Universidad del Bío-Bío
Facultad de Ingeniería
Departamento de Ingeniería Civil Industrial

Profesor Asignatura:
Danilo Gómez Correa

CERTAMEN N°1
TOPICOS AVANZADOS DE MINERIA DE DATOS

UNIVERSIDAD DEL BÍO-BÍO

Diego Yáñez Oyarce

Concepción, 16 de mayo de 2022

INDICE

INDICE DE ILUSTRACIONES

--------Summary descriptives table by 'cluster_kmean'---------

	1 N=28	2 N=16	3 N=29	4 N=14	5 N=13	p.overall
time_in_hospital	-0.41 (0.59)	1.18 (0.87)	-0.40 (0.64)	-0.37 (0.49)	-0.71 (0.35)	<0.001
num_lab_procedures	-0.10 (0.83)	1.10 (0.61)	0.63 (0.63)	-0.67 (0.76)	-1.20 (0.72)	<0.001
num_procedures	-0.71 (0.35)	0.49 (0.86)	-0.74 (0.27)	1.53 (0.90)	0.53 (0.94)	<0.001
num_medications	-0.26 (0.49)	1.15 (0.78)	-0.72 (0.54)	0.82 (0.77)	-0.37 (0.70)	<0.001
number_diagnoses	0.62 (0.35)	0.86 (0.13)	-1.09 (0.60)	0.49 (0.80)	-1.14 (0.62)	<0.001

OBJETIVOS

OBJETIVO GENERAL

1. Encontrar un patrón que permita agrupar individuos con alta probabilidad de tener diabetes.

OBJETIVOS SECUNDARIOS

1. Comprender la base de datos *"Diabetes 130-US hospitals for years 1999-2008 Data Set".*
2. Realizar la limpieza de la base de datos.
3. Construir y comparar cluster k-mean, Ward, mixto y clara.
4. Elegir y validar cluster que mejor describe a la base de datos.

PLAN DE PROYECTO

El procediendo el cual se va a seguir consta de las siguientes etapas:

- Análisis EDA: Se implementará un conjunto de técnicas para explorar, describir y resumir la naturaleza de las variables. Concretamente, se describirán las variables, se reajustarán (variables numéricas o categóricas), se tratarán los datos atípicos a través de los criterios de desviación estándar, método de los cuartiles, método de la entropía y coeficiente de variación, se eliminarán las variables que no aportan información, se tratarán los missing values con el método de imputación KNN y se determinará el grado de correlación entre las variables (Gomez, 2022).
- ACP: Se realizará el análisis de componentes principales para reducir las dimensiones y trabajar con aquellas variables e individuos que estén mejor representados.
- Clusters: Se realizará el análisis de conglomerados con al menos un cluster jerárquico y un cluster no jerárquico que permitirá agrupar a aquellos individuos que siguen patrones escondidos en la base de datos.
- Validación de los resultados: Se compararán los resultados del cluster que mejor se adapta a la base de datos con el label de la data original para sugerir un patrón que pueda predecir si un individuo puede ser un potencial paciente diagnosticado con diabetes.

ANALISIS EDA

DESCRIPCION Y NATURALEZA DE LAS VARIABLES

El data set reúne información de 101.766 consultas de pacientes en 130 hospitales estadounidenses entre 1999-2008. Estos pacientes poseen 50 atributos donde la variable label corresponde a si el individuo es diagnosticado con la enfermedad diabetes.

La función `summary` de R permite describir la base de datos donde nos encontramos con los siguientes atributos:

- **encounter_id:** Corresponde a la id de visita (variable nominal).
- **patient_nbr:** Corresponde a la id de cada paciente (variable nominal).
- **race:** Corresponde a la raza de cada paciente (variable categórica).
- **age:** Corresponde al rango de edad de cada paciente (variable categórica).
- **weight:** Corresponde al peso de cada paciente (variable numérica).
- **admission_type_id:** Corresponde a un identificador de admisión, por ejemplo 1 si fue por emergencia, 2 si fue urgencia, etc. (variable categórica).
- **discharge_disposition_id:** Corresponde a un identificador de derivación, por ejemplo, si fue derivado a casa, a otro hospital, etc. (variable categórica).
- **admission_source:** Corresponde a un identificador de donde viene el paciente, por ejemplo, si viene de otro hospital, del área de emergencias, etc. (variable categórica).
- **time_in_hospital:** Días entre que el paciente llega y se va (variable numérica).
- **payer_code:** Corresponde al método de pago del paciente (variable categórica).
- **medical_speciality:** Especialidad del medico que atiende al paciente (variable categórica).
- **num_lab_procedures:** Número de tests de laboratorios realizados en la consulta (variable numérica).
- **num_procedures:** Números de tests que no son de laboratorio realizados en la consulta (variable numérica).
- **num_medications:** Número de medicamentos suministrados en el encuentro (variable numérica),
- **number_outpatient:** Número de visitas del paciente en el año anterior (variable numérica).
- **number_emergency:** Número de visitas de emergencia del paciente en el año anterior (variable numérica).
- **number_inpatient:** Número de visitas hospitalarias del paciente en el año anterior (variable numérica).
- **diag1 diag2 diag3:** Corresponde al código ICD9 (diagnósticos de enfermedades internacionales), por ejemplo, V56 corresponde a diálisis renal (variable categórica).
- **number_diagnoses:** Número de diagnósticos ICD9 del paciente (variable numérica).
- **max_glu_serum:** Indica el resultado del test, si fue normal, mayor a 200 o si fue mayor a 300 (variable categórica).

- **A1Cresult:** Indica el resultado del test, si fue normal, si fue mayor a 8% o mayor a 7% (variable categórica).
- **change:** Indica si hubo cambios en los medicamentos suministrados al paciente (variable categórica).
- **readmitted:** Rango de reingreso al hospital. (variable categórica).
- **otras:** Indican si hubo cambios en las dosis de 24 medicamentos (variables categóricas).

ELIMINACION DE VARIABLES EXCEDENTES DE MISSING VALUES

Se eliminan variables que no aportan información, esto es, variables que poseen más del 30% de datos perdidos en sus registros. Bajo este criterio se elimina un total de 5 atributos (weight, payer_code, medical_speciality, max_glu_serum y AICresult)

TRATAMIENTO DE DATOS ATIPICOS

METODO DE DESVIACION ESTANDAR

Se eliminarán aquellos atributos numéricos que son prácticamente idénticos, esto es, atributos que poseen una desviación estándar menor a 0,1. Bajo este criterio se eliminan 3 variables (number_outpatient, number_emergency y number_inpatient).

METODO DE COEFICIENTE DE VARIACION

Se eliminarán aquellos atributos numéricos que poseen un coeficiente de variación menor al 30%. Bajo este criterio correspondería eliminar la variable number_diagnosis (coeficiente de variación 0.27), sin embargo, no se eliminará ya que a criterio del investigador este puede ser un atributo numérico que de alguna manera represente las variables categóricas diag1, diag2 y diag3, (variables las cuales quedarán fuera del ACP y se consideran importantes).

