Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación

Máquinas de aprendizaje

Reporte: Problema 1 con k-NN

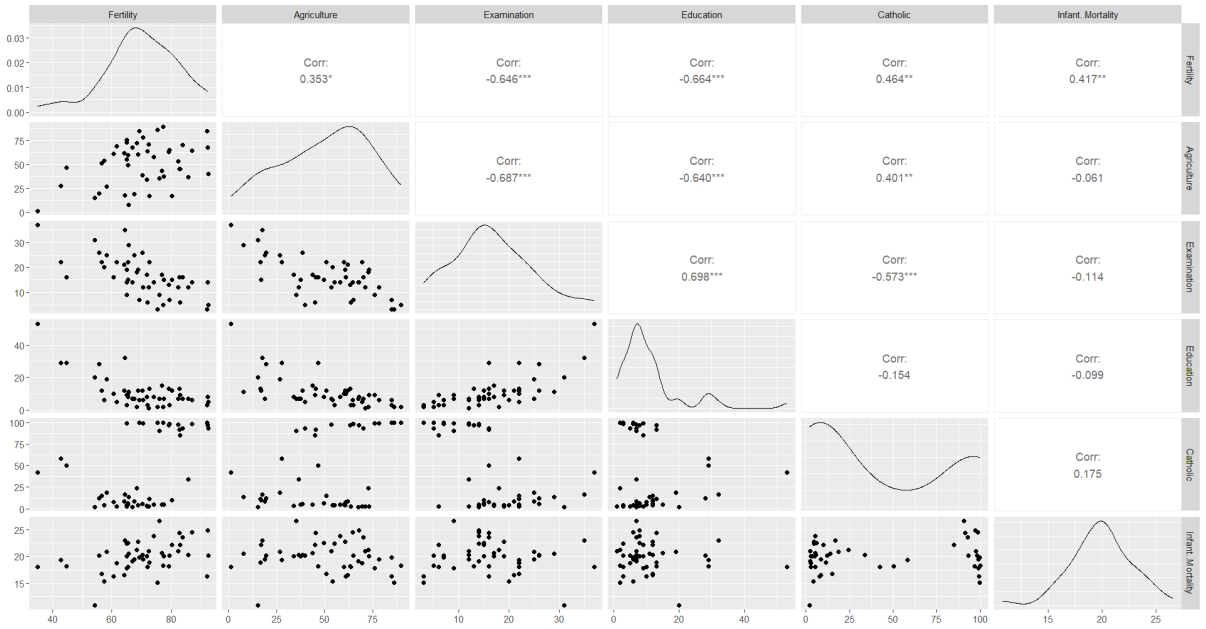
Docente: Abraham Sánchez López

Alumno	**Matrícula**
Taisen Romero Bañuelos	202055209

K-NN y la fertilidad en provincias francófonas

Bien, nuestro objetivo es encontrar las provincias con medidas de fertilidad similares usando k-NN.

La primera parte es explorar los datos, para ello, analicemos que nos dice su matriz de confusión.



Lo que podemos notar aquí es que hay más de una variable significativamente correlacionada con la fertilidad. Por ejemplo, la relación entre “Fertility” y “Catholic” tiene una correlación positiva de 0.464, lo que sugiere que en áreas con mayor proporción de gente católica, la fertilidad tiende a ser mayor. Pero en caso contrario, la relación entre “Fertility” y “Education” es inversa, o sea que a mayor nivel de educación es menor la fertilidad. Considerando que las variables marcadas con (*) tienen una relación significativa y las que tienen más asteriscos son significativamente más altos tenemos que estas son las variables más importantes:

