Liliana Ochoa Echeverri

Cc 1036927400

Avance 2 proyecto

Descripción del problema predictivo a resolver

La clasificación ATC es un sistema europeo de codificación de sustancias farmacéuticas y medicamentos con arreglo al sistema u órgano efector y al efecto farmacológico, las indicaciones terapéuticas y la estructura química de un fármaco [1p], de allí la importancia de estudiar el índice de ventas de sus ocho diferentes categorías para lograr obtener datos que indiquen el mayor consumo por órgano afectado ya que el ATC, recoge el sistema u órgano sobre el que actúa [3w]. En base a esto, se desea predecir si la cantidad vendida y exportada desde el sistema de punto de vista en el individuo farmacia es óptima y suficiente para la demanda que tendrá [2k].

Dataset que se va a utilizar

El dataset que se va a utilizar fue extraído de Kaggle y se llama Datos de Ventas Farmacéuticas [2k], este dataset contiene cerca de consta de 600 000 datos transaccionales recopilados en 6 años (período 2014-2019), que indican la fecha y la hora de venta, la marca del medicamento farmacéutico y la

cantidad vendida, exportados desde el sistema de punto de venta en el individuo farmacia. El grupo seleccionado de medicamentos del conjunto de datos (57 medicamentos) se clasifica en las siguientes categorías del Sistema de clasificación químico terapéutico anatómico (ATC):

- M01AB - Productos antiinflamatorios y antirreumáticos, no esteroides, derivados del ácido acético y sustancias relacionadas
- M01AE - Productos antiinflamatorios y antirreumáticos, no esteroides, derivados del ácido propiónico
- N02BA - Otros analgésicos y antipiréticos, ácido salicílico y derivados
- N02BE/B - Otros analgésicos y antipiréticos, Pirazolonas y Anilidas
- N05B - Medicamentos psicolépticos, Medicamentos ansiolíticos
- N05C - Drogas psicolépticas, hipnóticas y sedantes
- R03 - Medicamentos para enfermedades obstructivas de las vías respiratorias
- R06 - Antihistamínicos para uso sistémico

Los datos de ventas son remuestreados a los períodos horarios, diarios, semanales y mensuales. Losdatos ya están preprocesados, donde el procesamiento incluyó la detección y el tratamiento de valores atípicos y la imputación de datos faltantes.

Además, el dataset contiene 4 archivos con datos diarios, semanales, anuales y saleshourly contenidos en 44 columnas donde se podrán predecir mejor las estadísticas según su predicción.

Análisis de los datos (Avance 2)

Al hacer un análisi profundo de los datos, se determina que lo primero que se debe realizar es una **limpieza de los mismos** por medio de Hot Encoding, de tal manera que se implemente la estrategia de tener todos los datos numérico con la estrategia de crear una columna para cada valor distinto que exista en la característica que estamos codificando y, para cada registro, marcar con un 1 la

columna a la que pertenezca dicho registro y dejar las demás con 0. Esto se hizo con Meses, día, nombre del día, año y hora.

Después de tener los datos “limpio”, se realizó una **validación de los datos**, donde se trabajó con algoritmos de clustering, dado que son algoritmos basados en distancias, para ello se **escalaron** los datos para prescindir de las unidades de medida de las diferentes features.

Seguido de esto, se llegó a la conclusión de trabajar con el **algoritmo de K-Means**, pues el objetivo de este proyecto es llegar a conocer la categoría farmacológica que más se venda y por ende el órgano que más consume dicho f´ramaco, y esto se puede hacer con el algoritmo de K-means, ya que este trabaja iterativamente para asignar a cada “punto” (las filas de nuestro conjunto de entrada forman una coordenada) uno de los “K” grupos basado en sus características. Son agrupados en base a la similitud de sus features (las columnas). Como resultado de ejecutar el algoritmo tendremos:

- Los “centroids” de cada grupo que serán unas “coordenadas” de cada uno de los K conjuntos que se utilizarán para poder etiquetar nuevas muestras.
- Etiquetas para el conjunto de datos de entrenamiento. Cada etiqueta perteneciente a uno de los K grupos formados.

Para terminar con este avance del proyecto se trabajó con la **métríca de evaluación** del algotitmo **llamada Elbow** consiste básicamente en verificar la evolución de la suma de los cuadrados del error para varios valores de K y verificar cual es el que brinda un mejor agrupamiento.

Lo que queda faltando del proyecto es implementar el algoritmo de clasificación para llegar al objetivo de esta práctica, el cual veremos en el informe final.

Los gráficos yel avance de lo anterior descrito están reflejados en el Notebook

Bibliografía

- inec.gob.pa, «CONCEPTOS Y DEFINICIONES,» [En línea]. Available: https://www.inec.gob.pa/archivos/P4361CONCEPTOS.pdf.
- M. M. H. RODRIGUEZ, «APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS,» 2015. [En línea]. Available: https://m.riunet.upv.es/bitstream/handle/10251/65082/memoria.pdf?sequence=1&isAllowed=y. [Último acceso: 28 Marzo 2022].
- G. M. Hidalgo, «ELEMENTOS QUE PARTICIPAN EN LA INCIDENCIA DE ACCIDENTES DE,» [En línea]. Available: http://creandoconciencia.org.ar/enciclopedia/accidentologia/relevamiento-de-rastros/ELEMENTOS-QUE-PARTICIPAN-EN-LA-INCIDENCIA-DE-ACCIDENTES-DE-TRANSITO.pdf. [Último acceso: 03 Marzo 2022].
- M. Meléndrez, «Factores relativos a las causas de los accidentes,» [En línea]. Available: https://1library.co/article/factores-relativos-a-las-causas-de-los-accidentes.q05pen9y. [Último acceso: 28 Marzo 2022].