a los hombres (M) y a las mujeres (F).
- **Edad:**Determina la edad del paciente, sabemos que nuestra población varia en edad de 60 a 90 años.
- **Años de educación:** Determina cuantos años de estudio tuvo una persona.
- **Estrato socio económico:** Determina el estrato socio económico en el que esta una persona.
- **Mínimo examen de estado mental(MMSE):**Es un examen que contribuye a la detección de demencia en las personas, analizando la función cognitiva evaluando diferentes competencias, tiene como puntuación máxima 30 puntos y considera una función cognitiva "normalçon 24 puntos.[3]
- **Clasificación de demencia clínica (CDR):** Clasifica el estado mental de las personas en cinco niveles, el nivel 0; considerándolo una persona sana, el nivel 0.5; una demencia cuestionable, nivel 1; una demencia leve, nivel 2; una demencia moderada y el nivel 3; una demencia grave.[4]
- **Estimado del volumen intracraneal(ETIV) :** Es el tamaño del cerebro que varia entre 1132 a 2010 mm^3.

3. Análisis estadístico

Utilizando R estudio, comenzamos a analizar la base de datos MRI Alzheimer csv, realizando pruebas como; anova, manova, two way manova, PCA, entre otras. A continuación mostraremos los resultados de las pruebas:

Manova

Analizamos los efectos que pueden llagar a producirse mediante las variables independientes, que son: el examen mínimo mental(MMSE), la edad, los años de educación y el tamaño del cerebro con respecto a una variable dependiente que en este caso es el grupo en el que cada persona puede ser clasificada como (Dementes/ No dementes / Convertidos).

Hipótesis nula: Las variables examen mínimo mental(MMSE), la edad, los años de educación y el tamaño del cerebro no tienen ningún efecto en que la persona sea clasificada como demente, no demente o convertido.
Hipótesis alternativa: Al menos una de esas cuatro variables afecta esa clasificación como demente, no demente o convertido.

```
           Df   Wilks approx F num Df den Df   Pr(>F)
Group       2 0.59624   9.8106      8    266 6.09e-12 ***
Residuals 136
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

En este procedimiento obtuvimos que la hipótesis nula se rechaza, por lo cual al menos una de las cuatro variables mencionadas anteriormente afecta la clasificación de las personas.

Two way manova

Mediante esta prueba podríamos observar que existe una interacción entre el sexo, que lo tomaremos como nuestro factor uno y el estrato socio-económico, que es nuestro factor dos con respecto a unas variables independientes de la base de datos, las cuales son; edad, años de educación, el examen de estado mínimo mental(MMSE) y el volumen estimado del cerebro.

Hipótesis nula: No se presenta una interacción entre sexo y estrato socio-económico.
Hipótesis alternativa: El sexo y el estrato económico si interactúan.

```
                      Df   Wilks approx F num Df den Df    Pr(>F)
M.F                    1 0.63127   19.276      4    132 1.652e-12 ***
Socioeconomic.Status   1 0.49783   33.288      4    132 < 2.2e-16 ***
M.F:Socioeconomic.Status  1 0.93999    2.107      4    132   0.08347 .
Residuals            135
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note que obtuvimos una significación estadística tan alta, donde si tomamos un nivel de significación estadística de 0.05 no se podría rechazar nuestra hipótesis nula,por lo tanto concluimos que el sexo y el estrato socio-económico no tienen una interacción, pero individualmente si se causa algún efecto entre las variables independientes y los factores por separado.

Anova

- Analizando las variables independientes CDR y grupo, las cuales se refieren a el Ranting de demencia clínica y la clasificación (Demente/ No demente / Convertido) respectivamente, observando si se causa algún efecto en la variable dependiente,la cual es el examen mínimo mental (MMSE).

  Hipótesis nula: El CDR y el grupo no presentan un efecto significativo en el MMSE.
  Hipótesis alternativa: Al menos una de las dos (CDR y grupo) o ambas presentan un efecto significativo en los resultados del examen MMSE.

```
                              Df Sum Sq Mean Sq F value Pr(>F)
Clinical.Dementia.Rating       1 1156.0  1156.0 135.362 <2e-16 ***
Group                          2   50.1    25.1   2.936 0.0565 .
Clinical.Dementia.Rating:Group 1   34.2    34.2   3.999 0.0475 *
Residuals                    134 1144.4     8.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

De acuerdo con la tabla, podemos concluir que el CDR si trae efectos significativos en los resultados del examen

MMSE, ya que el nivel de significación es demasiado pequeño, donde la hipótesis nula se puede rechazar con un coeficiente de confianza de casi el 99 %. Por otro lado, el grupo no afectó el MMSE, pero la interacción entre ambas variables (CDR y grupo) si lograron afectar sus resultados.

- Ahora, observaremos las variables independientes edad y años de educación y si estas afectan o no el volumen intracranial estimado.

  Hipótesis nula: La edad y los años de educación no presentan un efecto significativo en el volumen intracranial estimado del cerebro.
  Hipótesis alternativa: Al menos una de las dos (edad y años de educación) o ambas presentan un efecto significativo en el volumen intracranial del cerebro.