- Examination
- Education
- Catholic
- Infant. Mortality

Una vez normalizados los datos aplicamos k-NN con k=7 (ya que la raíz de 47 es aproximadamente 7).

```
> cross_table
$t
             y
x             Avenches Delemont Entremont Grandson Moudon Neuveville Orbe Vevey Veveyse Yverdon
  Boudry             1        0         0        0      0          0    0     0       0       0
  Herens             0        0         0        0      0          0    0     0       1       0
  La Chauxdfnd       0        0         0        1      0          0    0     0       0       0
  Le Locle           0        0         0        0      0          0    0     0       0       1
  Martigwy           0        1         0        0      0          0    0     0       0       0
  Monthey            0        0         1        0      0          0    0     0       0       0
  Neuchatel          1        0         0        0      0          0    0     0       0       0
  Rive Droite        0        0         0        0      0          0    0     1       0       0
  Rive Gauche        0        0         0        0      0          0    1     0       0       0
  Sierre             0        0         1        0      0          0    0     0       0       0
  Sion               0        0         0        0      0          0    0     0       1       0
  St Maurice         0        0         0        0      0          0    0     0       1       0
  V. De Geneve       0        0         0        0      0          0    1     0       0       0
  Val de Ruz         0        0         0        0      1          0    0     0       0       0
  ValdeTravers       0        0         0        0      0          1    0     0       0       0

$prop.row
             y
x             Avenches Delemont Entremont Grandson Moudon Neuveville Orbe Vevey Veveyse Yverdon
  Boudry             1        0         0        0      0          0    0     0       0       0
  Herens             0        0         0        0      0          0    0     0       1       0
  La Chauxdfnd       0        0         0        1      0          0    0     0       0       0
  Le Locle           0        0         0        0      0          0    0     0       0       1
  Martigwy           0        1         0        0      0          0    0     0       0       0
  Monthey            0        0         1        0      0          0    0     0       0       0
  Neuchatel          1        0         0        0      0          0    0     0       0       0
  Rive Droite        0        0         0        0      0          0    0     1       0       0
  Rive Gauche        0        0         0        0      0          0    1     0       0       0
  Sierre             0        0         1        0      0          0    0     0       0       0
  Sion               0        0         0        0      0          0    0     0       1       0
  St Maurice         0        0         0        0      0          0    0     0       1       0
  V. De Geneve       0        0         0        0      0          0    1     0       0       0
  Val de Ruz         0        0         0        0      1          0    0     0       0       0
  ValdeTravers       0        0         0        0      0          1    0     0       0       0


$prop.col
             y
x             Avenches  Delemont Entremont  Grandson    Moudon Neuveville      Orbe     Vevey   Veveyse   Yverdon
  Boudry       0.5000000 0.0000000 0.0000000 0.0000000 0.0000000  0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
  Herens       0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0.0000000 0.0000000 0.0000000 0.3333333 0.0000000
  La Chauxdfnd 0.0000000 0.0000000 0.0000000 1.0000000 0.0000000  0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
  Le Locle     0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0.0000000 0.0000000 0.0000000 0.0000000 1.0000000
  Martigwy     0.0000000 1.0000000 0.0000000 0.0000000 0.0000000  0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
  Monthey      0.0000000 0.0000000 0.5000000 0.0000000 0.0000000  0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
  Neuchatel    0.5000000 0.0000000 0.0000000 0.0000000 0.0000000  0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
  Rive Droite  0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0.0000000 0.0000000 1.0000000 0.0000000 0.0000000
  Rive Gauche  0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0.0000000 0.5000000 0.0000000 0.0000000 0.0000000
  Sierre       0.0000000 0.0000000 0.5000000 0.0000000 0.0000000  0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
  Sion         0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0.0000000 0.0000000 0.0000000 0.3333333 0.0000000
  St Maurice   0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0.0000000 0.0000000 0.0000000 0.3333333 0.0000000
  V. De Geneve 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0.0000000 0.5000000 0.0000000 0.0000000 0.0000000
  Val de Ruz   0.0000000 0.0000000 0.0000000 0.0000000 1.0000000  0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
  ValdeTravers 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  1.0000000 0.0000000 0.0000000 0.0000000 0.0000000

$prop.tbl
             y
x             Avenches   Delemont  Entremont   Grandson     Moudon Neuveville       Orbe      Vevey    Veveyse    Yverdon
  Boudry      0.06666667 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
  Herens      0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.06666667 0.00000000
  La Chauxdfnd 0.00000000 0.00000000 0.00000000 0.06666667 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
  Le Locle    0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.06666667
  Martigwy    0.00000000 0.06666667 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
  Monthey     0.00000000 0.00000000 0.06666667 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
  Neuchatel   0.06666667 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
  Rive Droite 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.06666667 0.00000000 0.00000000
  Rive Gauche 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.06666667 0.00000000 0.00000000 0.00000000
  Sierre      0.00000000 0.00000000 0.06666667 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
  Sion        0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.06666667 0.00000000
  St Maurice  0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.06666667 0.00000000
  V. De Geneve 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.06666667 0.00000000 0.00000000 0.00000000
  Val de Ruz  0.00000000 0.00000000 0.00000000 0.00000000 0.06666667 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
  ValdeTravers 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.06666667 0.00000000 0.00000000 0.00000000 0.00000000
```

