Centro de Investigación en Matemáticas (CIMAT). Unidad Monterrey

Tarea 2

Marcelo Alberto Sanchez Zaragoza

26 de febrero de 2021

1. Problema 1

Considera una matriz de datos X_{nxd}. PCA puede formularse también como el problema de encontrar un subespacio (ortonormal) de baja dimensión de forma tal que se minimicen los errores de las proyecciones de los datos en tal subespacio.
Si consideramos una base ortonormal u_j, j=1,...,d, ya vimos que una observación x_i puede expresarse como una combinación lineal

$$x_i = \sum_{j=1}^{d} \alpha_{ij} u_j$$

Por la ortogonalidad de u_j, podemos expresar $\alpha_{ij} = x_j^t u_j$. Entonces

$$x_i = \sum_{j=1}^{d} (x_i^t u_j) u_j$$

Ahora, considera una aproximación basada en los primeros p < d vectores de la base de acuerdo al modelo lineal:

$$\widehat{x_i} = \sum_{j=1}^{p} z_{ij} u_j + \sum_{j=p+1}^{d} b_j u_j$$

Observa que los coe
cientes z_{ij} *dependen* de la observación i, mientras que b_j son constantes para todas las observaciones.

Considera minimización de la siguiente función de costo:

$$L = \frac{1}{n} \sum_{i=1}^{n} \left\| x_i - \widehat{x_i} \right\|^2$$

1.1. Solución

Inciso a)
Muestra que el minimo de la función de costo es:

$$\begin{aligned} z_{ij} &= x_i^t u_j, \quad j = 1, .., p \\ b_j &= \bar{x}^t u_j, \quad j = p+1, ..., d \\ x_i - \widehat{x_i} &= \sum_{j=p+1}^{d} [(x_i - \bar{x})^t u_j] u_j \end{aligned}$$

Vamos a partir de la función de costos y se va a ir desarrollando poco a poco los elementos, ya que buscamos minimizar, debemos encontrar una expresión sencilla.

$$L = \frac{1}{n}\sum_{i=1}^{n}\left\|x_i - \widehat{x_i}\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x_i})^t(x_i - \widehat{x_i})$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left\{x_i^t x_i - x_i^t\widehat{x_i} - \widehat{x_i}^t x_i + \widehat{x_i}^t\widehat{x_i}\right\}$$

Se va a realizar la observación detallada de cada elemento que se encuentra dentro de las llaves por lo que empezaremos con $x_i^t x_i$.

$$x_i^t x_i = \left[\sum_{j=1}^{d}\alpha_{ij}u_j\right]^t\left[\sum_{j=1}^{d}\alpha_{ij}u_j\right]$$

$$= [\alpha_{i1}u_1 + \alpha_{i2}u_2 + ... + \alpha_{id}u_d]^t\,[\alpha_{i1}u_1 + \alpha_{i2}u_2 + ... + \alpha_{id}u_d]$$
$$= \left[\alpha_{i1}u_1^t + \alpha_{i2}u_2^t + ... + \alpha_{id}u_d^t\right][\alpha_{i1}u_1 + \alpha_{i2}u_2 + ... + \alpha_{id}u_d]$$

Si tomamos un elemento del primer corchete, como $\alpha_{i1}u_1^t$ y lo multiplicamos por todo lo que se encuentre en el segundo corchete: [$\alpha_{i1}u_1$ + $\alpha_{i2}u_2$ +....+ $\alpha_{id}u_d$ +....+ $\alpha_{id}u_d$], se va a encontrar que el único término que va a permanecer es: $\alpha_{i1}^2 u_i^t u_i$.
Ya que los vectores u_j pertenecen a una base ortonormal y cumple con lo siguiente:

$$u_j^t u_i = 1, \quad si\ j = i$$
$$u_j^t u_i = 0, \quad si\ j \neq i$$

Así nuestro término $\alpha_{i1}^2 u_i^t u_i$ en realidad queda como: α_{i1}^2 por lo anterior mencionado.

Ahora nuestra expresión $x_i^t x_i$ la podemos volver a escribir como:

$$x_i^t x_i = \alpha_{i1}^2 + \alpha_{i2}^2 + ... + \alpha_{ij}^2 + ... + \alpha_{id}^2$$

Donde si se observa solo es una suma de alphas elevadas al cudrado. Ahora analizamos el siguiente término $x_i^t \widehat{x_i}$.

$$x_i^t \widehat{x_i} = \left[\sum_{j=1}^{d} \alpha_{ij} u_j\right]^t \left[\sum_{j=1}^{p} z_{ij} u_j + \sum_{j=p+1}^{d} b_j u_j\right]$$

$$= [\alpha_{i1}u_1 + \alpha_{i2}u_2 + ... + \alpha_{id}u_d]^t \, [z_{i1}u_1 + z_{i2}u_2 + ... + z_{ip}u_p + b_{p+1}u_{p+1} + ... + b_d u_d]$$

Si tomamos un elemento del primer corchete, como $\alpha_{i1}u_1^t$ y lo multiplicamos por todo lo que se encuentre en el segundo corchete: $[z_{i1}u_1 + z_{i2}u_2 + ... + z_{ip}u_p + b_{p+1}u_{p+1} + ... + b_d u_d]$, se va a encontrar que el único término que va a permanecer es: $\alpha_{i1} z_{i1} u_i^t u_i$.
Como se menciono que cada u_j pertenecen a una base ortonormal nuestro término queda de la siguiente forma $\alpha_{i1} z_{i1} u_i^t u_i = \alpha_{i1} z_{i1}$.
Ahora nuestra expresión $x_i^t \widehat{x_i}$ la podemos volver a escribir como:

$$x_i^t \widehat{x_i} = \alpha_{i1} z_{i1} + \alpha_{i2} z_{i2} + ... + \alpha_{ip} z_{ip} + \alpha_{i,p+1} b_{p+1} + ... + \alpha_d b_d$$