```
                      Df  Sum Sq Mean Sq F value Pr(>F)
Age                    1    3527    3527   0.114 0.7360
Years.of.Education     1  183519  183519   5.940 0.0161 *
Age:Years.of.Education 1   56149   56149   1.817 0.1799
Residuals            135 4171143   30897
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

  Teniendo en cuenta la tabla podemos concluir que la única variable independiente que presenta efectos en el volumen intracranial estimado del cerebro son los años de educación con un coeficiente de confianza del 98 %, mientras que la edad o la interacción entre ambas no afecta este volumen.

Clasificación

- Decidimos hacer el problema de clasificación con el rating de demencia clínica, de esta manera observamos las medias de edad, años de educación y puntuación en el MMSE en cada uno de los resultados del CDR (0,1,0.5,2) y sus probabilidades con el fin de determinar si se realizo una buena clasificación de esta variable.
  Con los resultados obtuvimos que el 51 % de las personas tenían un CDR de 0, el 33 % lo tiene de 0.5, el 12 % de 1 y las personas que tienen un CDR de 2 son casi irrelevantes en la base de datos. Por otro lado, observamos que la edad de las personas que sacaron 0 a 1 es de 75 a 78 años mientras que las que sacaron 2 es de 85, mucho mas alta. También observamos que los años de educación no fueron muy relevantes en el CDR ya que sus medias son similares y en cuanto al examen MMSE si se presentan menores notas en las calificaciones 1 y 2, por lo que interpretaríamos que tienen correlación. La predicción total muestra lo siguiente:

```
      0 0.5  1  2
0    69   3  0  0
0.5  27  17  3  0
1     1   3 13  0
2     0   1  1  1
```

  Con lo cual interpretamos que dadas las 3 variables el programa detecta un fallo en la clasificación de varias personas. 69 personas de 72 fueron calificadas correctamente con el CDR=0, mientras que se detectó que 27 personas que fueron puntuadas como 0.5 deberían estar en 0. En cuanto a las otras calificaciones se observan pocos errores.

- Ahora realizamos un problema de clasificación teniendo el cuenta los grupos (Demente / No demente / Convertido), tomando las variables de edad, examen mínimo mental MMSE y el volumen intracranial. En este observamos que el 51 % de las personas son no dementes, el 38 % dementes y apenas el 1 % son convertidos. En cuanto a los promedios observamos que los dementes y no dementes mantienen edades similares (77 y 78). Además, el MMSE se ve mucho mas bajo en las personas dementes y estas presentan un volumen intracranial mayor al resto. Ahora veremos la prediccion total:

```
             Converted Demented Nondemented
Converted            0        3          11
Demented             0       29          25
Nondemented          0        0          71
```

Con esta concluimos que las personas convertidas son poco relevantes y por tanto el programa las clasificó como no dementes en su mayoría. Por otro lado, sólo 29 personas fueron clasificadas correctamente como dementes mientras que 25 debían ser dementes, y las no dementes son correctas en su totalidad.

Análisis de componentes principales (PCA)

Determinado los componentes principales, veremos como se pueden analizar estos datos usando la correlación entre las variables.tomaremos las siguientes variables: edad,años de educación, estrato socioeconómico, clasificación de demencia clínica (CDR), examen del estado mínimo mental(MMSE) y el volumen intracraneal estimado del cerebro.

```
    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6
1.3960605 1.2572741 1.0056758 0.9302323 0.55319555 0.53622353
0.3248308 0.2634564 0.1685640 0.1442220 0.05100422 0.04792261
0.3248308 0.5882872 0.7568512 0.9010732 0.95207739 1.00000000
```

Note que, tomando las cuatro primeras componentes es suficiente para explicar el 90 % de la variación de los datos.

