# Proyecto final: Análisis de alergias de usuarios de un recetario web

INSTITUTO POLITÉCNICO NACIONAL

Alumno: Hector Isaac Roman Vazquez ———— Boleta: 2021670099
Alumno: Jesús Eduardo Guijarro Saldaña ——— Boleta: 2021670220
Alumno: Omar Montoya Romero ——————- Boleta: 2021670251
Alumno: Arnold Torres Maldonado ——————- Boleta: 2021670117

Docente: Hector Alejandro Acuña Cid

16 de Junio del 2023

## Abstract

This report reflects the data analysis process of a web system developed by the authors of the document, the objective of the analysis is identify which is the allergy that is most repeated, as well as the number of most common allergies and use linear regression and the KNN algorithm to make a comparison between the results obtained from both methods. The conclusion is the linear regression generated very useful result to predict the number of users with six allergies which is negative due the collected data. Also the most repeated allergy is to pickles and the more frecuent allergies number it's zero allergies.

## Resumen

En este reporte se plasma el proceso de análisis de los datos de un sistema web desarrollado por los autores del documento con los objetivos de identificar cual es la alergia más repetida y el número de alergias más común, así como hacer uso de las herramientas de regresión lineal y el algoritmo KNN para observar las diferencias entre los mismos. Se concluye que la regresión lineal generó un resultado más útil para predecir el número de usuarios con seis alergias el cual es negativo debido a la cantidad de datos recolectados. Además, la alergia más repetida es pepinillos y el número de alergias más común es de cero.

## Introducción

En la siguiente investigación se utiliza una base de datos de aproximadamente 600 usuarios de los cuales se pueden recuperar los datos de los ingredientes a los que son alérgicos e incluso si alguno de los usuarios no cuenta con alergias, para de esta manera poder realizar una análisis utilizando regresión lineal y el algoritmo KNN, para poder hacer una comparativa entre ambos, mediante la metodología CRISP-DM para lograr un mejor análisis.

## Marco teórico

Una alergia es una reacción de su sistema inmunitario hacia algo que no molesta a la mayoría de las demás personas. Quienes tienen alergias suelen ser sensibles a más de una cosa. Las sustancias que suelen causar reacciones son: [4]

- Polen
- Ácaros del polvo
- Esporas de moho
- Caspa de animales
- Alimentos
- Picaduras de insectos
- Medicinas

Las alergias pueden provocar una serie de síntomas como goteos nasales, estornudos, picazón, sarpullidos, edema (hinchazón) o asma. Las alergias van de leves a severas. Una reacción severa llamada anafilaxia puede resultar fatal. Los médicos usan pruebas de piel y sangre para diagnosticar las alergias. Los tratamientos incluyen medicinas, inyecciones y evitar las sustancias que causan las alergias. [4]

Una alergia alimentaria es una reacción del sistema inmunitario que ocurre poco después de haber ingerido un determinado alimento. Incluso una pequeña cantidad del alimento que causa la alergia puede ocasionar signos y síntomas, como problemas digestivos, urticaria o inflamación de las vías respiratorias. En algunos casos, una alergia alimentaria puede ocasionar síntomas graves o, incluso, una reacción que puede poner en riesgo la vida, llamada anafilaxia. [5]

Se calcula que la alergia alimentaria afecta al 8 por ciento de los niños menores de 5 años y hasta al 4 por ciento de los adultos. A pesar de que no existe cura, algunos niños superan sus alergias alimentarias cuando crecen. [5]

La alergia alimentaria puede fácilmente confundirse con una reacción mucho más frecuente llamada intolerancia alimentaria. Si bien es molesta, la intolerancia alimentaria es una afección de menor gravedad que no involucra al sistema inmunitario si no que, significa que el cuerpo de la persona no puede digerir bien determinado alimento, o que un alimento en particular le irrita el sistema digestivo. Entre los síntomas de la intolerancia alimentaria, se incluyen los siguientes: náuseas, gases, retortijones, dolor abdominal, diarrea, irritabilidad y dolor de cabeza. [2]

los “Datasets”, son fundamentales para la revolución del procesamiento de datos por la que estamos pasando, y muchas veces, más sencillos de lo que parecen. Un Dataset no es más que un conjunto de datos tabulados en cualquier sistema de almacenamiento de datos estructurados. El término Dataset hace referencia a una única base de datos de origen, la cual se puede relacionar con otras, cada columna del Dataset representa una variable y cada fila corresponde a cualquier dato que estemos tratando. [1]