Las celdas con un “1” sugiere que tienen características similares. Por ejemplo, la celda de Le Locle y Yverdon tiene un 1.

```
$t
             y
x             Avenches Delemont Entremont Grandson Moudon Neuveville Orbe Vevey Veveyse Yverdon
  Boudry             1        0         0        0      0          0    0     0       0       0
  Herens             0        0         0        0      0          0    0     0       1       0
  La Chauxdfnd       0        0         0        1      0          0    0     0       0       0
  Le Locle           0        0         0        0      0          0    0     0       0       1
  Martigwy           0        1         0        0      0          0    0     0       0       0
  Monthey            0        0         1        0      0          0    0     0       0       0
  Neuchatel          1        0         0        0      0          0    0     0       0       0
```

Para comprobar esto veamos en la tabla si realmente tienen datos similares.

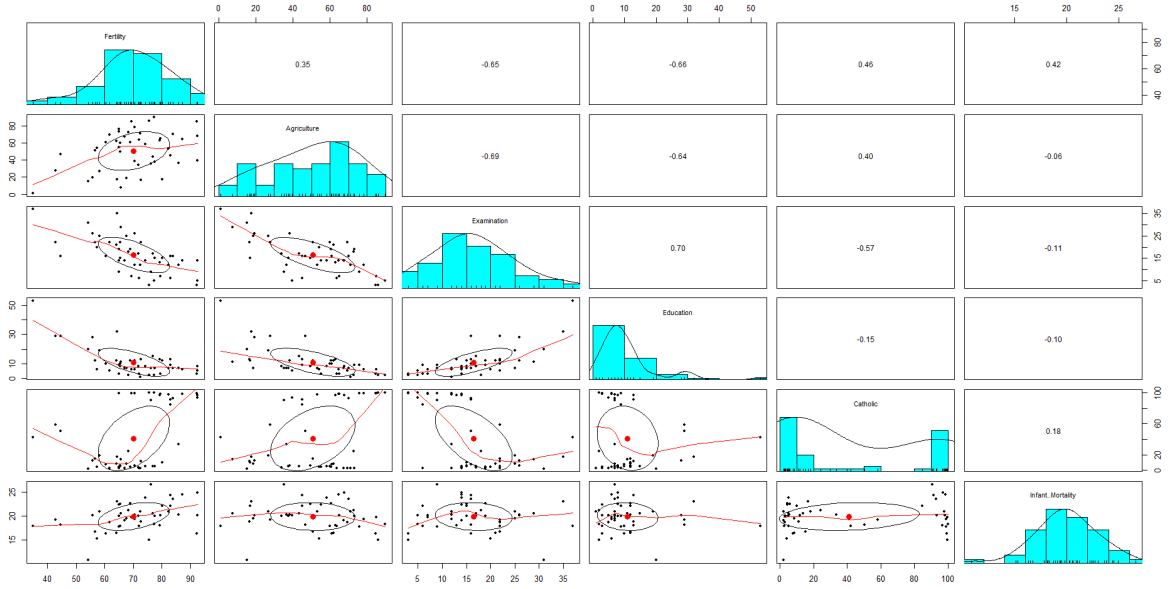
	X	Fertility	Agriculture	Examination	Education	Catholic	Infant..Mortality
30	Yverdon	65.4	49.5	15	8	6.10	22.5
41	Le Locle	72.7	16.7	22	13	11.22	18.9