comp1 <dbl>	comp2 <dbl>	comp3 <dbl>	comp4 <dbl>
0.06565485	0.07076862	0.97147730	0.20854739
0.76996867	-0.43726078	-0.12316489	0.21945095
-0.73388581	0.50271339	0.01150293	-0.24553452
0.65135330	0.64981505	0.01789636	-0.06095569
-0.54949809	-0.73052789	0.06644429	0.14233112
0.29499997	-0.41938253	0.21812515	-0.83031342

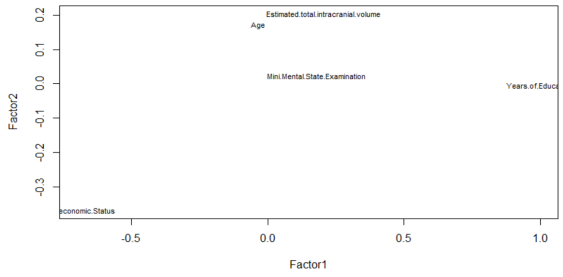
Note que el componente uno es un contraste entre la variable de años de educación y estrato socio-económico, la segunda componente se le puede retirar la variable de edad y se genera un overall, ya que todos no se encuentran tan alejados, la componente tres es solamente la edad y la componente cuatro es solamente la variable del volumen intracraneal estimado.

Análisis de factores (FA)

Mediante el análisis de factores, vamos a reducir las variables y visualizarlas utilizando su correlación, entonces tomaremos las siguientes variables: edad,años de educación, estrato socio-económico, examen del estado mínimo mental(MMSE) y el volumen intracraneal estimado del cerebro. Utilizando la estimación por máxima verosimilitud, en donde obtuvimos los siguientes factores, los cuales son suficientes para la base de datos.

```
                                Factor1 Factor2
Age                                       0.172
Years.of.Education               0.997
Socioeconomic.Status            -0.695  -0.370
Mini.Mental.State.Examination    0.181
Estimated.total.intracranial.volume  0.204   0.205
```

En los factores podemos notar, que en el factor uno se relacionan las variables, años de educación y el examen mínimo de estado mental, en el factor dos se relacionan más la edad, el estrato socio-económico y el volumen intracraneal estimado del cerebro. Obteniendo la siguiente gráfica:

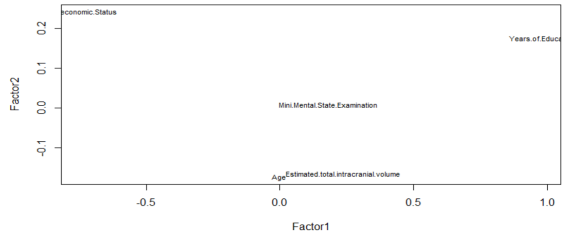


Rotación en análisis de factores

Realizamos una rotación para ver mejor la partición de las variables con respecto a los factores, donde obtendremos los siguientes factores:

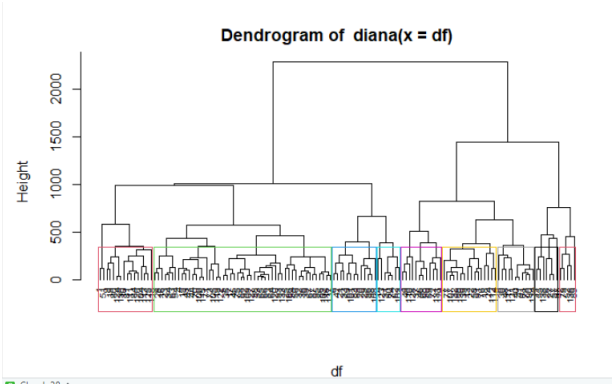
```
Factor1 Factor2
         -0.176
 0.982    0.177
-0.748    0.244
 0.182
 0.236   -0.167
```

Note que el factor uno va a relacionar de una mejor manera los años de educación, el examen mínimo de estado mental y el volumen intracraneal estimado del cerebro, el segundo factor relacionaría únicamente la edad y su estrato socio-económico.

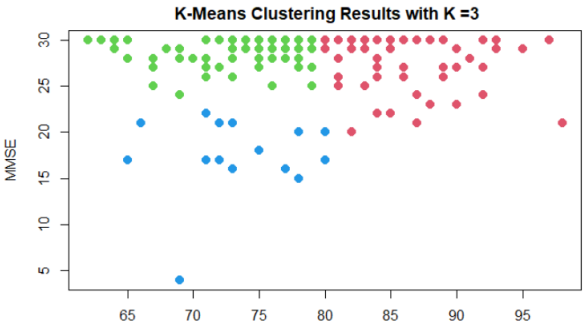


Clustering

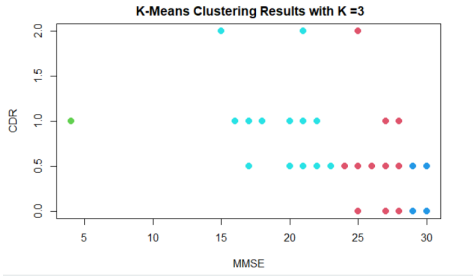
Realizamos un clustering jerárquico sobre todos los datos para identificar el número óptimo de clusters, que podemos usar para clasificar todas las variables.