METODO DE LA ENTROPIA

Este método permite eliminar variables categóricas que poseen muchos datos repetidos o con un grado de desorden muy bajo. En primera instancia se eliminaron aquellas variables que poseen entropías menores al 50%, pero luego de realizar esta limpieza, se ha realizado una inspección manual en la cual aún se pueden observar variables con un grado de desorden bajo. Lo anterior es por resultado de que existen variables que poseen más de dos categorías, por lo que las entropías podrían superar el 100%.

Producto de que casi no existen variables con solo dos categorías, se decide ser más exigente y eliminar las entropías menores al 70% (tomando las precauciones correspondientes) y además calcular las probabilidades de cada categoría en los distintos atributos. Si la probabilidad de una categoría es mayor al 60% se eliminará.

Por entropía se eliminan 21 variables y por probabilidades 4 variables (race, id_discharge_disposition, metmorfin y readmitted).

Ilustración 1: Ejemplo probabilidades atributo race.

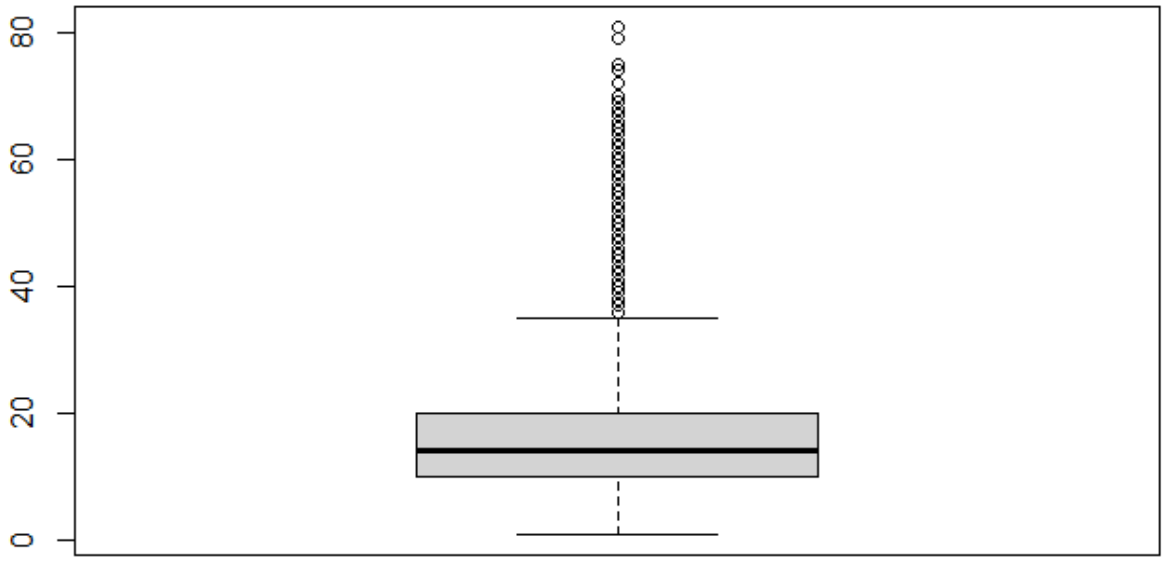
AfricanAmerican	Asian	Caucasian	Hispanic	Other
0.185237890	0.007143884	0.768880264	0.021805376	0.016932586

Fuente: Elaboración propia.

METODO DE LOS CUARTILES

Se utilizarán los gráficos de caja y bigotes (bplot) para eliminar aquellos valores de las variables numéricas que se encuentran fuera de rango y que podrían perjudicar el estudio. Estos valores serán cambiados por missing values y posteriormente imputados por método knn.

Ilustración 2: Ejemplo boxplot variable num_medications.



Fuente: Elaboración propia.

TRATAMIENTO DE MISSING VALUES

Se imputarán por método KNN los 19.680 datos perdidos presentes en la base de datos hasta este entonces, utilizando los 10 vecinos más cercanos. Luego de esto, se tendrá la base de datos limpia y completa.

CORRELACIONES ENTRE VARIABLES

La matriz de correlaciones permitirá analizar el grado en que las variables numéricas podrían explicar a las otras variables numéricas. Por ende, se creará un data frame solo con variables numéricas que permitirá realizar este estudio y posteriormente el ACP.

No existen relaciones inversas entre las variables, esto es, a medida que una variable aumenta, las otras también lo harán. Sin embargo, no existen correlaciones fuertes, esto es, el grado en que cada variable explica a las otra no es alto.

Si la correlación entre una variable y otra es menor a 0.3, se dirá que la correlación es débil, en el otro caso, si se encuentra entre 0.3 y 0.8 se dirá que es una correlación moderada.

Ilustración 3: Matriz de correlaciones.

	time_in_hospital	num_lab_procedures	num_procedures	num_medications	number_diagnoses
time_in_hospital	1.00	0.33	0.17	0.43	0.23
num_lab_procedures	0.33	1.00	0.02	0.24	0.16
num_procedures	0.17	0.02	1.00	0.30	0.06
num_medications	0.43	0.24	0.30	1.00	0.29
number_diagnoses	0.23	0.16	0.06	0.29	1.00

Fuente: Elaboración propia.

ANALISIS DE COMPONENTES PRINCIPALES

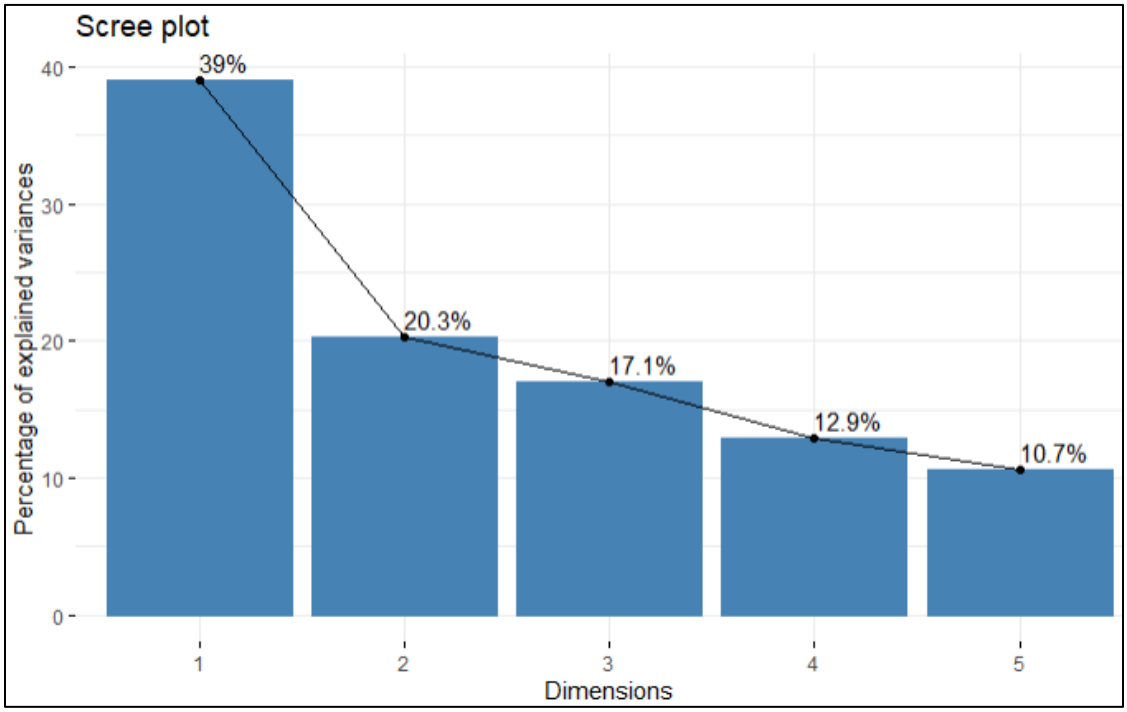
TEST DE BARTLETT

El test de Bartlett aplicado a la matriz de correlaciones arroja como resultado un p-value= 9,5453e-08, por lo tanto, se rechaza la hipótesis H0: Las muestras presentan varianzas iguales. Esto significa que existe la correlación suficiente entre las variables para poder realizar el ACP.