El siguiente término $\widehat{x_i}^t x_i$

$$\widehat{x_i}^t x_i = \left[\sum_{j=1}^{p} z_{ij} u_j + \sum_{j=p+1}^{d} b_j u_j\right]^t \left[\sum_{j=1}^{d} \alpha_{ij} u_j\right]$$

$$= [z_{i1}u_1 + z_{i2}u_2 + ... + z_{ip}u_p + b_{p+1}u_{p+1} + ... + b_d u_d]^t \, [\alpha_{i1}u_1 + \alpha_{i2}u_2 + ... + \alpha_{id}u_d]$$
$$= \left[z_{i1}u_1^t + z_{i2}u_2^t + ... + z_{ip}u_p^t + b_{p+1}u_{p+1}^t + ... + b_d u_d^t\right] [\alpha_{i1}u_1 + \alpha_{i2}u_2 + ... + \alpha_{id}u_d]$$

Todo esto producto es igual al calculado con $x_i^t \widehat{x_i}$ por lo que para $\widehat{x_i}^t x_i$ tendremos:

$$\widehat{x_i}^t x_i = \alpha_{i1} z_{i1} + \alpha_{i2} z_{i2} + ... + \alpha_{ip} z_{ip} + \alpha_{i,p+1} b_{p+1} + ... + \alpha_d b_d$$

Finalmente para $\widehat{x_i}^t\widehat{x_i}$

$$\widehat{x_i}^t\widehat{x_i} = \left[\sum_{j=1}^{p} z_{ij}u_j + \sum_{j=p+1}^{d} b_j u_j\right]^t \left[\sum_{j=1}^{p} z_{ij}u_j + \sum_{j=p+1}^{d} b_j u_j\right]$$

$$= [z_{i1}u_1 + z_{i2}u_2 + ... + z_{ip}u_p + b_{p+1}u_{p+1} + ... + b_d u_d]^t\,[z_{i1}u_1 + z_{i2}u_2 + ... + b_d u_d]$$
$$= \left[z_{i1}u_1^t + z_{i2}u_2^t + ... + z_{ip}u_p^t + b_{p+1}u_{p+1}^t + ... + b_d u_d^t\right][z_{i1}u_1 + z_{i2}u_2 + ... + b_d u_d]$$

Si tomamos un elemento del primer corchete, como $z_{i1}u_1^t$ y lo multiplicamos por todo lo que se encuentre en el segundo corchete: $[z_{i1}u_1 + z_{i2}u_2 +...+ z_{ip}u_p + b_{p+1}u_{p+1} +...+ b_d u_d]$, se va a encontrar que el único término que va a permanecer es: $z_{i1}^2 u_i^t u_i$, esto sucede por lo antes mencionado de la base ortonormal de los u_j.
Así podemos escribir nuestro término como:

$$\widehat{x_i}^t\widehat{x_i} = z_{i1}^2 + z_{i2}^2 + ... + z_{ip}^2 + b_{p+1}^2 + ... + b_d^2$$

Regresando a nuestra función de costos y remplazando lo que se encontro tenemos

$$L = \frac{1}{n}\sum_{i=1}^{n}\left\|x_i - \widehat{x_i}\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x_i})^t(x_i - \widehat{x_i})$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left\{x_i^t x_i - x_i^t\widehat{x_i} - \widehat{x_i}^t x_i + \widehat{x_i}^t\widehat{x_i}\right\}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[(\alpha_{i1}^2 + ... + \alpha_{id}^2) - 2(\alpha_{i1}z_{i1} + ... + \alpha_d b_d) + (z_{i1}^2 + ... + b_d^2)\right]$$

Para poder encontrar un valor minimo vamos a derivar la anterior expresión respecto a z_{ij}. Realizando dicha acción vamos a tener como resultado, note que se esta derivando respecto a un determinado i, j:

$$\frac{\partial L}{\partial z_{ij}} = -2\alpha_{ij} + 2z_{ij}$$

Igualando a cero vamos a tener lo siguiente: $\alpha_{ij} = z_{ij}$.
Ahora derivamos respecto a b_j

$$\begin{aligned}\frac{\partial L}{\partial b_j} &= \frac{1}{n}\sum_{j=1}^{n}(-2\alpha_{ij} + 2b_j) \\ &= -2\frac{1}{n}\sum_{j=1}^{n}(\alpha_{ij}) + \frac{n}{n}(2b_j) \\ &= -2\frac{1}{n}\sum_{j=1}^{n}(\alpha_{ij}) + 2b_j\end{aligned}$$

Igualando a cero tenemos lo siguiente: $b_j = \frac{1}{n}\sum_{j=1}^{n}(\alpha_{ij})$ pero recordemos que $\alpha_{ij} = x_i^t u_j$ por lo que tenemos que $b_j = \frac{1}{n}\sum_{i=1}^{n}(x_i^t u_j)$ pero esto corresponde a la media pero transpuesta por lo que realmente tenemos que $b_j = \bar{x}^t u_j$. Al analizar las segundas derivadas respecto a z_{ij}, b_j y las derivadas cruzadas se encontro que encontramos un punto minimo. Por lo que sustituyendo lo encontrado en $x_i - \widehat{x_i}$, tendremos lo siguiente:

$$\begin{aligned}x_i - \widehat{x_i} &= \sum_{j=1}^{d}\alpha_{ij}u_j - \left[\sum_{j=1}^{p} z_{ij}u_j\right] \\ &= \sum_{j=1}^{d}\alpha_{ij}u_j - \sum_{j=1}^{p}(x_i^t u_j)u_j - \sum_{j=p+1}^{d}(\bar{x}^t u_j)u_j \\ &= \sum_{j=1}^{d}\alpha_{ij}u_j - \sum_{j=1}^{p}\alpha_{ij}u_j - \sum_{j=p+1}^{d}(\bar{x}^t u_j)u_j \\ &= \sum_{j=p+1}^{d}(\alpha_{ij}u_j - (x_i^t u_j)u_j) \\ &= \sum_{j=p+1}^{d}[(x_i - \bar{x})^t u_j]u_j.\end{aligned}$$

Inciso b)

Partimos de lo siguiente:

$$L = \frac{1}{n}\sum_{i=1}^{n} \left\|x_i - \widehat{x_i}\right\|^2 = \frac{1}{n}\sum_{i=1}^{n} \left\|\sum_{j=p+1}^{d}[(x_i - \bar{x})^t u_j]u_j\right\|^2$$
$$= \frac{1}{n}\sum_{i=1}^{n} \left[\sum_{j=p+1}^{d}[(x_i - \bar{x})^t u_j]u_j\right]^t \left[\sum_{j=p+1}^{d}[(x_i - \bar{x})^t u_j]u_j\right]$$
$$= \frac{1}{n}\sum_{i=1}^{n} \left[\sum_{j=p+1}^{d} u_j^t[u_j^t(x_i - \bar{x})]\right] \left[\sum_{j=p+1}^{d}[(x_i - \bar{x})^t u_j]u_j\right]$$