Como podemos notar, hay cierta similitud respecto a algunas variables, no con todas, pero sí con aquellas que la matriz de correlación clasificó como significativas. Entonces, si hacemos una lista de las provincias más similares entre sí (aquellas marcadas con un 1) tenemos lo siguiente:

- Martigwy - Delemont
- Grandson - La Chauxdfnd
- Moudon - Val de Ruz
- Neuveville - ValdeTravers
- Vevey - Rive Droite
- Yverdon - Le Locle

Estos resultados fueron obtenidos con **k-NN**, pero, ¿Obtendremos los mismos resultados con otros métodos?, habrá que verificarlo.

Empecemos haciendo un gráfico de pares para visualizar la relación entre las variables.



Aquí podemos notar que la distribución Catholic es bimodal porque tiene dos picos, esto sugiere que los datos tienden a agruparse en dos grupos distintos, lo cual sería interesante de investigar con clustering. De momento, centrémonos en que hay una correlación fuerte con la variable **Examination** y **Education**.

Naive Bayes funciona mejor cuando hay etiquetas (clases conocidas) en el conjunto de entrenamiento para clasificar nuevas observaciones, sin embargo, nosotros tenemos medidas numéricas que representan esas etiquetas, ¿qué haremos entonces?, convertir esos valores numéricos en etiquetas.

	X	Fertility	Agriculture	Examination	Education	Catholic	Infant..Mortality	Fertility_Category
11	Veveyse	87.1	64.5	14	6	98.61	24.5	Alta
12	Aigle	64.1	62.0	21	12	8.52	16.5	Baja
13	Aubonne	66.9	67.5	14	7	2.27	19.1	Media
14	Avenches	68.9	60.7	19	12	4.43	22.7	Media
15	Cossonay	61.7	69.3	22	5	2.82	18.7	Baja
16	Echallens	68.3	72.6	18	2	24.20	21.2	Media
17	Grandson	71.7	34.0	17	8	3.30	20.0	Media

Esta estrategia nos permite obtener el siguiente modelo:

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = data_train, y = as.factor(data_train_labels))

A-priori probabilities:
as.factor(data_train_labels)
   Baja   Media    Alta
0.34375 0.31250 0.34375
```

```
Conditional probabilities:
                            Fertility
as.factor(data_train_labels)      [,1]       [,2]
                       Baja  0.4412648 0.07187088
                       Media 0.6170435 0.05595810
                       Alta  0.8509091 0.09324753

                            Agriculture
as.factor(data_train_labels)      [,1]      [,2]
                       Baja  0.5371341 0.2224034
                       Media 0.7301695 0.1678065
                       Alta  0.5186441 0.1799500

                            Examination
as.factor(data_train_labels)      [,1]      [,2]
                       Baja  0.5294118 0.1493932
                       Media 0.3000000 0.1822580
                       Alta  0.2754011 0.1200294

                            Education
as.factor(data_train_labels)       [,1]       [,2]
                       Baja  0.21153846 0.14442295
                       Media 0.09423077 0.07221147
                       Alta  0.14860140 0.06147287

                            Catholic
as.factor(data_train_labels)       [,1]       [,2]
                       Baja  0.05660798 0.05545175
                       Media 0.23250894 0.40830488
                       Alta  0.71697868 0.37626203