APLICACION DEL ACP

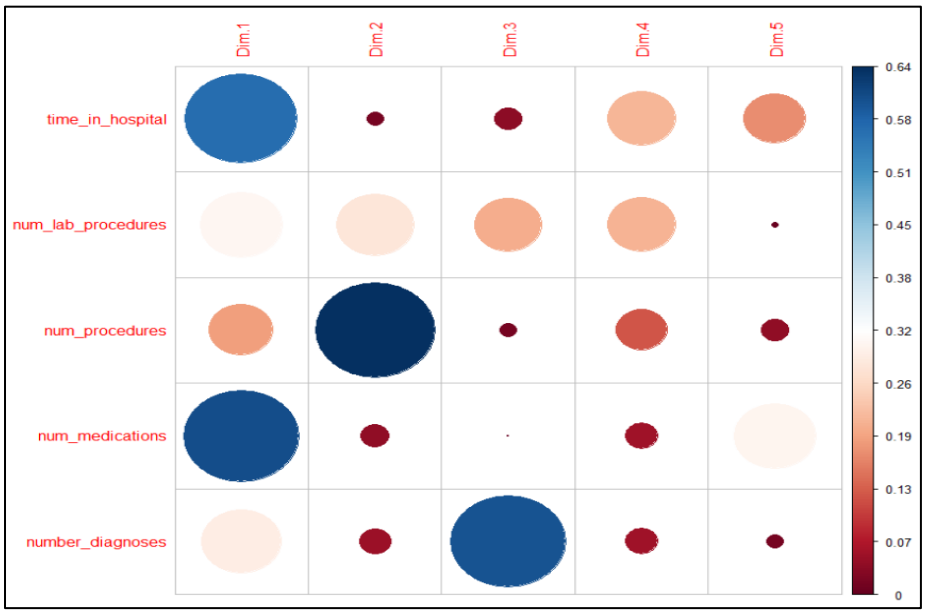
El ACP permite reducir las dimensiones a aquellas que explican en mejor forma la varianza.

Ilustración 4: Porcentaje de la variabilidad explicada por dimensión.



Fuente: Elaboración propia.

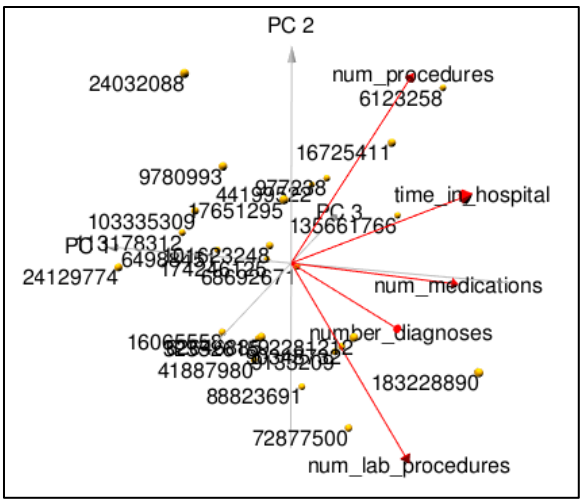
Ilustración 5: Explicación de las variables por dimensión.



Fuente: Elaboración propia.

Por el método del codo y apoyándose también en el grafico de explicación de las variables por dimensión, se toma la decisión de incorporar al estudio las primeras 3 dimensiones dado que de esta forma se tiene en gran porcentaje la explicación de la varianza.

Ilustración 6: PCA 3D.



Fuente: Ilustración propia.

De lo anterior, si el ángulo entre las variables es cercano a 90° significa que no existe una correlación entre ellas, si es cercana a 0° significa que poseen una correlación directa y si es cercana a 180° significa que la relación es inversa. Cada individuo representado en el

PCA 3D estará graficado siguiendo las mismas reglas anteriores con las variables existentes.

REPRESENTACION DE LAS VARIABLES

El ACP con las 3 primeras dimensiones arroja como resultado que todas las variables son bien representadas, por lo cual, no se eliminará ninguna bajo este criterio.

Ilustración 7: Ranking variables.

	Dim.1	Dim.2	Dim.3	Representacion	Bien_o_Mal
number_diagnoses	0.2908744	0.04762076	0.596699835	93.51950	BIEN
num_procedures	0.1871977	0.63880854	0.015588900	84.15952	BIEN
num_lab_procedures	0.3056494	0.27603944	0.204908602	78.65975	BIEN
num_medications	0.6032866	0.03956566	0.001389674	64.42419	BIEN
time_in_hospital	0.5632110	0.01507565	0.035442192	61.37289	BIEN

Fuente: Elaboración propia.

REPRESENTACION DE LOS INDIVIDUOS

Del mismo modo, el ranking de la representación de los individuos en 3 dimensiones se puede visualizar en la siguiente ilustración:

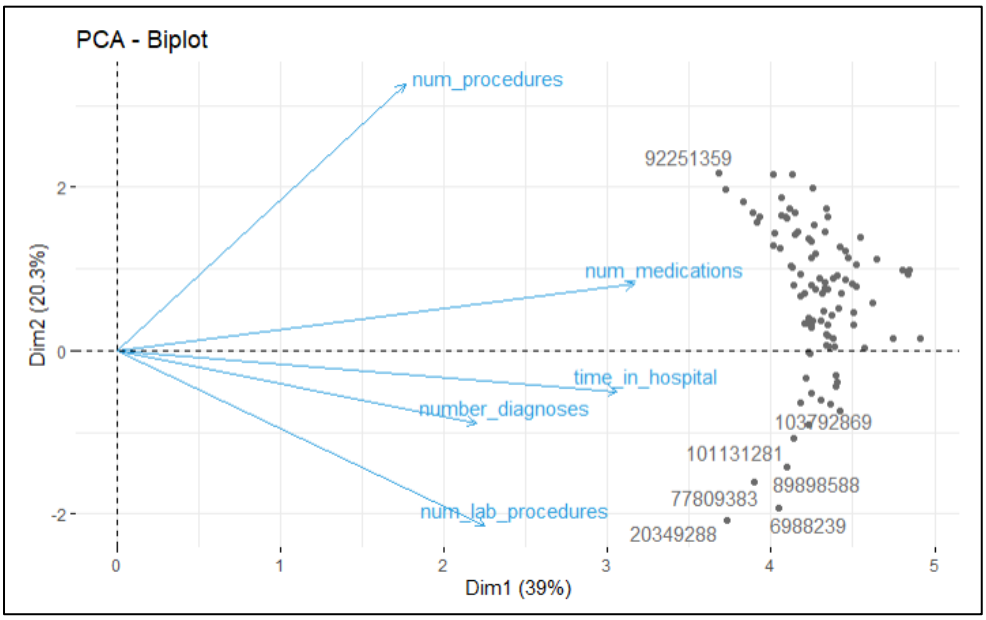
Ilustración 8: Ranking mejores y peores 5 individuos representados en PCA 3D.

