Reporte: minería de datos sobre alergias de usuarios de un sistema web de recetas

Jesus, Hector, Arnold, Omar

2023-06-09

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 mndefinedás 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 X es la herramienta más precisa, la alergia más repetida es a los pepinillos y el número de alergias más común es cero alergias.

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

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].

Para desarrollar el modelo de regresión lineal 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]:

1. Comprensión del problema: esta fase inicia por comprender los requisitos y objetivos del proyecto para establecer un contexto.
2. Entendimiento de los datos: consiste en explorar y analizar los datos disponibles para obtener una comprensión detallada de su contenido, calidad y estructura.

3. Preparación de los datos: fase donde se realizan tareas de limpieza, transformación y selección de datos para garantizar la calidad y adecuación de los mismos para el análisis.
4. Modelado: periodo de aplicar técnicas de minería de datos y construir modelos predictivos o descriptivos utilizando algoritmos apropiados.
5. Evaluación: fase para poner a prueba el rendimiento y la eficacia de los modelos construidos mediante métricas y pruebas para verificar su validez y precisión.
6. Despliegue: donde se implementan los resultados y las conclusiones del proyecto en el entorno.

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 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.

Experimentación y resultados

Comprensión del problema

Se trata de un análisis de datos llevado a cabo usando tablas de frecuencia de regresión lineal y del algoritmo KNN usando el conjunto de datos del sistema web de recetas.

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;password;alergias;numero_alergias
Valeria;Flores;valeriaflores117@example.com;Vwx234@#$;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.

```
# 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.

```
# 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
```

Uso del algoritmo KNN

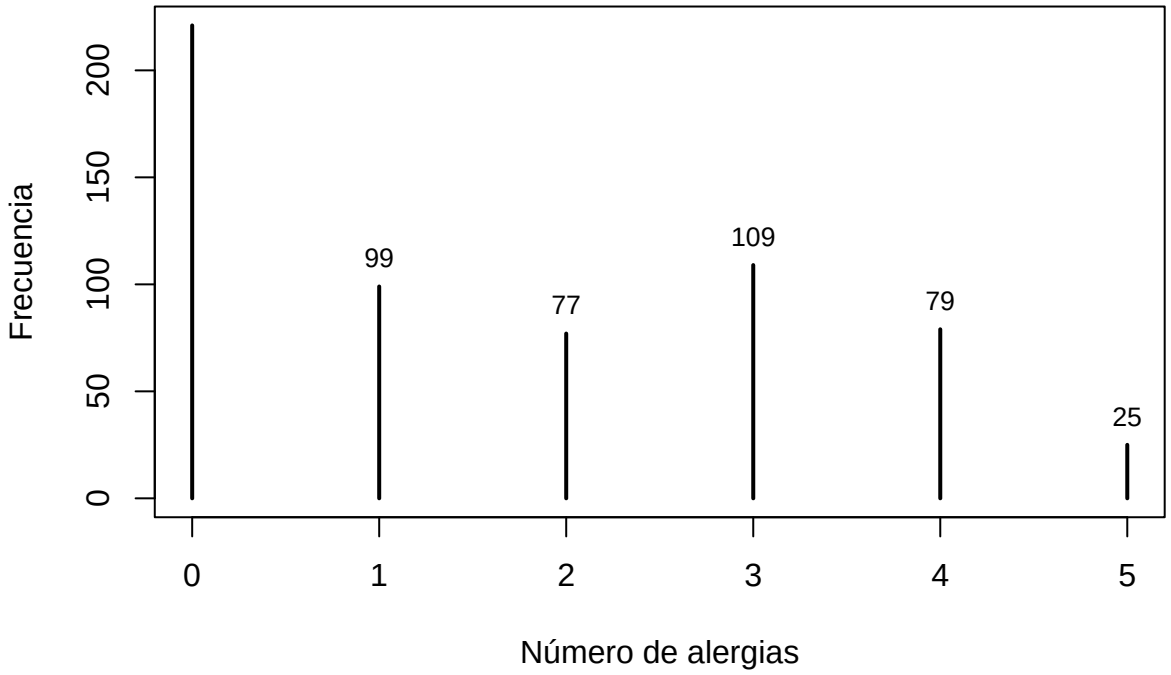
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

Uso de la técnica de regresión lineal

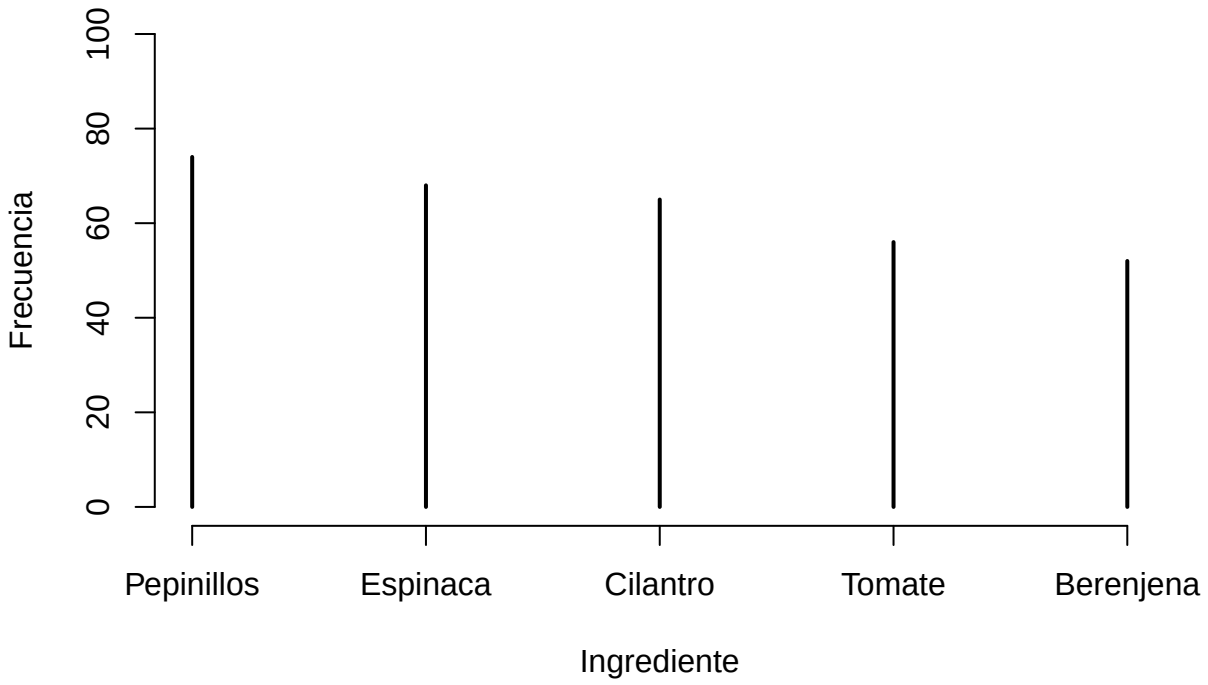
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

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.



Conclusiones

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

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.