Como se menciono anteriormente los u_j provienen de una base ortogonal por lo que vamos a tener los iguiente:

$$L = \frac{1}{n}\sum_{i=1}^{n} \left[\sum_{j=p+1}^{d}[(x_i - \bar{x})^t u_j]^2\right]$$
$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=p+1}^{d}(x_i^t u_j - \bar{x}^t u_j)^2$$

También observe como podemos desarrollar nuestra expresión:

$$L = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=p+1}^{d}(x_i^t u_j - \bar{x}^t u_j)^2$$

$$= \sum_{j=p+1}^{d}\frac{1}{n}\sum_{i=1}^{n}(x_i^t u_j - \bar{x}^t u_j)^t(x_i^t u_j - \bar{x}^t u_j)$$

$$= \sum_{j=p+1}^{d}\frac{1}{n}\sum_{i=1}^{n}u_j^t(x_i - \bar{x})(x_i^t - \bar{x}^t)u_j$$

$$= \sum_{j=p+1}^{d}u_j^t\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^t u_j$$

Observe que $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^t = S$, por lo que finalmente vamos a tener:

$$L = \sum_{j=p+1}^{d}u_j^t S u_j$$

2. Problema 2

Índices de Marginación de las localidades en Nuevo León para el año 2010

El índice de Marginacion (IM) desarrollado por el Consejo Nacional de Población (CONAPO) permite identificar, por áreas geográficas, la intensidad de las privaciones y exclusión social de la población. La intensificación de esas carencias configura entornos más adversos para el devenir educativo de los habitantes, en particular de los infantes. Se trata de una medida continua que aumenta de valor en tanto mayores porcentajes de

la población viven en localidades pequeñas, en viviendas inadecuadas, con falta de acceso a la educación y con ingresos monetarios reducidos. El IM se utiliza ampliamente en México para establecer jerarquías entre las unidades territoriales, según la intensidad de las carencias estructurales padecidas por sus pobladores y, de este modo, ofrece un criterio sólido para priorizar acciones de política social en los distintos niveles de gobierno.

Con este informe se pretende calcular el índice de marginación para 2,037 localidades del estado de Nuevo León con base al Censo de Población y Vivienda de 2010 realizado por el Instituto Nacional de Estadística y Geografía (INEGI). Para encontrar dicho índice se van a calcular antes ocho indicadores socioeconomicos que nos brindaran información para ir construyendo el índice para cada localidad.
Para analizar el índice de marginación en el estado de Nuevo León se toma la base del Censo de Población y Vivienda del 2010. La base se conforma de un total de 5,265 localidades y 625 características, que reflejado en el archivo son los renglones y columnas respectivamente. En la base de datos se observa que muchos de las localidades no cuentan con varios datos que son importantes para realizar el analisis. Para tratar el problema de los valores faltantes se consulto el reporte oficial de los indicadores socioeconomicos, e índice de marginación de Nuevo León que proporciona la CONAPO.
El cual no considera 3,228 localidades, para las cuales no fue posible calcular el índice de marginación, debido a que estas localidades cuentan con una o dos viviendas o no registran información.
Una vez que realizamos el filtrado en la base de datos se reduce la cantidad de localidades, también solo tomamos aquellas caracteristicas que nos sirve para encontrar cada uno de los indicadores, el total de caracteristicas con las que vamos a trabajar al final son solo 24.

2.1. Indicadores de Marginación

Siguiendo la métodologia del anexo técnico que proporciona CONAPO, nos sugiere que tomemos el valor medio del resto de las localidades para

rellenar aquellos indicadores que no se puedan calcular, es decir, aquellas localidades que si cuentan con la mayoria de las caracteristicas pero por falta de un dato o dos no se pudiera encontrar uno de sus indicadores. Ahora vamos a mostrar las caracteristicas que cada indicador necesita para poder ser calculado.

1) Porcentaje de población de 15 años o más analfabeta
    * Población de 15 años o más
    * Población de 15 años o más analfabeta
2) Porcentaje de población de 15 años o más sin primaria completa
    * Población de 15 años o más sin escolaridad
    * Población de 15 años o más con primaria incompleta
    * Población de 15 años o más con primaria completa
    * Población de 15 años o más con secundaria incompleta
    * Población de 15 años o más con secundaria completa
    * Población de 18 años o más con educación pos-básica
3) Porcentaje de viviendas particulares habitadas sin excusado
    * Viviendas particulares habitadas totales
    * Viviendas particulares habitadas que disponen de sanitario
4) Porcentaje de viviendas particulares habitadas sin energía eléctrica
    * Viviendas particulares habitadas que disponen de luz eléctrica
    * Viviendas particulares habitadas que no disponen luz eléctrica
5) Porcentaje de viviendas particulares habitadas sin agua entubada
    * Viviendas particulares habitadas que disponen de agua entubada fuera de la vivienda
    * Viviendas particulares habitadas que disponen de agua entubada dentro de la vivienda

6) Promedio de ocupantes por cuarto en viviendas particulares habitadas
    * Promedio de ocupantes por cuarto en viviendas particulares habitadas
7) Porcentaje de viviendas particulares habitadas con piso de tierra
    * Viviendas particulares habitadas con piso de tierra
    * Viviendas particulares habitadas con piso diferente de tierra
8) Porcentaje de viviendas particulares habitadas que no disponen de refrigeradora
    * Viviendas particulares habitadas totales
    * Viviendas particulares habitadas que disponen de refrigerador