	Dim.1	Dim.2	Dim.3	Representacion	Bien_o_Mal
32950737	0.0040783577	0.5895540828	4.063572e-01	99.9989664	BIEN
93869154	0.5997849360	0.0771190981	3.230818e-01	99.9985870	BIEN
88531686	0.8995169539	0.0921627568	8.302785e-03	99.9982496	BIEN
4670703	0.4023423763	0.4114335509	1.862058e-01	99.9981727	BIEN
75268737	0.4023423763	0.4114335509	1.862058e-01	99.9981727	BIEN
96359850	0.0031942226	0.0038031711	1.978711e-05	0.7017181	MAL
86187249	0.0008113810	0.0019867481	3.370175e-03	0.6168304	MAL
749196	0.0041952723	0.0016335625	1.205268e-05	0.5840888	MAL
44594892	0.0041952723	0.0016335625	1.205268e-05	0.5840888	MAL
98462592	0.0000331360	0.0007321003	4.437846e-03	0.5203083	MAL
31516434	0.0002021574	0.0010393508	3.749968e-03	0.4991476	MAL

Fuente: Elaboración propia

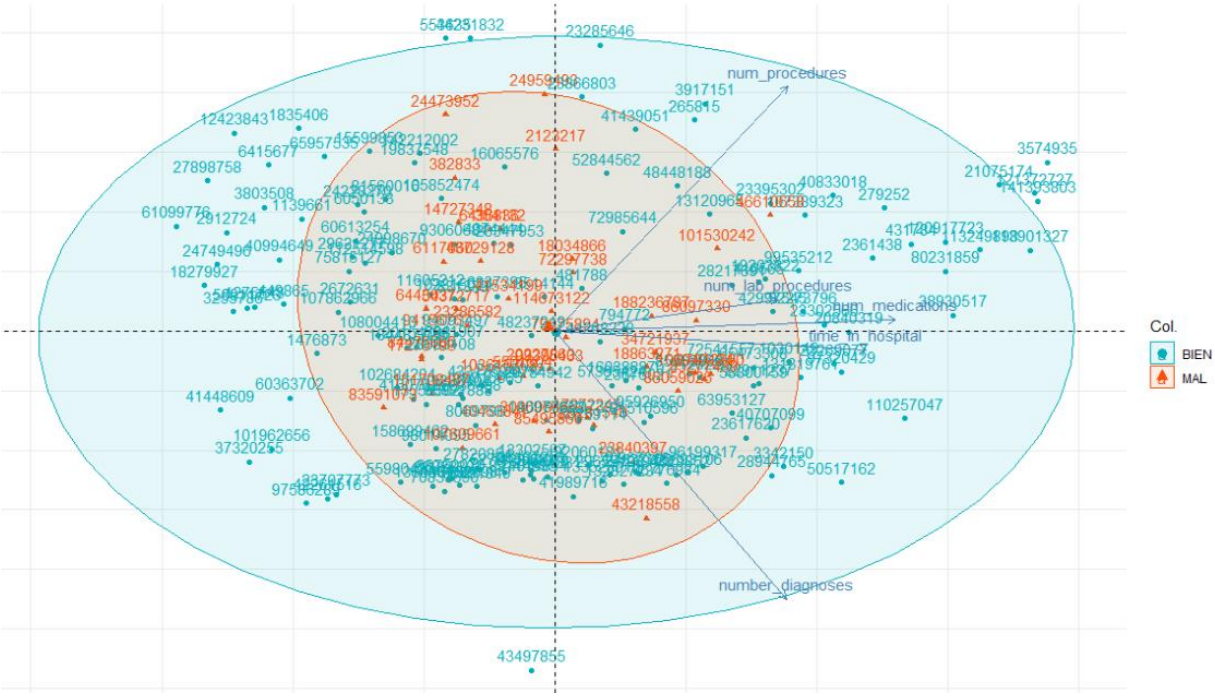
En el siguiente gráfico se puede visualizar a los 100 individuos que más contribuyen en el PCA en 2D, en donde, si el ángulo que forman los vectores de las variables es de 90°, significa que aquellas variables no tienen correlación, si el ángulo es de 180° significa que la correlación es inversa y si es más cercana a 0° significa que poseen una correlación directa.

Ilustración 9: 100 individuos mejor representados en PCA 2D.



Fuente: Elaboración propia.

Ilustración 10: Representación bidimensional de 200 individuos al azar (PCA 3D).



Fuente: Elaboración propia.

Todos aquellos individuos que están mal representados por las primeras tres dimensiones serán eliminados para el estudio de cluster. Por lo tanto, el número de pacientes que servirán para encontrar patrones en el data set es 54.946.

CLUSTERS

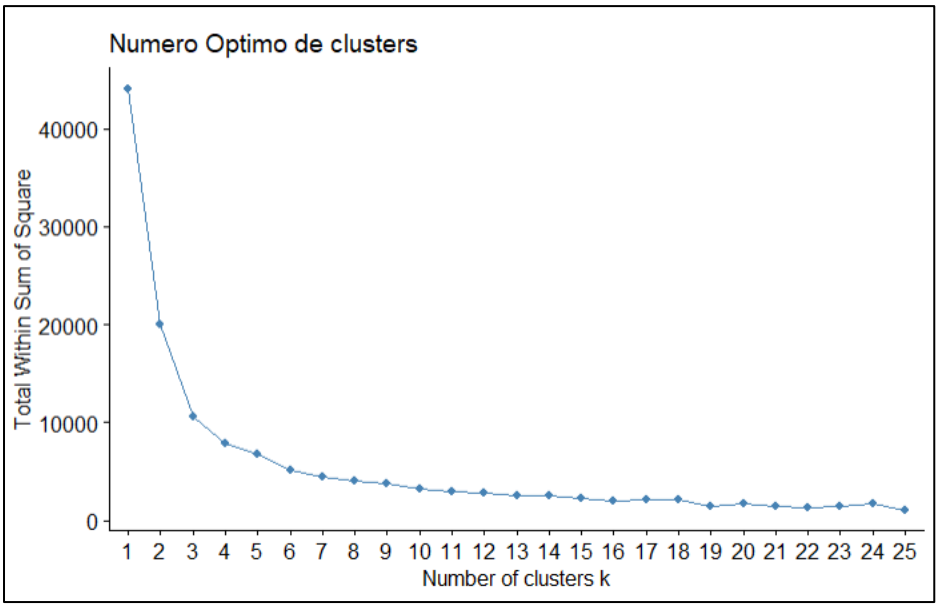
Para realizar el estudio de clusters, se considerará una muestra de 100 individuos que están bien representados para efectos de obtener gráficos que se puedan visualizar de manera óptima y para no incurrir en errores por almacenamiento en clusters como el mixto que es más robusto en recursos computacionales.