Existen cuatro tipos de Datasets catalogados según su origen y formato, los cuales son usados según las necesidades de los modelos de datos a trabajar. [1]

- Archivo: es un fichero independiente en el que se almacena toda la información con la que se va a trabajar del Dataset. Tiene como ventajas, la seguridad y rapidez para el trabajo con los datos, ya que siempre se explotan y se visualizan de manera local, sin embargo la escalabilidad y conexión con otros Datasets que no están almacenados en la misma máquina se dificulta.
- Folder: es la suma de diferentes Datasets almacenados en una misma carpeta, los cuales están conectados entre ellos. Estos archivos deben compartir un mismo formato como puede ser .csv, .mif o dxf.
- de datos: este tipo de Dataset puede llegarse a confundir con el archivo, pero se diferencia por su nivel de especialidad, es decir, son bases de datos con formatos específicos diseñadas para programas puntuales. Por ejemplo las bases de datos de Oracle, las cuales solo funcionan para sus desarrollos.
- Web: es la compilación de datos que se almacenan dentro de un sitio web del Dataset. El nombre que se le asigna por defecto a este Dataset es el correspondiente a la URL.

# Materiales y métodos

## R y RStudio

Para la investigación se usó el lenguaje de programación R, R es un lenguaje de programación estadística ampliamente utilizado para análisis de datos, modelado estadístico y visualización. Es una plataforma gratuita y de código abierto que ofrece una amplia gama de herramientas y paquetes especializados para el procesamiento y análisis de datos. RStudio, por otro lado, es un entorno de desarrollo integrado (IDE) diseñado específicamente para trabajar con R. Proporciona una interfaz gráfica fácil de usar que facilita la escritura, ejecución y depuración de código en R. RStudio también ofrece características adicionales como paneles de visualización, administración de proyectos y colaboración en línea, lo que lo convierte en una herramienta popular entre los usuarios de R [8].

## Modelo CRISP-DM

Para desarrollar los modelos y análisis correspondiente hicimos uso de la metodología Cross Industry Standard Process for Data Mining (CRISP-DM por sus siglas), esta es una metodología para la minería de datos que se utiliza comúnmente en el análisis de datos empresariales. La metodología consta de seis fases [6]:

### Comprensión del negocio o problema

El objetivo de esta fase es alinear los objetivos del proyecto de data mining con los objetivos del negocio. Tratando así de evitar embarcarnos en un proyecto de minería de datos que no produzca ningún efecto real en la organización.

En esta fase deberemos ser capaces de:

- Establecer los objetivos de negocio.
- Evaluar la situación actual.
- Fijar los objetivos a nivel de minería de datos.
- Obtener un plan de proyecto.

### Comprensión de los datos

Dos puntos clave en esta fase: conocer los datos, estructura y distribución, y la calidad de estos.

En esta fase deberemos ser capaces de:

- Ejecutar procesos de captura de datos.
- Proporcionar una descripción del juego de datos.
- Realizar tareas de exploración de datos.
- Gestionar la calidad de los datos, identificando problemas y proporcionando soluciones.

### Preparación de los datos

El objetivo final de esta fase es obtener los datos finales sobre los que aplicarán los modelos.

En esta fase deberemos ser capaces de:

- Establecer el universo de datos con los que trabajar.
- Realizar tareas de limpieza de datos.
- Construir un juego de datos apto para ser usado en modelos de minería de datos.
- Integrar datos de fuentes heterogéneas si es necesario.

### Modelado

El objetivo último de esta fase es construir un modelo que nos permita alcanzar los objetivos del proyecto.

En esta fase deberemos ser capaces de:

- Seleccionar las técnicas de modelado más adecuadas para nuestro juego de datos y nuestros objetivos.
- Fijar una estrategia de verificación de la calidad del modelo.
- Construir un modelo a partir de la aplicación de las técnicas seleccionadas sobre el juego de datos.
- Ajustar el modelo evaluando su fiabilidad y su impacto en los objetivos anteriormente establecidos.

**Evaluación del modelo**

En esta fase se centra en evaluar el grado de acercamiento del modelo a los objetivos.

En esta fase deberemos ser capaces de:

- Evaluar el modelo o modelos generados hasta el momento.
- Revisar todo el proceso de minería de datos que se ha llevado hasta este punto.
- Establecer los siguientes pasos a tomar, tanto si se trata de repetir fases anteriores como si se trata de abrir nuevas líneas de investigación.