Cada indicador cuenta con una expresión para ser calculada donde solo hay que sustituir los valores que nos pide la expresión pero antes de tomar los datos hay que consultar el Diccionario de datos SCINCE del INEGI, donde encontramos el nombre del campo, dicho nombre de campo es el que nos ayuda a encontrar alguna caracteristica que viene en nuestra base de datos. La expresión y los nombres de campo que se utilizaron se mostraran al final en un apendice, donde se dara una breve explicación de como sustituir los datos.
En el cuadro 2.1 nos muestra de manera resumida cierta información sobre los indicadores. Por ejemplo, la media del procentaje de población de 15 años o más analfabeta es cerca del 7.75 %, este dato refleja cierta información valiosa. La información de la tabla nos pueda dar un panorama general sobre el estado, cabe recalcar que estos indicadores fueron comparados con los que se nos propocionados y los resultados son cercanos.
En la siguiente figura 2.1 se puede visualizar la relación que se encontro en ciertos indicadores, es decir, hay cierto comportamiento relacionado entre ellas, la escala va de -1 a 1. Por ejemplo hay una correlación de 0.55, entre el indicador 4 y 8. En otra palabras podemos decir que el procentaje de viviendas particulares habitadas sin energia eléctrica y el porcentaje de

	PC	Min	Media	Mediana	Max
I1	1	0.0	7.74	5.60	77.78
I2	2	0.0	39.33	38.25	100.00
I3	3	0.0	7.39	1.41	100.00
I4	4	0.0	4.49	0.00	100.00
I5	5	0.0	40.47	29.09	100.00
I6	6	0.4	1.24	1.20	10.00
I7	7	0.0	5.77	0.00	100.00
I8	8	0.0	22.27	12.20	100.00

Cuadro 2.1: Datos sobre los Indicadores.

viviendas particulares habitas que no disponen de refrigeración, guardan una relación positiva.
Esta información también se podria entender que si alguna de ellas cambia, este cambio podria afectar a la otra.

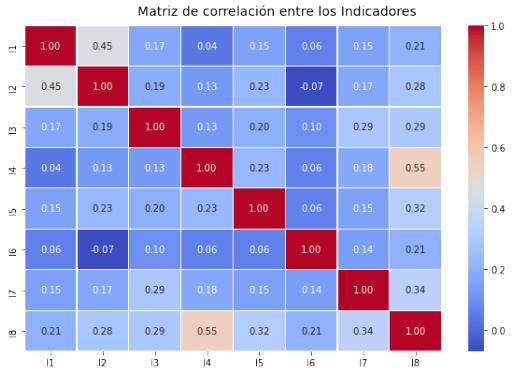

Figura 2.1: Correlaciones de Indicadores.

2.2. Construcción del Índice de Marginación

Los datos que se mostraron en la figura 2.1 además de dar información que se podría interpretar es necesaria para encontrar nuestro Índice de Marginación para cada localidad. Siguiendo la métodologia del anexo técnico que proporciona CONAPO, habla sobre encontrar los componentes principales para cada indicador socieconomico, para la parte del cálculo nos podriamos apoyar de algún software, en este caso se utilizo PYTHON. La forma de encontrar estos componentes principales se propocionara al final en un apéndice.
Una vez que tenemos estos valores los vamos a utilizar para poder encontrar el peso relativo que tomo cada indicador. Estos valores se presentan en el siguiente cuadro 2.2, donde también se agregar una columna que corresponde al poderado de lectura, el peso viene representado por el coeficiente.

	Coeficiente	Ponderado de Lectura
I1	0.312217	0.768690
I2	0.360996	0.888785
I3	0.344548	0.848288
I4	0.364811	0.898177
I5	0.347140	0.854669
I6	0.151352	0.372634
I7	0.356156	0.876869
I8	0.500142	1.231366

Cuadro 2.2: Datos sobre los Indicadores.

Al final solo nos restaria realizar operaciones sencillas para encontrar cada indice para todas nuestras localidades, se explica en el apendice con mayor detalle. Finalmente tenemos la gráfica en la figura de nuestros índices de marginación de las localidades.

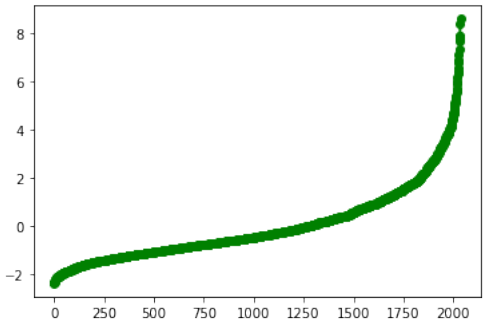

Figura 2.2: Índices de Marginación de las localidades.

2.3. Conclusiones

Al realizar todos los cálculos y encontrando cada uno de los indicadores e Índices de Marginación, fue posible acercarnos a los valores reales de cada indicador e Índice, de las localidades pero al 100 % no por la falta de algunos datos. Al tomar el valor medio e imputar esta cantidad para aquellas localidades que les hiciera falta nos causo cierta variación porque en algunos revisamos nuestro resultado con que el dato real y habia en unos casos mucha diferencia.
Al realizar una pequeña investigación se encontro que hay más indicadores que se podrían tomar en cuenta para calcular el Índice de Marginación, incluso tomar en cuenta estos nuevos indicadores encontrar nuevas relaciones entre ellos.

2.4. Apéndice

En esta sección se pretende exponer las expresiones que se emplearon para encontrar los distintos valores que se mencionaron antes.

* EDU28 - Población de 15 años o más
* POB20 - Población de 15 años o más analfabeta

* p15yn_se - Población de 15 años o más sin escolaridad
* p15pri_in - Población de 15 años o más con primaria incompleta
* p15pri_co - Población de 15 años o más con primaria completa
* p15sec_in - Población de 15 años o más con secundaria incompleta
* p15sec_co - Población de 15 años o más con secundaria completa
* p18yn_pb - Población de 18 años o más con educación pos-básica
* VIV2 - Viviendas particulares habitadas totales
* VIV9 - Viviendas particulares habitadas que disponen de sanitario
* VIV15 - Viviendas particulares habitadas que disponen de luz eléctrica
* VIV14 - Viviendas particulares habitadas que no disponen luz eléctrica
* VIV17 - Viviendas particulares habitadas que disponen de agua entubada fuera de la vivienda
* VIV17 - Viviendas particulares habitadas que disponen de agua entubada dentro de la vivienda
* VIV5_R - Promedio de ocupantes por cuarto en viviendas particulares habitadas
* VIV6 - Viviendas particulares habitadas con piso de tierra
* VIV2-VIV6 - Viviendas particulares habitadas con piso diferente de tierra
* VIV2 - Viviendas particulares habitadas totales
* VIV26 - Viviendas particulares habitadas que disponen de refrigerador

Una vez que tenemos nuestros datos definidos vamos a mostrar las distintas expresiones que necesitamos para cada uno de nuestros indicadores, índices de marginación y demás calculos.