Cabe destacar que se utilizará la misma muestra en todos los clusters,, ya sean para variables numéricas o categóricas (solo considerando las variables que se requieran).

SELECCIÓN DEL NUMERO OPTIMO DE CLUSTERS

Para seleccionar el numero óptimo de clusters se utilizará el método de la matriz de ruta mínima (WSS- Total Within Sum of Square), en el cual se puede visualizar el quiebre a partir del número de cluster 4 o 5.

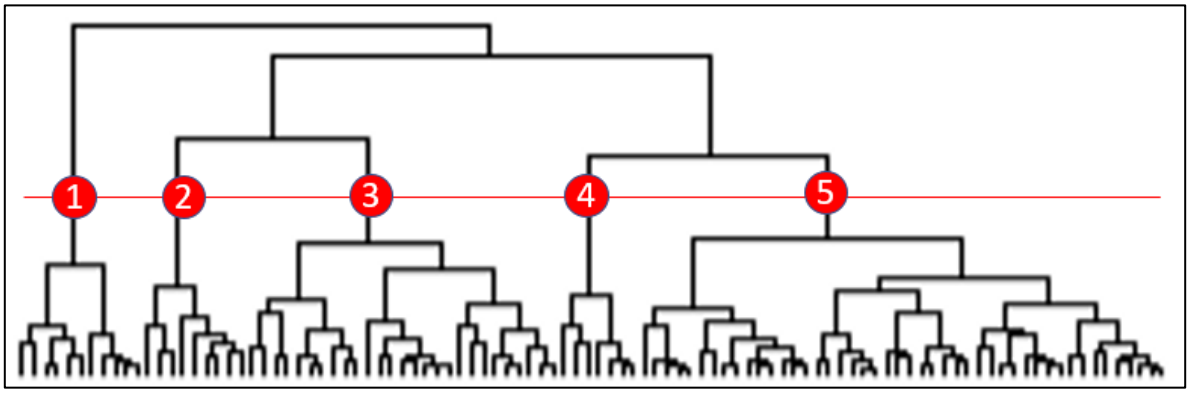
Ilustración 11: Selección del número de clusters – WSS.



Fuente: Elaboración propia.

Apoyándose en el mapa de calor de la matriz de distancia euclidiana, se puede confirmar la decisión de utilizar 5 clusters.

Ilustración 12: Selección del número de clusters – heatmap.

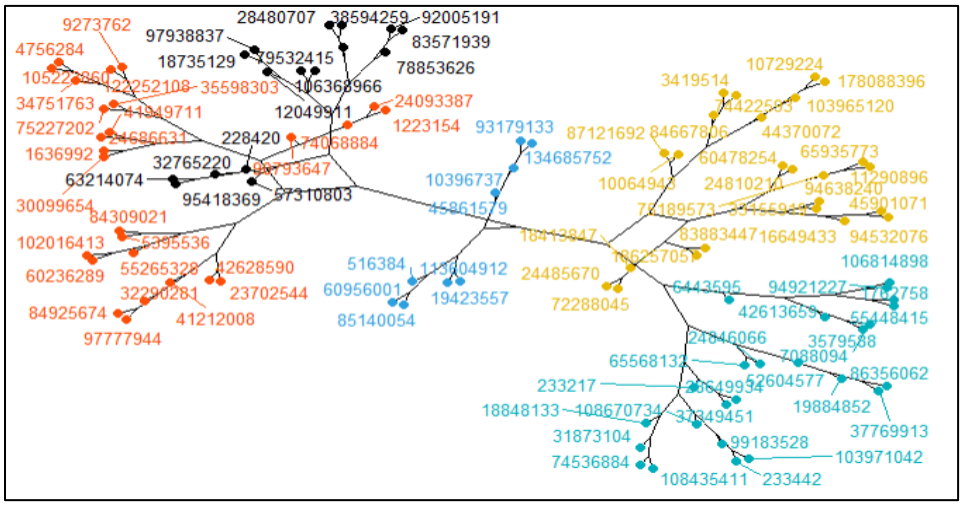


Fuente: Elaboración propia.

CLUSTER JERARQUICO VARIANZA MINIMA DE WARD

Este método busca agrupar los casos minimizando la varianza dentro de cada cluster (de la Fuente). El resultado del cluster en nuestra muestra de 100 pacientes es el siguiente:

Ilustración 13: Método de Ward

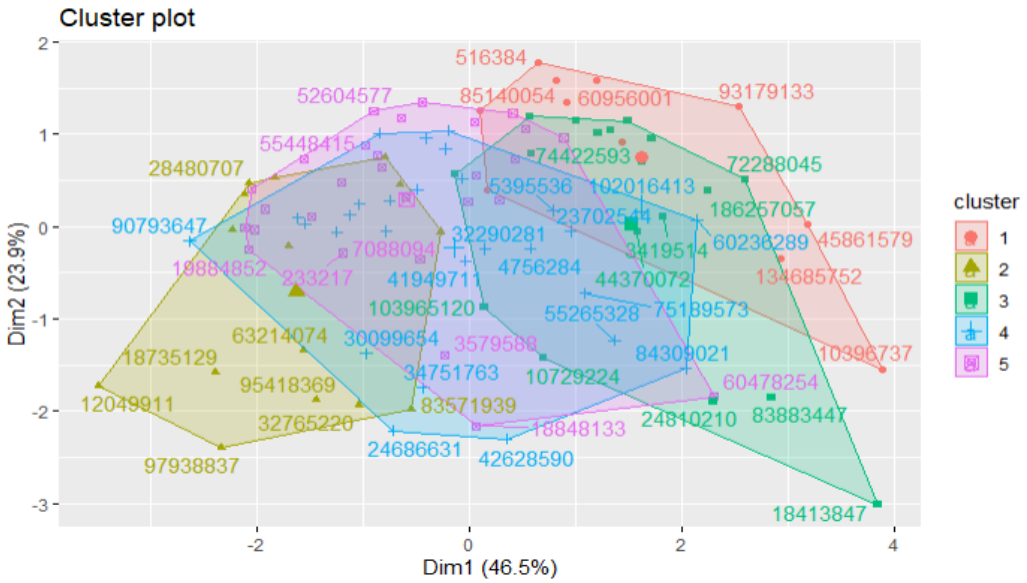


Fuente: Elaboración propia.

CLUSTER NO JERARQUICO – KMEAN

La agrupación basándose en la minimización de la suma de distancias entre cada individuo y el centroide se verá en la última tabla de resumen. Se puede observar la composición de cada cluster en donde el centroide posee un tamaño más grande que los demás.

Ilustración 14: Método K-Mean.