Luego, realizamos un clustering con la técnica de k-means sobre 2 variables para poder compararlas y ver que tipo de correlaciones hay, como por ejemplo la relación entre la edad y la puntuación en el examen del CDR que determina el grado de demencia.



el método clustering también lo podemos enfocar con el fin de ver la correlación entre los exámenes de MMSE y CDR de la siguiente manera:



4. Conclusiones

Conclusión general

A lo largo del estudio, logramos encontrar diversas variables de la base de datos que no influían en la clasificación de demencia de la persona, pero fueron de gran ayuda para aclarar nuestras dudas sobre que tan influyentes pueden llegar a ser y cuales son a las que debemos prestar mas atención. Por otra parte, encontramos variables que serán de mucha ayuda, como los exámenes CDR y MMSE, que a pesar de no ser precisos, nos pueden ayudar a clasificar el nivel de demencia de la persona.

Conclusiones especificas

- Al menos una de las variables entre MMSE, edad, años de educación y tamaño del cerebro, influyen en que una persona esté clasificada como demente, no demente o convertido.
- El sexo y el estrato socio-económico no interactúan entre si, pero individualmente afectan algunas de las siguientes variables: la edad, años de educación, MMSE y volumen del cerebro.
- La calificación del CDR influye en los resultados del examen MMSE. Por ejemplo, cuando una persona tiene un CDR de 0, es mas probable que pase la prueba MMSE con una calificación mayor a 24.
- Aunque el hecho de que una persona sea demente, no demente o convertido no afecta su MMSE, la unión entre ambas características (grupo y CDR) si logra afectar los resultados de la prueba.
- El volumen intracranial del cerebro depende de los años de educación de acuerdo a los resultados de nuestra investigación.
- Las mitad de las personas clasificadas con demencia deberían estar en “No demente” y la clasificación de “Convertidos” no es muy relevante por lo que podría ser eliminada sin problema.
- Encontramos mediante el análisis de componentes principales, que se puede representar en cuatro componentes, donde la primera explica un contraste entre la variable de años de educación y estrato socio-económico, la segunda componente puede generar un overall con todas las variables menos la edad, la tercera componente explica más que todo la variable de edad y la cuarta explica la variable del volumen intracraneal estimado, estas componentes explican juntas el 90 % de la variación de los datos.
- Con el uso de análisis de factores podemos realizar una división teniendo en cuenta su correlación con las demás variables, donde debido a una rotación de factores vemos claramente que las variables años de educación, el examen MMSE y el volumen del cerebro tienen una correlación entre ellas.
- Con la ayuda del clustering y la técnica de K-means, pudimos determinar que el rango de edades con mayor probabilidad de Alzheimer es entre los 65 y 80 años, a partir de los 80 la probabilidad baja de manera considerable.
- El método de clustering nos facilita ver la relación que existe entre los exámenes que determinan el desgaste mental de las personas.
- El clustering nos ayudo a determinar, que el estrato socio-económico no tiene una influencia notable sobre el desgaste mental.

5. Referencias

[1] Mayo Clinic (2020).*Enfermedad del Alzheimer* https://www.mayoclinic.org/es-es/diseases-conditions/alzheimers-disease/symptoms-causes/syc-20350447(Visitado: *29 dic. 2020*)

[2] https://www.kaggle.com/jboysen/mri-and-alzheimers?select=oasis_longitudinal.csv

[3] Cochrane (2016).*Mini-Mental State Examination (MMSE) para la detección de la demencia en las personas de 65 años o mayores* https://www.cochrane.org/es/CD011145/DEMENTIA_mini-mental-state-examination-mmse-para-la-deteccion-de-la-demencia-en-las-personas-de-65-anos-o(*13 enero 2016*)

[4] Hipocampo (2007).*Clinical Dementia Rating (CDR) de Hughes* https://www.hipocampo.org/hughes.asp(*28 nov. 2007*)

[5]https://rpubs.com/mjimcua/clustering-jerarquico-en-r