Indicador 1: Porcentaje de población de 15 años o más analfabeta.

$$I1 = \frac{EDU28}{POB} * 100$$

Indicador 2: Porcentaje de población de 15 años a más sin primaria completa.

$$I2 = \frac{p15ym_{se} + p15pri_{in}}{p15ym_{se} + p15pri_{in} + p15pri_{co} + p15se_{in} + p18ym_{pb}} * 100$$

Indicador 3: Porcentaje de viviendas particulares habitadas sin excusado.

$$I3 = \frac{VIV2 - VIV19}{VIV2} * 100$$

Indicador 4: Porcentaje de viviendas particulares habitadas sin energía eléctrica.

$$I4 = \frac{VIV15}{VIV15 + VIV14} * 100$$

Indicador 5: Porcentaje de viviendas particulares habitadas que no disponen de agua entubada.

$$I5 = \frac{VIV17}{VIV16 + VIV17} * 100$$

Indicador 6: Promedio de ocupantes por cuarto de viviendas particulares habitadas sin excusado.

$$I6 = VIV5_R$$

Indicador 7: Porcentaje de viviendas particulares habitadas con piso de tierra.

$$I7 = \frac{VIV6}{VIV6 + (VIV2 - VIV6)} * 100$$

Indicador 8: Porcentaje de viviendas particulares habitadas que no disponen de refrigerador.

$$I8 = \frac{VIV2 - VIV26}{VIV2} * 100$$

Cada indicador fue construido con la anterior expresión donde solo se hizo una sustitución sencilla, es decir, una vez que definamos la localidad donde queremos encontrarle sus indicadores resta buscar las caracteriticas necesarias y realizar las operaciones. Este proceso puede ser algo lento hacerlo uno por uno por lo que se puede automatizar con ayuda de un software como PYTHON.

Construcción del Índice.

Para poder encontrar el índice de marginación tenemos lo siguiente: Note que w_1 el vector de pesos del primer componente principal y sea X la matriz que contiene las observaciones, donde realizamos:

$$z_{1j} = w_1^t X$$

En otras palabras realizar la anterior acción equivale a algo llamada multiplicación matricial pero si no se tiene conocimiento de lo antes mencionado igual se puede recurrir a la ayuda de un software para realizar dichas operaciones, solo definir bien los elementos involucrados.

3. Problema 3

Rostros LFW

El conjunto de datos que se consideraron fueron los datos Ladelled Faces in the Wild(LFW), que consiste en fotografías de rostros recolectados de internet y contenido en la biblioteca sklearn. Para el análisis que nos solicitan vamos a considerar solo aquellas personas que tienen al menos 70 fotografías de su rostro. La cantidad de imágenes que se juntaron fue de 1288, donde el número de personas total fue de 7. Los nombres que se encontraron fueron:

a) Ariel Sharon

b) Colin Powell

c) Donald Rumsfeld

d) George W Bush

e) Gerhard Schroeder

f) Hugo Chavez

g) Tony Blair

El lenguaje de programación con el cual nos apoyamos fue PYTHON. Las imágenes que encontramos tienen un tamaño de 125x94 pixeles, en la figura 3.1 se tomo un conjunto de 15 imágenes para ilustarlas y observar algunos rostros.

Figura 3.1: Imágenes en la base de datos.

Como primer tarea se desea tomar el 80 % de los datos, estos datos que vamos a tomar los llamaremos datos de entrenamiento. Lo anterior fue posible ocupando la función train_test_split que nos facilita dividir los datos. Este 80 % de los datos se selecciona de manera aleatoria. Una vez que tenemos el conjunto de datos de entrenamiento vamos a trabajar con ellos. Cada imagen puede ser expresada como una matriz, la cual estara compuesta por filas y columnas con información sobre cada imagen, este conjunto de datos son los que nos ayudan a visualizar las imagenes de la

figura 3.1.
Como primer paso se estandarizan los datos de entrenamiento, en la figura 3.2 se ven algunas imágenes estandarizadas, esta figura nos ayuda a ilustrar el proceso. Se observa que en algunas de ellas hay ciertas sombras nuevas que incluso hacen que la imagen se vea más oscura.

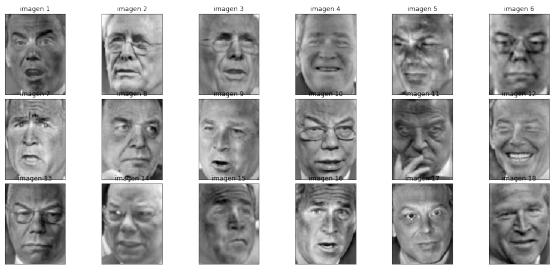

Figura 3.2: Imágenes estandarizadas.

Una vez que tenemos nuestros datos estandarizados se realizara un proceso llamado PCA que nos ayudara a encontrar los componentes principales, lo que se quiere dar a entender cuando se habla de componentes principales son aquellas caracteristicas de los datos, en este caso imagenes, que nos ayuden a clasificar futuras fotografías, es decir, aquellos rasgos unicos que ayudan a hacer cierta diferencia entre un grupo de imágenes de un individuo y otro.
Para este ejercicio se calcularon un total de 150 componentes principales con las imagenes de las cuales se muestran los primeros dos componentes y se ilustran en la figura 3.3.

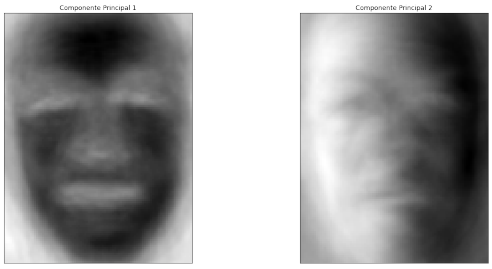

Figura 3.3: Primeros dos componentes principales.