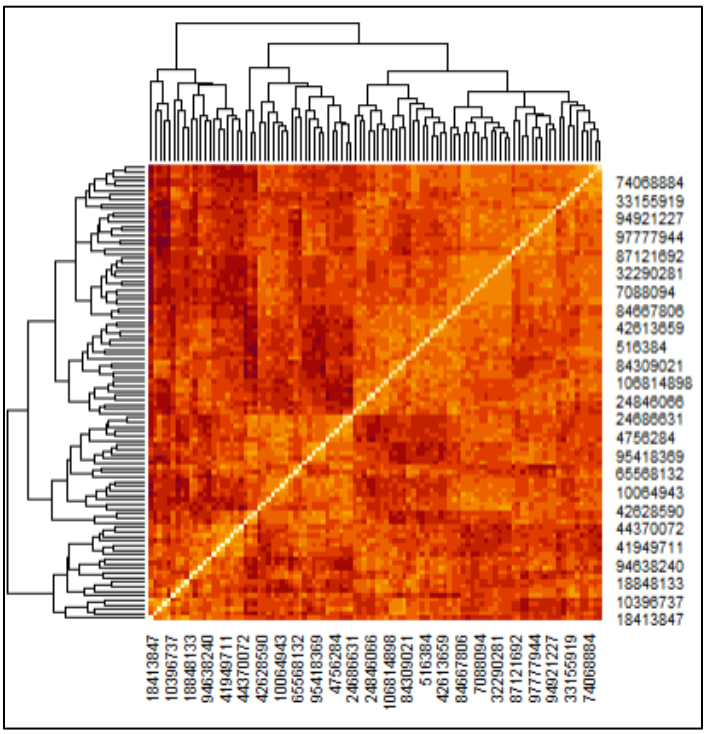


Fuente: Elaboración propia.

CLUSTER NO JERARQUICO – MIXTO

Dado que la base de datos posee en su mayoría variables categóricas, se cree (sin ver los resultados) que este será el cluster que mejor podrá describir patrones para identificar pacientes con un potencial diagnóstico de diabetes. En el gráfico se puede observar la agrupación de individuos bajo este método, para los cuales cuando el cuadrado de las medias es más grande, más oscuro será en el gráfico. Este método buscará agrupar en cada cluster a aquellos individuos que se encuentren más cerca.

Ilustración 15: Método Mixto.



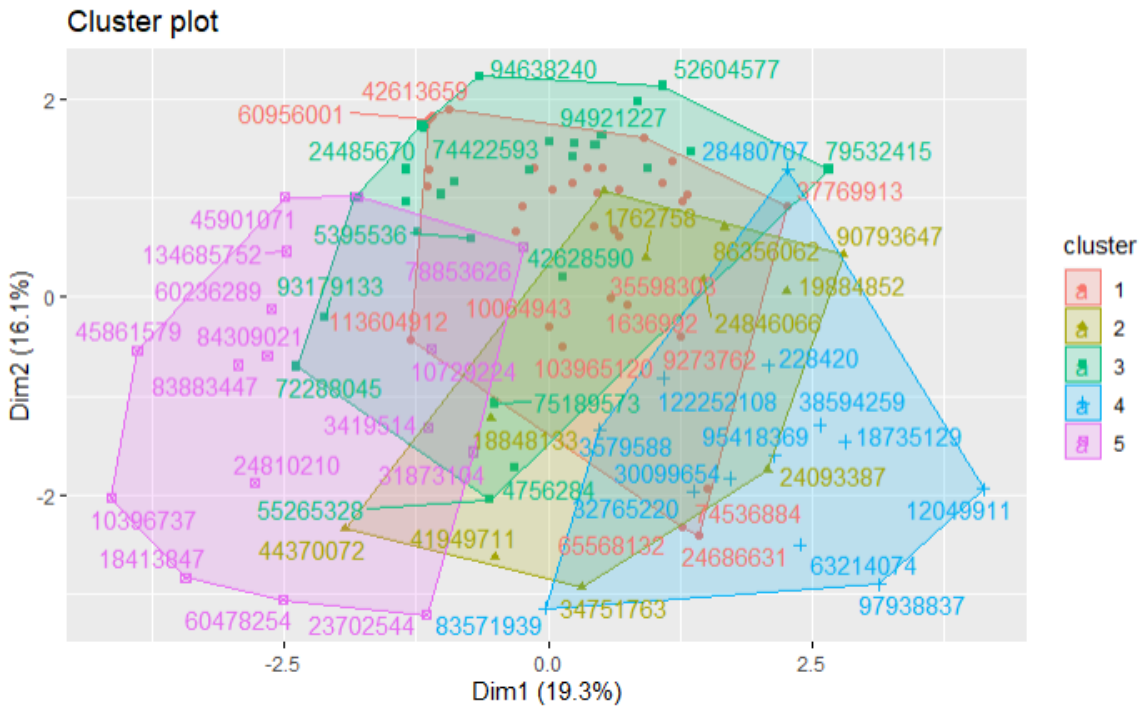
Fuente: Elaboración propia.

CLUSTER NO VISTO EN CLASES – CLUSTER CLARA

Este método, es un cluster no jerárquico y permite considerar bases de datos muy grandes. Este método aplica el algoritmo de los K-Medoids (minimiza la distancia entre puntos que se añadirán a un grupo con el centro del mismo grupo) para encontrar clusters óptimos acorde a muestras (Amat, 2017)

La agrupación de individuos gráficamente se puede ver a continuación:

Ilustración 16: Método CLARA.



Fuente: Elaboración propia.

RESULTADOS Y CONCLUSIONES

Para visualizar los resultados, se ha generado una tabla incorporando en nuestra muestra la variable label a predecir y los resultados de cada cluster. De esta forma se analizará cual es el método que mejor agrupa a los pacientes que padecen o pueden padecer diabetes.

Ilustración 17: Extracto tabla clusters versus label.

	cluster_kmean	cluster_mixto	cluster_clara	diabetes_med
18119214	3	3	1	Yes
9107163	3	1	2	No
23310684	1	4	1	Yes
68524299	3	2	3	No
1927881	3	3	4	Yes
55959237	1	2	3	No
86400099	3	5	1	No
90761598	5	3	2	Yes
34176501	5	5	1	Yes
50571495	2	3	4	Yes

Fuente: Elaboración propia.

Se debe determinar cuales son los clusters que mejor aplican para la data diabetes, para ello se ha realizado una tabla de agrupaciones en donde se compara si el paciente es medicado con diabetes versus todos los clusters asignados por cada método.

Ilustración 18: Comparación de clusters.

```
--------Summary descriptives table by 'diabetes_med'

_____________________________________________________
                    No          Yes       p.overall
                   N=22         N=78
_____________________________________________________
cluster_kmean:                                 0.064
    1           2 (9.09%)   9 (11.5%)
    2           2 (9.09%)   26 (33.3%)
    3           9 (40.9%)   19 (24.4%)
    4           3 (13.6%)   15 (19.2%)
    5           6 (27.3%)   9 (11.5%)
cluster_mixto:                                <0.001
    1           13 (59.1%)  17 (21.8%)
    2           0 (0.00%)   18 (23.1%)
    3           9 (40.9%)   11 (14.1%)
    4           0 (0.00%)   16 (20.5%)
    5           0 (0.00%)   16 (20.5%)
cluster_clara:                                 0.006
    1           7 (31.8%)   25 (32.1%)
    2           0 (0.00%)   11 (14.1%)
    3           10 (45.5%)  18 (23.1%)
    4           5 (22.7%)   8 (10.3%)
    5           0 (0.00%)   16 (20.5%)
_____________________________________________________
```

Fuente: Elaboración propia.