**Implementación**

El objetivo de esta fase es realizar la implementación de los resultados obtenidos de forma que sea propagado a los usuarios finales así como el mantenimiento de este una vez la implementación haya finalizado.

En esta fase deberemos ser capaces de:

- Diseñar un plan de despliegue de modelos y conocimiento sobre nuestra organización.
- Realizar seguimiento y mantenimiento de la parte más operativa del despliegue.
- Revisar el proyecto en su globalidad con el objetivo de identificar lecciones aprendidas. [9]

## La regresión lineal

Por otra parte, la regresión lineal es una técnica conocida de modelado estadístico que se utiliza para analizar la relación entre dos variables continuas. Se utiliza para predecir el valor de una variable dependiente a partir del valor de una o más variables independientes, esta asume que la relación entre las variables es lineal, lo que significa que el cambio en la variable independiente tiene un cambio proporcional en la variable dependiente. Una forma de aplicar la regresión lineal es mediante el método de mínimos cuadrados, este es un método matemático utilizado para encontrar una línea de regresión que mejor se ajusta a los datos [7]. El objetivo del método es minimizar la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por la línea de regresión.

## El algoritmo KNN (K vecinos más cercanos)

Es un algoritmo de clasificación supervisada que usa la proximidad (por ejemplo la distancia euclidiana) para hacer clasificaciones y predicciones sobre agrupaciones de datos partiendo del supuesto de que se pueden encontrar puntos similares cerca el uno del otro. Para hacer predicciones sobre clasificaciones se toma el promedio de los k vecinos más cercanos, estas predicciones se pueden hacer para valores continuos mientras que las clasificaciones se hacen solo para valores discretos.

## Conjunto de datos de alergias

El conjunto de datos que se analizó proviene de un sistema web de recetas donde los usuarios se dan de alta y especifican los ingredientes a los cuales son alérgicos, el sistema los guarda en una base de datos es ahí de donde nosotros obtuvimos el conjunto para trabajar. Algunas de las ventajas de este algoritmo son que es fácil de implementar dado que solo requiere de un valor de k que elige el analista y una métrica de distancia. Otra es que representa un buen material didáctico dada su simplicidad y precisión por lo dicho anteriormente [9].

# Experimentación y resultados

## Comprensión del problema

Se genera una comparación entre los resultados que se pueden obtener de el análisis de usuarios con distintas alergias por medio de tablas de frecuencia, regresion lineal y el algoritmo KNN haciendo uso de la metodologia CRISP-DM. Los objetivos fueron predecir el número de usuarios que iban a tener 6 alergias, encontrar el ingrediente al que los usuarios en general son más alérgicos y el número de alergias más frecuente entre los usuarios

Los datos se obtienen utilizando un conjunto de datos obtenidos mediante un sistema web, el sistema consta de un recetario, donde una vez que se ingresa a la plataforma, el usuario puede seleccionar si se tienen alergias a algun alimento, dichos datos se almacenan en una base de datos y de ahi fue posible extraerlos en como un dataset para poder realizar el análisis.

## Entendimiento de los datos

Usamos un conjunto de datos con 7 variables las cuales eran:

- Names: los nombres de los usuarios registrados en el sistema.
- lastNames: los apellidos de dichos usuarios.
- email: correo electrónico de los usuarios.
- password: contraseña de los usuarios.
- alergias: ingredientes a los que cada usuario era alérgico.
- numero_alergias: número de ingredientes a los que el usuario era alérgico.

## Preparación de los datos

El archivo CSV que contenía los datos tenía las columnas separadas por punto y coma y los ingredientes estaban separados por comas de la siguiente manera:

```
Names;lastNames;email;alergias;numero_alergias
Valeria;Flores;valeriaflores117@example.com;Camarones, Aceitunas;2
```

Por lo tanto para hacer los análisis correspondientes importamos dicho archivo especificando el caracter que separaba las columnas con la línea de código siguiente:

```
data <- read.csv("usuarios.csv", sep=";")
```

## Modelado

### Encontrando la alergia más frecuente

Para encontrar la alergia más frecuente hicimos un vector que contenía todos los ingredientes repetidos.

```
# strsplit para dividir los nombres separados por coma
# en elementos individuales
ingredientes_separados <- strsplit(data$alergias, ",")
# unlist para crear un vector para poner los nombres separados
ingredientes <- unlist(ingredientes_separados)
```