La imagen que se encuentra a la izquierda pertenece a la primer componente principal y la del lado derecho a la segunda componente principal. En el primer componente se logra ver un rostro donde las partes que aparecen más claras son donde están ubicados los ojos, nariz y boca, incluso se puede agregar la forma del rostro. Para el segundo componente principal se ve muy poco el rostro pero esta componente nos puede ayudar a distinguir ciertas sombras que tuviera la imágen, porque en esta se ve un lado muy claro y el otro más oscuro. Como se menciono antes, con ayuda de estos componentes principales se pretende encontrar una distinción entre los grupos de imágenes de cada individuo, para lograr esto los datos que nos proporcionan son colocados en matrices al igual que los componentes, para finalmente ayudarnos a visualizar los datos en un gráfico de dispersión, donde facilmente se pueden observar patrones, grupos o algún comportamiento de los datos.

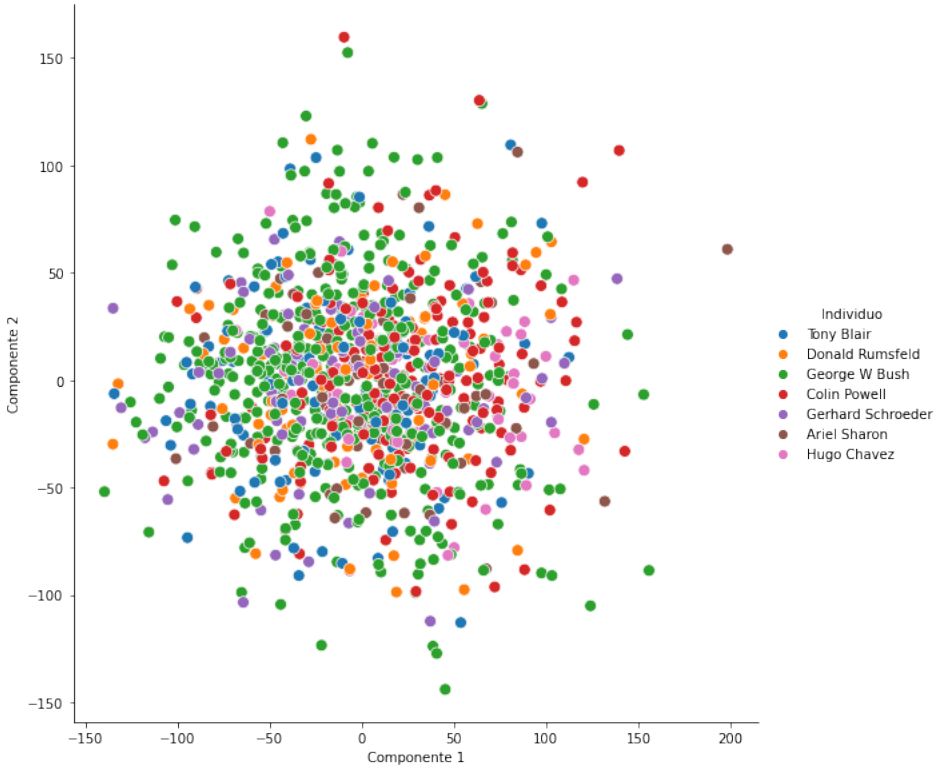

Figura 3.4: Proyección de los datos en los primeros dos componentes.

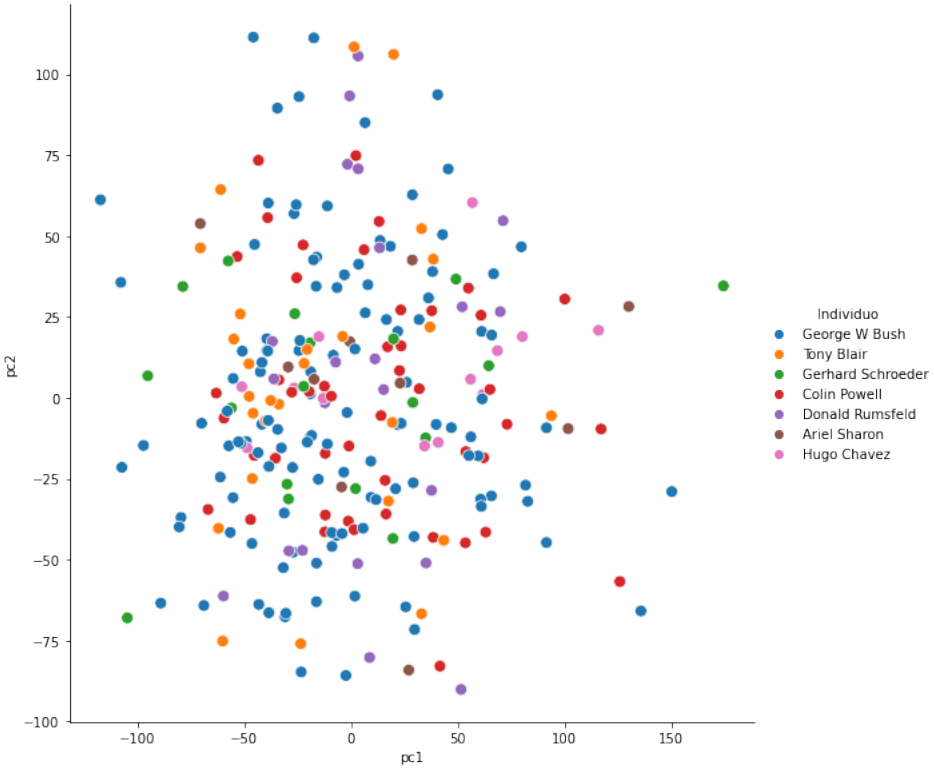

Figura 3.5: Proyección de los datos de prueba en los primeros dos componentes.

En la figura 3.4 y 3.5 nos muestra como estan los datos en estas primeras dos componentes, nos ayuda a distinguir entre las imágenes de cada individuo, colocando un color en especifico. En ambos gráficos no se observa algún grupo conformado solo por un color, en algunos casos se observa que puntos del mismo color están cerca y en otros, puntos de distinto color estan más pegados y se pensaria que no deberia ser así pero puede que en esas imágenes coincidiera la forma que puso la boca, los ojos, etc., el

individuo al momento de tomarle la fotografía y esta caracteristica es la que resalta al menos para estos dos primeros componentes.
Ahora que pasaría si tomamos una nueva observación X y sin saber a que individuo pertenerce, se pretender encontrarle un grupo, en este caso se ilustra esta idea en la figura 3.6.