El análisis anterior señala que los clusters que mejor predicen a un paciente con diabetes son la partición 2, 4 y 5 del mixto y la 2 y 4 del clara dado que en aquellos clusters se han agrupado **SOLO** individuos que padecen diabetes. Además, el cluster k-mean entrega en el cluster 2 un patrón mucho más flexible que los dos anteriores pero que al validar en la data original, se concluye que logra predecir con éxito el 78% de los casos de pacientes que poseen diabetes.

A partir de lo anterior, se analiza como inciden las variables del data set en cada uno de los clusters mencionados:

Ilustración 19: Patrones cluster k-mean.

--------Summary descriptives table by 'cluster_kmean'---------

	1 N=28	2 N=16	3 N=29	4 N=14	5 N=13	p.overall
time_in_hospital	-0.41 (0.59)	1.18 (0.87)	-0.40 (0.64)	-0.37 (0.49)	-0.71 (0.35)	<0.001
num_lab_procedures	-0.10 (0.83)	1.10 (0.61)	0.63 (0.63)	-0.67 (0.76)	-1.20 (0.72)	<0.001
num_procedures	-0.71 (0.35)	0.49 (0.86)	-0.74 (0.27)	1.53 (0.90)	0.53 (0.94)	<0.001
num_medications	-0.26 (0.49)	1.15 (0.78)	-0.72 (0.54)	0.82 (0.77)	-0.37 (0.70)	<0.001
number_diagnoses	0.62 (0.35)	0.86 (0.13)	-1.09 (0.60)	0.49 (0.80)	-1.14 (0.62)	<0.001

Fuente: Elaboración propia.

De la aplicación del cluster k-mean se puede apreciar que un patrón de variables que describe a pacientes que podrían tener diabetes es:

- Individuos que poseen más de 1 día en el hospital.
- Individuos que poseen más de 1 procedimiento de laboratorio.
- Individuos que poseen más de 1 procedimiento que no es de laboratorio.
- Individuos que poseen más de 1 medicación.
- Individuos con al menos un diagnóstico del ICD 9.

Ilustración 20: Patrones cluster mixto.

--------Summary descriptives table by 'cluster_mixto'---------

	1 N=30	2 N=18	3 N=20	4 N=16	5 N=16	p.overall
gender:						<0.001
Female	21 (70.0%)	0 (0.00%)	11 (55.0%)	16 (100%)	8 (50.0%)	
Male	9 (30.0%)	18 (100%)	9 (45.0%)	0 (0.00%)	8 (50.0%)	
age:						.
[10-20)	0 (0.00%)	1 (5.56%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	
[20-30)	0 (0.00%)	1 (5.56%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	
[30-40)	1 (3.33%)	1 (5.56%)	2 (10.0%)	0 (0.00%)	1 (6.25%)	
[40-50)	0 (0.00%)	2 (11.1%)	5 (25.0%)	3 (18.8%)	2 (12.5%)	
[50-60)	5 (16.7%)	4 (22.2%)	2 (10.0%)	2 (12.5%)	4 (25.0%)	
[60-70)	3 (10.0%)	3 (16.7%)	4 (20.0%)	5 (31.2%)	3 (18.8%)	
[70-80)	9 (30.0%)	5 (27.8%)	4 (20.0%)	3 (18.8%)	2 (12.5%)	
[80-90)	9 (30.0%)	1 (5.56%)	2 (10.0%)	3 (18.8%)	4 (25.0%)	
[90-100)	3 (10.0%)	0 (0.00%)	1 (5.00%)	0 (0.00%)	0 (0.00%)	
admission_type_id:						.
1	27 (90.0%)	16 (88.9%)	0 (0.00%)	16 (100%)	0 (0.00%)	
2	3 (10.0%)	2 (11.1%)	9 (45.0%)	0 (0.00%)	7 (43.8%)	
3	0 (0.00%)	0 (0.00%)	10 (50.0%)	0 (0.00%)	7 (43.8%)	
5	0 (0.00%)	0 (0.00%)	1 (5.00%)	0 (0.00%)	2 (12.5%)	
admission_source_id:						<0.001
1	0 (0.00%)	0 (0.00%)	18 (90.0%)	0 (0.00%)	12 (75.0%)	
2	1 (3.33%)	0 (0.00%)	1 (5.00%)	0 (0.00%)	0 (0.00%)	
3	0 (0.00%)	1 (5.56%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	
4	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (18.8%)	
5	1 (3.33%)	0 (0.00%)	1 (5.00%)	0 (0.00%)	0 (0.00%)	
6	1 (3.33%)	2 (11.1%)	0 (0.00%)	1 (6.25%)	0 (0.00%)	
7	27 (90.0%)	14 (77.8%)	0 (0.00%)	15 (93.8%)	1 (6.25%)	
10	0 (0.00%)	1 (5.56%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	
time_in_hospital	3.17 (1.51)	3.67 (1.81)	2.65 (1.76)	4.12 (2.28)	5.75 (3.07)	<0.001
num_lab_procedures	41.2 (18.6)	51.4 (12.6)	32.1 (19.8)	48.9 (21.4)	52.1 (20.9)	0.006
num_procedures	0.50 (0.90)	0.89 (1.64)	1.55 (1.36)	0.38 (1.02)	1.94 (1.84)	0.002
num_medications	12.5 (5.49)	12.9 (7.19)	12.0 (5.79)	13.4 (5.70)	17.6 (6.35)	0.060
number_diagnoses	7.57 (1.65)	6.94 (2.21)	6.20 (2.21)	7.62 (1.50)	7.06 (1.77)	0.109
insulin:						.
Down	0 (0.00%)	4 (22.2%)	0 (0.00%)	5 (31.2%)	3 (18.8%)	
No	23 (76.7%)	3 (16.7%)	17 (85.0%)	3 (18.8%)	0 (0.00%)	
Steady	7 (23.3%)	11 (61.1%)	3 (15.0%)	3 (18.8%)	4 (25.0%)	
Up	0 (0.00%)	0 (0.00%)	0 (0.00%)	5 (31.2%)	9 (56.2%)	
change:						<0.001
Ch	0 (0.00%)	14 (77.8%)	0 (0.00%)	16 (100%)	16 (100%)	
No	30 (100%)	4 (22.2%)	20 (100%)	0 (0.00%)	0 (0.00%)	

Fuente: Elaboración propia.

De la aplicación del cluster mixto se puede apreciar que un patrón de variables que describe a pacientes que podrían tener diabetes es:

- Individuos independientes de su género mayores de 40 años.
- Que ingresan por emergencia, urgencia o por elección.
- Que ingresan a través del área de emergencias o por referencia médica.
- Que poseen 4 o más días en el hospital.
- Que se les aplican más de 49 procedimientos de laboratorios.
- Que poseen cambios en la insulina o que poseen insulina alta.
- Que poseen cambios en su medicación.

Ilustración 20: Patrones cluster CLARA.