                            Infant..Mortality
as.factor(data_train_labels)      [,1]      [,2]
                       Baja  0.4689298 0.2218163
                       Media 0.5740506 0.1590849
                       Alta  0.7577675 0.1368345
```

Aquí podemos notar que hay variables con más "peso" que otras, por ejemplo Catholic, Infant. Mortality y Agriculture son las que más peso tienen en la etiqueta de alta fertilidad. Si nos damos cuenta, aquí se ve reflejada la correlación negativa con **Examination**, ya que justamente tiene muy poco peso en cuanto a fertilidad, es decir, en las colonias con más valor de dicha variable experimentan una **fertilidad menor**.

Continuando con Bayes, ahora probemos el modelo para hacer una predicción y con eso evaluaremos la precisión del modelo.

```
> print(confusion_matrix)
          Actual
Predicted Baja Media Alta
    Baja     3     4    0
    Media    1     1    1
    Alta     1     0    4
```

```
Precisión del modelo: 0.5333333
```

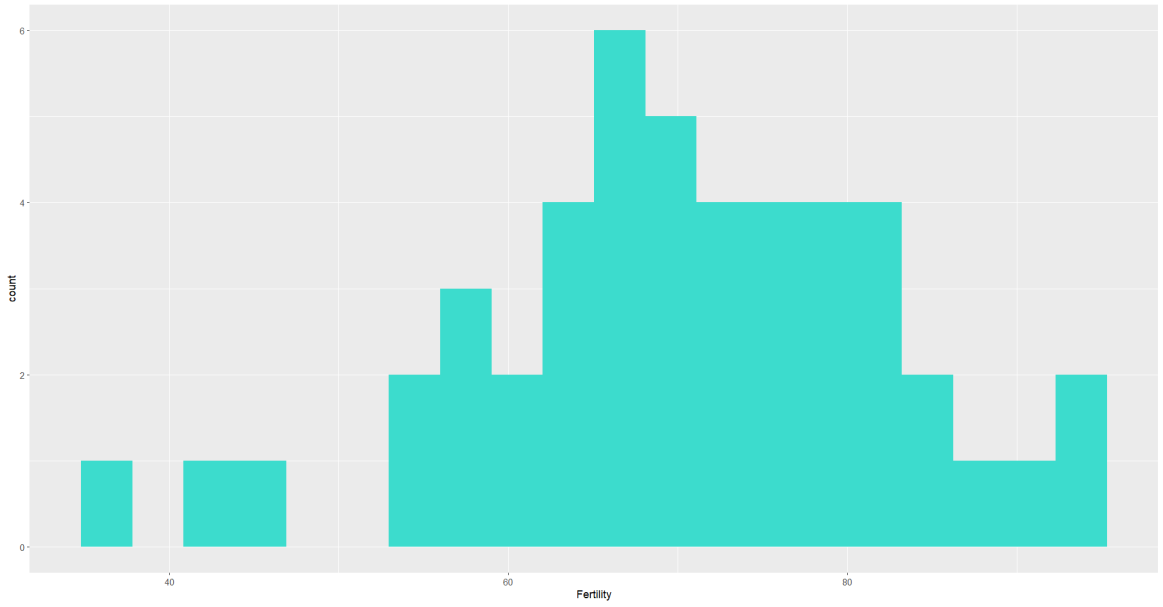
Bueno, la predicción pese a no ser buena tampoco es mala, pero es lo que se podría esperar dadas las pocas entradas del dataset. Aún así, no dejemos que eso nos quite ánimo y comparemos los resultados con los de k-NN.