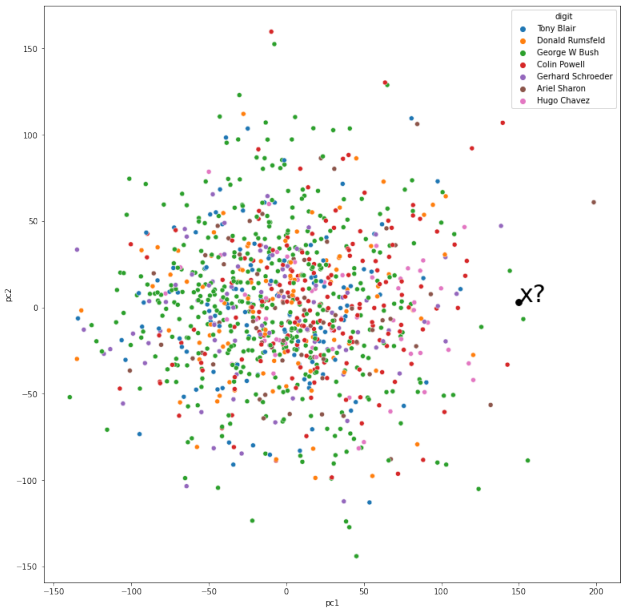

Figura 3.6: Proyección de los datos y una nueva observación.

No vamos a tomar esta representación de los datos para realizar dicha tarea, antes hay que revisar que tanta varianza aporta cada una de los componentes principales, ya que se desea conservar al menos un 80 % de la varianza explicada que tienen los datos. La varianza que nos proporcionan los datos nos ayudara a hacer la distinción entre cada grupo de imágenes.

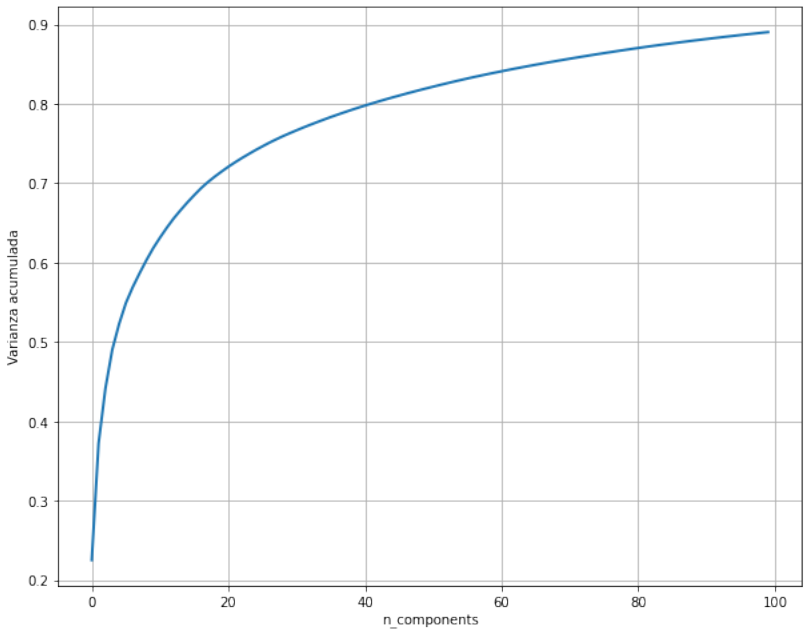

Figura 3.7: Varianza aportada por cada componente principal.

En la figura 3.7 nos muestra que al tomar 40 componentes principales podemos conservar el 80 % de la varianza explicada, por lo que tomaremos las primeras 40 componentes principales que obtuvimos y con ellas trabajar el caso de una nueva observación. Ya que hemos dicho cuantas compontes principales se van a ocupar nos falta mencionar el criterio o método a ocupar, para este caso se empleara el método del vecino más cercano para indentificar a un "sujeto"que pertenece al 20 % restante de las imágenes. Se ocupara la distancia euclideana para determinar que es cerca y que es lejos.
Como estamos hablando de 40 componentes principales no es tan sencillo visualizar estas operaciones pero podemos mostrar el resultado final del proceso. En este caso vamos a mostrar las imagenes que esten involucradas,

es decir, la imagen original, la proyección y el vecino más cercano. Cabe recalcar que para realizar esta tarea igual se estandarizados los datos que estamos intentando clasificar.

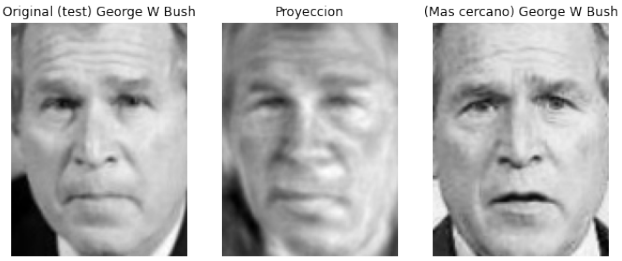

Figura 3.8: Ejemplo de vecino cercano 1.

En la figura 3.8 se observa que si coincide la imagen que es nueva con el vecino más cercano, si prestamos atención hay cierta coincidencia en la forma que tienen la boca y puede que este rasgo nos ayudara a clasificar mejor nuestra imagen original. En la figura 3.8 se muestra la proyección que se realizo de dicha imagen, en ella se ven nuevas sombras pero aún se distingue el rostro.

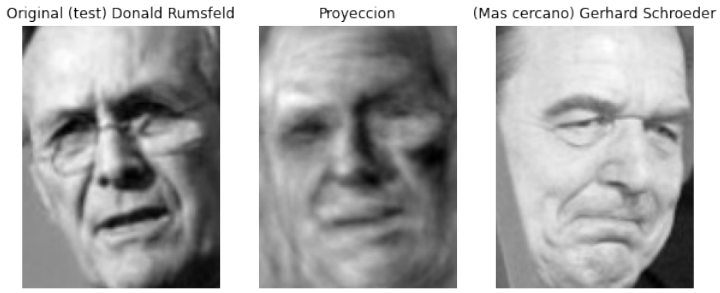

Figura 3.9: Ejemplo de vecino cercano 2.