--------Summary descriptives table by 'cluster_clara'---------

	1 N=32	2 N=11	3 N=28	4 N=13	5 N=16	p.overall
gender:						<0.001
Female	25 (78.1%)	1 (9.09%)	18 (64.3%)	3 (23.1%)	9 (56.2%)	
Male	7 (21.9%)	10 (90.9%)	10 (35.7%)	10 (76.9%)	7 (43.8%)	
age:						.
[10-20)	0 (0.00%)	1 (9.09%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	
[20-30)	0 (0.00%)	1 (9.09%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	
[30-40)	0 (0.00%)	2 (18.2%)	1 (3.57%)	2 (15.4%)	0 (0.00%)	
[40-50)	2 (6.25%)	2 (18.2%)	2 (7.14%)	3 (23.1%)	3 (18.8%)	
[50-60)	1 (3.12%)	4 (36.4%)	8 (28.6%)	1 (7.69%)	3 (18.8%)	
[60-70)	7 (21.9%)	0 (0.00%)	6 (21.4%)	1 (7.69%)	4 (25.0%)	
[70-80)	10 (31.2%)	0 (0.00%)	7 (25.0%)	2 (15.4%)	4 (25.0%)	
[80-90)	8 (25.0%)	1 (9.09%)	4 (14.3%)	4 (30.8%)	2 (12.5%)	
[90-100)	4 (12.5%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	
admission_type_id:						.
1	22 (68.8%)	6 (54.5%)	22 (78.6%)	1 (7.69%)	8 (50.0%)	
2	7 (21.9%)	1 (9.09%)	3 (10.7%)	5 (38.5%)	5 (31.2%)	
3	1 (3.12%)	4 (36.4%)	3 (10.7%)	7 (53.8%)	2 (12.5%)	
5	2 (6.25%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (6.25%)	
admission_source_id:						.
1	7 (21.9%)	4 (36.4%)	4 (14.3%)	11 (84.6%)	4 (25.0%)	
2	1 (3.12%)	0 (0.00%)	0 (0.00%)	1 (7.69%)	0 (0.00%)	
3	1 (3.12%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	
4	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	3 (18.8%)	
5	0 (0.00%)	0 (0.00%)	2 (7.14%)	0 (0.00%)	0 (0.00%)	
6	2 (6.25%)	1 (9.09%)	0 (0.00%)	0 (0.00%)	1 (6.25%)	
7	21 (65.6%)	6 (54.5%)	21 (75.0%)	1 (7.69%)	8 (50.0%)	
10	0 (0.00%)	0 (0.00%)	1 (3.57%)	0 (0.00%)	0 (0.00%)	
time_in_hospital	3.22 (1.48)	3.09 (2.59)	3.93 (1.76)	1.85 (0.99)	6.31 (2.63)	<0.001
num_lab_procedures	44.6 (20.0)	45.2 (9.15)	50.6 (16.1)	19.1 (16.3)	51.9 (19.1)	<0.001
num_procedures	0.31 (0.74)	0.73 (1.42)	0.64 (1.03)	1.77 (1.42)	2.50 (1.86)	<0.001
num_medications	10.6 (3.66)	8.73 (4.31)	16.5 (5.41)	8.92 (3.97)	20.6 (5.23)	<0.001
number_diagnoses	6.91 (1.87)	5.18 (1.40)	8.18 (1.19)	5.54 (1.81)	8.25 (1.44)	<0.001
insulin:						.
Down	4 (12.5%)	2 (18.2%)	4 (14.3%)	0 (0.00%)	2 (12.5%)	
No	17 (53.1%)	1 (9.09%)	18 (64.3%)	10 (76.9%)	0 (0.00%)	
Steady	11 (34.4%)	6 (54.5%)	5 (17.9%)	2 (15.4%)	4 (25.0%)	
Up	0 (0.00%)	2 (18.2%)	1 (3.57%)	1 (7.69%)	10 (62.5%)	
change:						<0.001
Ch	12 (37.5%)	9 (81.8%)	8 (28.6%)	1 (7.69%)	16 (100%)	
No	20 (62.5%)	2 (18.2%)	20 (71.4%)	12 (92.3%)	0 (0.00%)	

Fuente: Elaboración propia.

De la aplicación del cluster clara se puede apreciar que un patrón de variables que describe a pacientes que podrían tener diabetes es:

- Individuos principalmente mujeres que ingresan por emergencia, urgencia o por elección.
- Que llegan a través de referencias médicas, del área de emergencia
- Que poseen más de 19 procedimientos de laboratorios.
- Que poseen más de 1 procedimiento que no son de laboratorio.
- Con cambios en la insulina o sin tenerla medicada.

Para validar el patrón sugerido a partir del cluster mixto, se comparará la predicción con el valor real de la data set original. Se tiene que, para 1697 individuos de las características descritas anteriormente, el patrón predice bien el 100% de los casos. De igual manera se realizó para el patrón sugerido por el cluster k-mean, entregando un acierto del 78% de los casos pero que es más flexible al solo incorporar variables numéricas.

Ilustración 21: Validación del patrón sugerido con cluster mixto.

	age	admission_type_id	admission_source_id	time_in_hospital	num_lab_procedures	insulin	change	diabetesMed
1	[60-70)	3	2	4	70	Steady	Ch	Yes
2	[50-60)	3	2	6	65	Up	Ch	Yes
3	[60-70)	3	2	4	49	Steady	Ch	Yes
4	[50-60)	3	2	6	69	Up	Ch	Yes
5	[40-50)	2	2	7	58	Steady	Ch	Yes
6	[60-70)	2	2	6	75	Steady	Ch	Yes
7	[50-60)	3	2	8	52	Up	Ch	Yes
8	[70-80)	2	2	6	74	Up	Ch	Yes
9	[60-70)	3	2	10	66	Steady	Ch	Yes
10	[70-80)	2	2	11	78	Steady	Ch	Yes
11	[70-80)	3	1	8	62	Up	Ch	Yes
12	[60-70)	2	1	12	95	Steady	Ch	Yes
13	[50-60)	2	1	7	59	Steady	Ch	Yes
14	[70-80)	2	1	6	70	Up	Ch	Yes
15	[70-80)	3	1	4	55	Steady	Ch	Yes
16	[40-50)	3	1	4	79	Up	Ch	Yes
17	[50-60)	3	1	4	54	Steady	Ch	Yes
18	[60-70)	3	1	8	70	Steady	Ch	Yes
19	[70-80)	2	2	4	76	Steady	Ch	Yes
20	[70-80)	1	1	9	79	Steady	Ch	Yes
21	[60-70)	3	1	6	74	Steady	Ch	Yes
22	[50-60)	3	2	5	65	Steady	Ch	Yes
23	[50-60)	1	1	5	55	Steady	Ch	Yes

Fuente: Elaboración propia.

Con lo descrito anteriormente, se logra dar con 3 patrones que pueden predecir si un paciente tiene riesgo de poseer diabetes, por lo tanto, se da cumplimiento al objetivo general.

BIBLIOGRAFIA

Amat, J. (Septiembre de 2017). *Clustering y heatmaps: aprendizaje no supervisado*. Obtenido de Ciencia de Datos: https://www.cienciadedatos.net/documentos/37_clustering_y_heatmaps

de la Fuente, S. (s.f.). *Analisis de Conglomerados - Análisis Cluster.* Universidad Autónoma de Madrid, Madrid.

Gomez, D. (05 de 2022). clase "Análisis EDA". *Exploratory Data Analysis*.