Para empezar realizamos una tabla de frecuencias para los ingredientes que nos permitió observar cuáles se repetían más, después esta tabla se ordenó de manera descendente según su frecuencia para después tomar el valor en la posición número 1, por último se imprime el resultado.

```r
# Cargar el dataset
data <- read.csv("usuarios.csv", sep=";")

# Calcular la frecuencia de las alergias
# strsplit para dividir los nombres separados por coma
# en elementos individuales
ingredientes_separados <- strsplit(data$alergias, ",")

# unlist para crear un vector para poner los nombres separados
ingredientes <- unlist(ingredientes_separados)
tabla_de_frecuencia <- table(ingredientes)

# ordenar de mayor a menor las frecuencias
alergias_ordenadas <- sort(tabla_de_frecuencia, decreasing = TRUE)
alergia_mas_comun <- names(alergias_ordenadas)[1] # alergia ubicada en el #1

# Mostrar el resultado
cat("La alergia más común es:", alergia_mas_comun)
```

```
## La alergia más común es: Pepinillos
```

### Encontrando el número de alergias más frecuente

Para esta sección también se hizo uso de una tabla de frecuencias que ordenamos de manera descendente según el número de veces que se repetía una cantidad de alergias. De nueva cuenta se seleccionó el valor en la posición número 1 para después imprimir el resultado.

```r
# Cargar el dataset
data <- read.csv("usuarios.csv", sep=";")
# elegir la columna numero_alergias del DS
num_alergias <- data$numero_alergias
# tabla de frecuencias de los numeros de alergias
frecuencia <- table(num_alergias)
# ordenar frecuencias de la tabla de manera descendente
frecuencia_num_ordenadas <- sort(frecuencia,decreasing = TRUE)
# elegir la alergia #1
num_alergia_mas_frecuente <- names(frecuencia_num_ordenadas)[1]

# imprimir solución
cat("Es más común que los usuarios tengan", num_alergia_mas_frecuente, "alergias")
```

```
## Es más común que los usuarios tengan 0 alergias
```

### Predicciones usando regresión lineal y KNN

Para realizar la predicción usamos las frecuencias que arrojó el análisis de la sección anterior y los hicimos dataframe, pasamos la columna num_alergias para el eje x y la de la frecuencia para el eje y. Después generamos un modelo de regresión lineal e hicimos la predicción de la frecuencia de usuarios que iban a tener 6 alergias.

```r
# Cargar el dataset
data <- read.csv("usuarios.csv",sep=";")
# elegir la columna numero_alergias del DS
num_alergias <- data$numero_alergias
# tabla de frecuencias de los numeros de alergias
frecuencia_num_alergias <- table(num_alergias)
```