```
> print(alta_predichas)
            X Fertility Agriculture Examination Education Catholic Infant..Mortality Fertility_Category
35     Monthey      79.4        64.9           7         3    98.22             20.2              Alta
37      Sierre      92.2        84.6           3         3    99.46             16.3              Alta
38        Sion      79.3        63.1          13        13    96.83             18.1              Alta
43  Val de Ruz      77.6        37.6          15         7     4.97             20.0              Alta
47 Rive Gauche      42.8        27.7          22        29    58.33             19.3              Baja
```

Estas son las provincias clasificadas como "Alta" por el modelo (con la predicción). Aquí podemos notar (otra vez) que la precisión no es tan buena, pues, Rive Gauche tiene una fertilidad baja, pero si observamos las otras colonias, vemos que en efecto hay una fuerte similitud en **Catholic** y **Agriculture**, y un poco también en **Infant. Mortality**. Sin embargo, sólo una de las colonias coincide con las marcadas por k-NN (**Val de Ruz**). Quizá se deba a que el modelo se centró en las colonias con una fertilidad de entre 77 y 79.

Para no quedar inconformes, habrá que comparar resultados con regresión lineal.

Empecemos con un histograma de la fertilidad.



Se distribuye de forma casi normal pero con un par de brechas en la izquierda.

Una vez que ajustamos el modelo de regresión obtenemos los siguientes resultados

```
> summary(model_lm)

Call:
lm(formula = Fertility ~ Agriculture + Examination + Education +
    Catholic + Infant..Mortality, data = data[1:32, ])

Residuals:
    Min      1Q  Median      3Q     Max
-14.457  -2.990   1.162   3.730  10.146

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       65.63258   10.93362   6.003 2.44e-06 ***
Agriculture       -0.14432    0.08474  -1.703  0.10048
Examination       -0.44015    0.29889  -1.473  0.15286
Education         -0.43153    0.35680  -1.209  0.23738
Catholic           0.11178    0.03602   3.103  0.00458 **
Infant..Mortality  1.01042    0.37417   2.700  0.01202 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.242 on 26 degrees of freedom
Multiple R-squared:  0.7204,    Adjusted R-squared:  0.6666
F-statistic:  13.4 on 5 and 26 DF,  p-value: 1.648e-06
```

Nuestra R cuadrada es de 0.72, lo cual es bueno considerando el tamaño de los datos (el explica el 72% de variabilidad). En cuanto a los valores respectivos de las variables, básicamente dice lo mismo que los análisis previos, Catholic e Infant. Mortality tienen una correlación positiva significativa con la fertilidad.

Para encontrar las provincias similares haremos una predicción y luego agrupamos según la similitud de sus residuales (los residuales vendrían siendo la diferencia entre los valores reales y los predichos por el modelo).

```
> print(similar_provinces)
   Fertility Agriculture Examination  Education   Catholic Infant..Mortality
33 0.7356522  1.00000000  0.05882353 0.01923077 1.00000000         0.4746835
34 0.6173913  0.87005650  0.26470588 0.09615385 0.98937149         0.5443038
35 0.7721739  0.71977401  0.11764706 0.03846154 0.98180889         0.5949367
38 0.7704348  0.69943503  0.29411765 0.23076923 0.96760347         0.4620253
39 0.6156522  0.42033898  0.67647059 0.21153846 0.03546244         0.6012658
40 0.5339130  0.07344633  0.76470588 0.19230769 0.11895759         0.6139241
41 0.6556522  0.17514124  0.55882353 0.23076923 0.09269290         0.5126582
42 0.5113043  0.18531073  0.94117647 0.59615385 0.15094532         0.7721519
43 0.7408696  0.41129944  0.35294118 0.11538462 0.02881962         0.5822785
44 0.5669565  0.19774011  0.64705882 0.11538462 0.06642821         0.5506329
```

En conclusión, las colonias que comparten similitud según nuestro modelo de regresión lineal son las siguientes:

- Herens
- Martigwy
- Monthey
- Sion

- Boudry
- La Chauxdfnd
- Le Locle
- Neuchatel
- Val de Ruz
- ValdeTravers

Otra vez tenemos ciertas disparidades con respecto a los métodos anteriores, sin embargo, me parece que este método es el más eficaz ya que abarca un rango más amplio de similitud (no como el método anterior).

Como extra, anexo los siguientes gráficos que pueden resultar representativos de la información que ya se ha mencionado pero vista de una forma más bonita.