En la figura 3.9 es un ejemplo de que el método tiene ciertas fallas ya que ambas imágenes cuentan con una persona diferente pero coincide en la forma de la boca y también podemos agregar los pómulos ligeramente más marcados en ambos. En la misma figura 3.9 se muestra la proyección que se realizo de dicha imagen. En ella podemos notar que hay cierto cambio brusco de la imagen e incluso se pierde cierta forma del rostro. Como dato relevante se encontro que cerca del 56 % de los casos nuestro método clasifico de manera adecuada las imagenes nuevas que ingresamos. Finalmente vamos a observar que sucede cuando tomamos en cuenta un conjunto de datos que no pertenecen al conjunto incial de 7 personas, en la figura 3.10 mostramos los datos que tomamos, el conjunto de datos es pequeño.

Figura 3.10: Nuevas imagenes

En la figura 3.11 mostramos como el método del vecino más cercano intenta buscar ciertas similitudes en las imagenes nuevas que le proporcionamos.

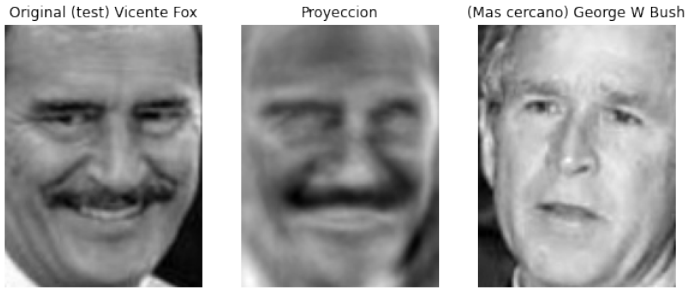

Figura 3.11: Ejemplo con nuevas imagenes

Es claro que el método va a tener fallas pero para poder prevenir casos como los que muestra la figura 3.11 podemos definir que distancia es considerada como mínima para que un nuevo elemento pertenezca a un determinado grupo de datos o imágenes. Se puede encontrar un valor medio de distancia que presenten los datos de entrenamiento de un solo individuo y en base a ese valor decir si la nueva observación pertenece o no al conjunto, si se decide que no pertenece al grupo crear un nuevo conjunto para este nuevo individuo y evitar que nos regrese un vecino "cercano".

4. Problema 4

Ahora, considera la matriz $K_{nxn} = X^t X$.

4.1. Solución

Inciso a)

Muestra que es equivalente realizar PCA en S o en K, es decir, que $(\lambda^{-1/2}Xu,\lambda)$ es un par eigenvector-eigenvalor normalizado de K, y a su vez, $(\lambda^{-1/2}X^tv,\lambda)$ es un par eigenvector-eigenvalor normalizado de S, donde u y v son vectores propios de S y K, respectívamente.

Podemos partir de $K = XX^t$, donde:

$$\begin{aligned} Kv &= \lambda v \\ (XX^t)v &= \lambda v \end{aligned}$$

Ahora vamos a multiplicar ambos lado por X^t, así tenemos:

$$\begin{aligned} X^t(XX^t)v &= X^t(\lambda v) \\ \implies (X^tX)(X^t\lambda) &= \lambda(X^tu) \end{aligned}$$

Si tomamos $X^tv = u$, podemos escribir lo anterior como:

$$\implies (X^tX)u = \lambda u$$

Hemos encontrado que la nueva matriz X^tX tiene como valor propio el mismo λ. Pero ahora si partimos de la matriz $S = X^tX$, tenemos lo siguiente:

$$\begin{aligned} Su &= \lambda u \\ (X^tX)u &= \lambda u \end{aligned}$$

Volvemos a multiplicar ambos lados pero ahora por la matriz X, así :

$$\begin{aligned} X(X^tX)u &= X\lambda u \\ \implies (XX^t)(X\lambda) &= \lambda(Xu) \end{aligned}$$

Si tomamos $Xu = v$, de esta forma podemos escribir lo anterior como:

$$\implies (XX^t)v = \lambda v$$

Finalmente encontramos que la matriz XX^t tiene el mismo valor propio λ.

Al igual es fácil observar que esto se cumple si tomamos la descomposición SVD ya que contamos con una matriz que no es simetrica.
Partimos de $X = UDA^t$, donde U y A son matrices ortogonales, la matriz D es una matriz diagonal, con elementos $(\sqrt{\lambda_1}, \sqrt{\lambda_2}, ..., \sqrt{\lambda_r})$

$$S = (UDA^t)^t(UDA^t) = (AD^tU^t)(UDA^t) = AD^2A^t = X^tX$$
$$K = (UDA^t)(UDA^t)^t = (UDA^t)(AD^tU^t) = UD^2U^t = XX^t$$

Donde se puede observar que las matrices tienen los mismos valores propios pero diferentes vectores propios. Igual observe que dichos vectores

Inciso b)
Verifica experimentalmente el resultado del inciso previo en el conjunto de imágenes LFW que usaste en el ejercicio anterior. ¿En qué casos es recomendable usar K?
Al realizar la prueba para el caso de la matriz S y K, vemos que en un caso tenemos una matriz de tamaño (1030,1030) y otra de tamaño de (11750,11750) respectivamente, para poder encontrar los valores propios tenemos el mismo resultado pero hay un factor importante que se hizo notar al momento del calculo que fue el tiempo para esperar dicho resultado.

Para la matriz de S el calculo no tardo tanto debido a la dimensión de la misma, mientras que para la matriz K tuvimos un mayor tiempo al momento de entontrar los valores propios.
Para la matriz K, se buscaron los componentes principales y se pueden ver en la figura 4.1, donde hay una clara diferencia entre los componentes principales que en un principio se mostraron. Algo que hay que tomar en cuenta es que cada matriz nos dio vectores propios distintos.

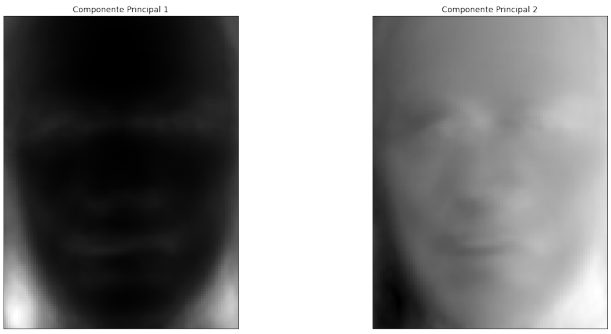

Figura 4.1: Componente principales para la matriz K

Ocupar la matriz K podría ser necesaria para el caso donde tenemos un número mayor de registros que de caracteristicas o variables.