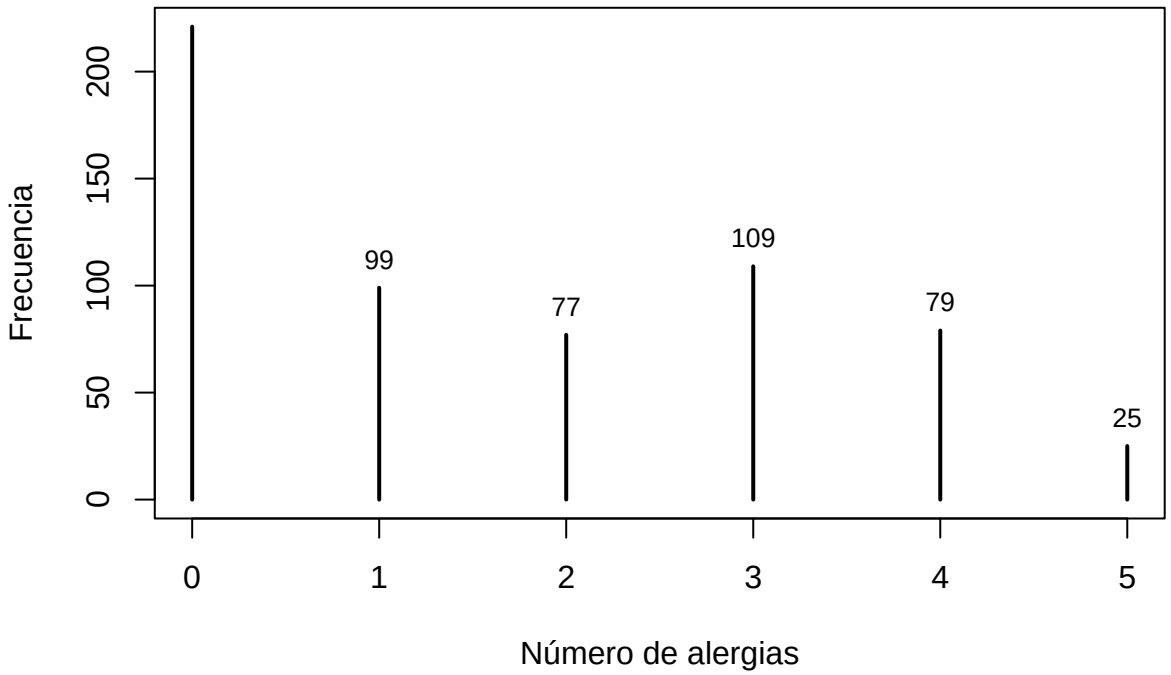
```r
# ordenar frecuencias de la tabla de manera descendente
frecuencia_num_alergias_ordenadas <- sort(frecuencia_num_alergias, decreasing = TRUE)

# Predicción utilizando Regresión Lineal
datos <- data.frame(frecuencia_num_alergias_ordenadas)

x <- as.numeric(datos$num_alergias)
y <- datos$Freq

# Ajuste del modelo de regresión lineal
modelo_rl <- lm(y ~ x)

# Realizar predicción utilizando regresión lineal
numero_de_alergias <- 7 # posición 7 del vector de alergias ==> numero_de_alergias=6
prediccion_rl <- predict(modelo_rl, data.frame(x = numero_de_alergias))
```

Hicimos lo mismo de arriba para la predicción usando el algoritmo KNN:

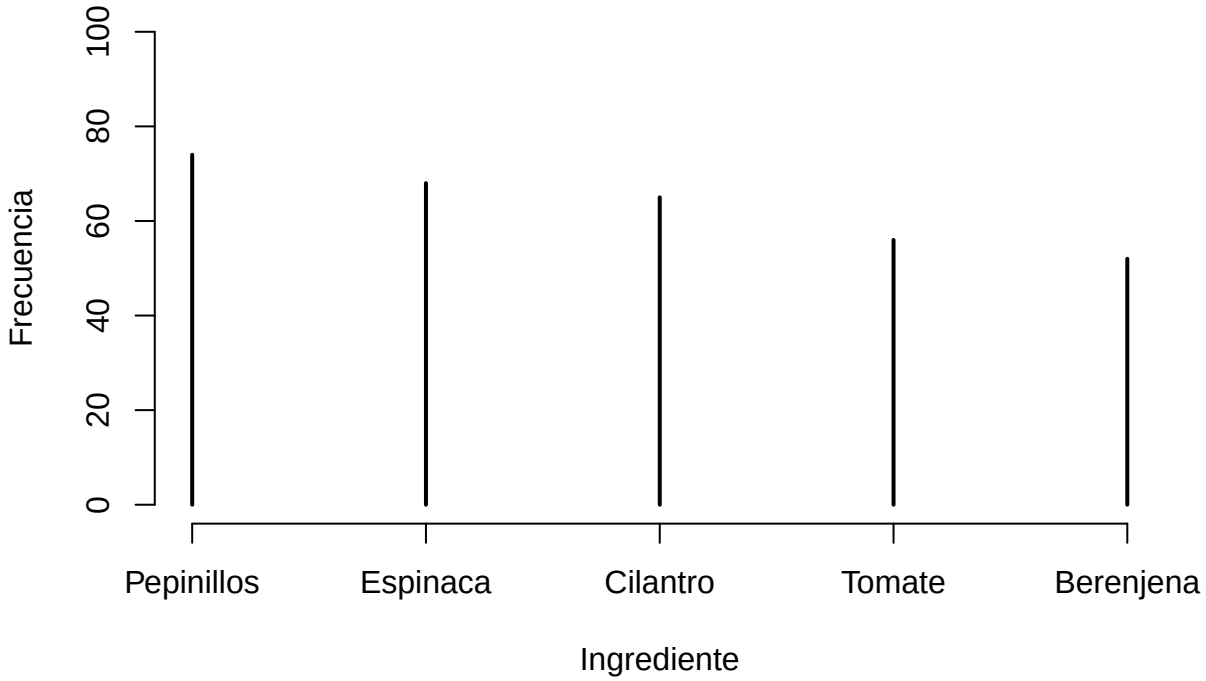
```r
library(class) # biblioteca que contiene la función knn()

# Cargar el dataset
data <- read.csv("usuarios.csv",sep=";")
# elegir la columna numero_alergias del DS
num_alergias <- data$numero_alergias
# tabla de frecuencias de los numeros de alergias
frecuencia_num_alergias <- table(num_alergias)
# ordenar frecuencias de la tabla de manera descendente
frecuencia_num_alergias_ordenadas <- sort(frecuencia_num_alergias, decreasing = TRUE)

# Predicción utilizando Regresión Lineal
datos <- data.frame(frecuencia_num_alergias_ordenadas)

x <- as.numeric(datos$num_alergias)
y <- datos$Freq

# Predicción utilizando KNN
datos_entrenamiento <- data.frame(
  x = as.numeric(as.character(datos$num_alergias)),
  y = as.numeric(as.character(datos$Freq))
)

datos_prueba <- data.frame(x = numero_de_alergias)

# Ajuste del modelo KNN
k <- 2 # Valor de K
prediccion_knn <- knn(train = datos_entrenamiento[, "x"],
                      test = datos_prueba,
                      cl = datos_entrenamiento$y,
                      k = k)
```

## Resultados del modelo

El análisis que se realizó a partir de tablas de frecuencia muestra que el número de alergias más frecuente es el de cero con 221 usuarios, seguido de tres, uno, cuatro, dos y cinco alergias.




También muestra que el ingrediente al cual los usuarios son más alérgicos son los pepinillos seguido de las espinacas, el cilantro, el tomate y la berenjena.




El modelo de regresión lineal generado arrojó un ajuste de R cuadrado de 0.7 lo que indica que el data frame usado contenía una muestra suficiente de datos para lograr la predicción propuesta.

```
summary(modelo_rl)
```

```
##
## Call:
## lm(formula = y ~ x)
##
```

```
## Residuals:
##       1       2       3       4       5       6
##  41.048 -39.638 -18.324  -7.010  22.305   1.619
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  211.267     29.954   7.053  0.00213 **
## x            -31.314      7.691  -4.071  0.01521 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.18 on 4 degrees of freedom
## Multiple R-squared:  0.8056, Adjusted R-squared:  0.757
## F-statistic: 16.58 on 1 and 4 DF,  p-value: 0.01521
```

Dicho esto, el modelo predijo que el número de usuarios con seis alergias sería de -8 de lo cual podemos deducir que no existirían usuarios con ese número de alergias (al menos basado en los datos utilizados). En la gráfica al final de esta sección no se alcanza a apreciar ese valor porque es negativo y la función usada para hacer la gráfica usa solo valores positivos.

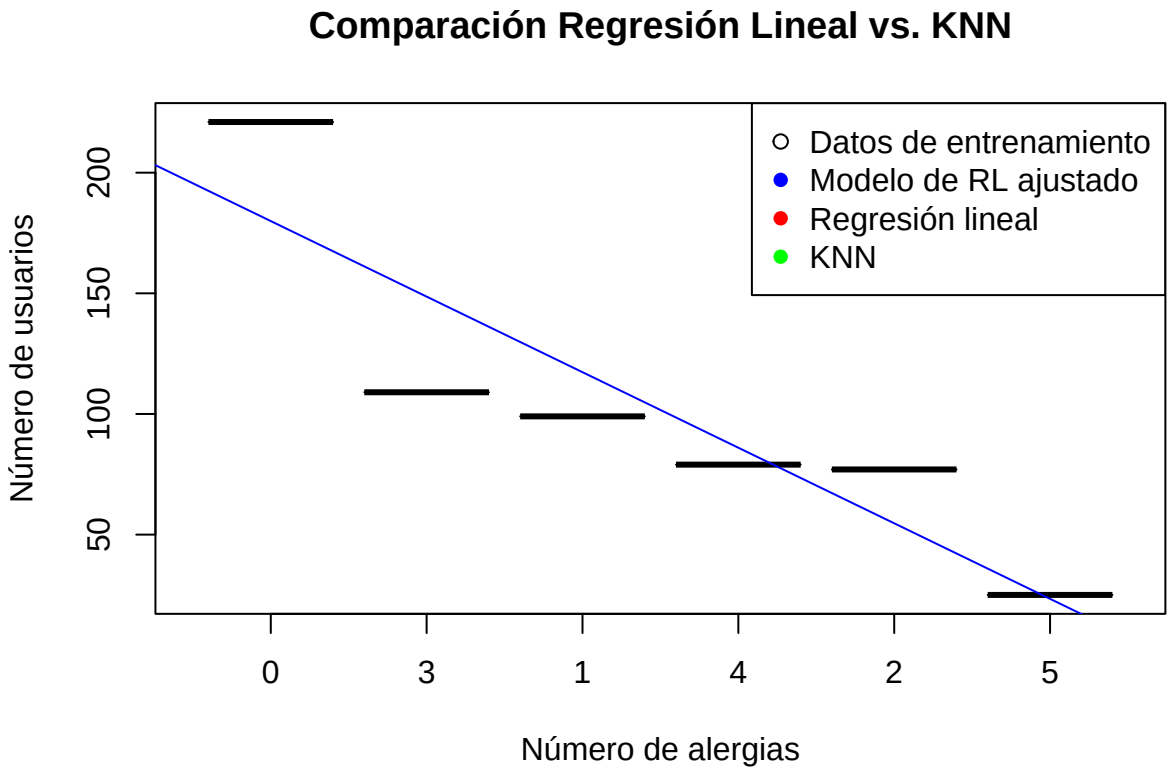
```
# Imprimir la predicción utilizando regresión lineal
cat("Predicción usando Regresión Lineal para numero_de_alergias = 6:", prediccion_rl, "\n")
```

```
## Predicción usando Regresión Lineal para numero_de_alergias = 6: -7.933333
```

Por otro lado el modelo generado por el algoritmo KNN con $k = 1, 2, 3, 4$ arrojó como predicción que el número de usuarios con seis alergias sería de 1, 1, 3 y 2 respectivamente, por lo que dejamos $k = 2$ como valor final debido a que estaba más cerca a la predicción que arrojó el modelo de regresión lineal. Este resultado se muestra en la siguiente gráfica al final de esta sección.

```
# Ajuste del modelo KNN
k <- 2 # Valor de K
prediccion_knn <- knn(train = datos_entrenamiento[, "x"],
                      test = datos_prueba,
                      cl = datos_entrenamiento$y,
                      k = k)

# Imprimir la predicción del modelo de generado por KNN
cat("Predicción usando KNN para numero_de_alergias = 6:", prediccion_knn, "\n")
```

```
## Predicción usando KNN para numero_de_alergias = 6: 3
```




## Conclusiones

Es comprensible que el algoritmo KNN y la regresión lineal usan métodos distintos para hacer predicciones pero no debería variar tanto el resultado de la predicción que arrojan. Por lo tanto afirmamos que estas dos técnicas tienen un resultado más preciso según el conjunto de datos donde se apliquen.

## Referencias

[1] "¿Qué son los Datasets? [4 sitios donde encontrarlos]", KeepCoding Bootcamps, 27-feb-2020.

[2] "¿En qué se diferencia una alergia alimentaria de una intolerancia alimentaria?", Kidshealth.org. [En línea]. Disponible en: https://kidshealth.org/es/parents/allergy-intolerance.html. [Consultado: 07-jun-2023].

[3] "Alergia", Ear, Nose and Throat, 2002.

[4] "Alergias", Mayoclinic.org, 05-ago-2022. [En línea]. Disponible en: https://www.mayoclinic.org/es-es/diseases-conditions/allergies/symptoms-causes/syc-20351497. [Consultado: 07-jun-2023].

[5] A. A. Medina, S. M. Armentia, y S. F. Cortés, "Alergia alimentaria", Medicine, vol. 13, núm. 28, pp. 1572–1578, 2021.

[6] IBM. (2021). Conceptos básicos de ayuda de CRISP-DM. [En línea]. Disponible en: https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview. [Consultado: 07-jun-2023].

[7] Online Statistics Education: A Multimedia Course of Study. (2019). Linear Regression. Disponible en: [http://onlinestatbook.com/2/regression/intro.html]. [Consultado: 07-jun-2023].

[8] Wickham, H., & Grolemund, G. (2017). R for Data Science. O'Reilly Media.

[9] IBM. (2021). ¿Qué es KNN?. [En línea]. Disponible en: https://www.ibm.com/mx-es/topics/knn. Consultado: 07-jun-2023].