Proyecto 2: Aprendizaje no supervisado

A00517244 Camila Navarro
A01705747 Enrique García
A01197399 Diana Cadena
Tecnológico de Monterrey
Ingeniería en Ciencia de Datos y Matemáticas
Monterrey, Nuevo León, México

RESUMEN

El aprendizaje no supervisado consiste es un proceso de agrupamiento, con base en la comparación de distancias que hay entre distintos datos. El cual es aplicado para este documento, donde se aborda un análisis de sentimiento sobre la opinión de el efecto que les genero una serie de medicamentos a gente que le fueron recetados. Cabe destacar, el uso de técnicas para procesamiento de lenguaje en formato textual.

Palabras clave *– nlp, área de salud, aprendizaje no supervisado, clustering*

ACM Reference Format:
A00517244 Camila Navarro, A01705747 Enrique García, A01197399 Diana Cadena. 2022. Proyecto 2: Aprendizaje no supervisado. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1. INTRODUCCIÓN

El *Machine Learning* es el estudio de métodos y herramientas de utilidad para identificar patrones en los datos. [16] El identificar estos patrones sirve ya sea para comprender mejor lo que describen o para hacer predicciones. Esta área de estudios tiene una gran variedad de aplicaciones y entre ellas se encuentra el área de salud, el enfoque general del proyecto. En su mayoría, las aplicaciones del ML entran dentro del aprendizaje supervisado, no supervisado y de refuerzo. [16] En este caso, se enfocará el reporte en el aprendizaje no supervisado ya que se busca enfocarse en el procesamiento del lenguaje natural.

El uso de *Machine Learning* dentro del área de salud es amplia y puede llegar a considerarse compleja, sobre todo si es que se están analizando una gran cantidad de datos, lo cual sería lo ideal debido a la complejidad del cuerpo humano. Generalmente hablando, la aplicación del ML es de gran utilidad para hacer más eficientes los procesos y consultas dentro de esta área, sobre todo tomando en consideración que la oferta debe de cumplir con la demanda. [2]

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, July 2017, Washington, DC, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

El diagnóstico pronto y asertivo de ciertas enfermedades o condiciones es de suma importancia para un buen tratamiento y posible recuperación.

No obstante, el dataset en específico a utilizarse no es específicamente sobre alguna enfermedad en particular sino sobre los medicamentos recetados con base a ciertos síntomas y las experiencias de lxs pacientes. Este dataset tiene como nombre *UCI ML Drug Review dataset*, el cual ha sido analizado múltiples veces con enfoques distintos. Como se menciona anteriormente, el dataset provee información acerca la satisfacción general de lx paciente con el medicamento así como posibles efectos secundarios y su experiencia. Como complementario a esto, se incluye una calificación dentro del rango [1, 10]. Además, claramente, se incluye el nombre del medicamento recetado así como la condición que se presentaba para su misma receta. [3]

Los datos que proporciona dicho dataset fueron recopilados de distintos sitios farmacéuticos web, con la intención de estudiar y analizar distintos aspectos. Hablando específicamente sobre las reseñas contenidas en el archivo, estas hacen referencia tanto a la eficacia del medicamento así como sus efectos secundarios. Al contar con información en lenguaje natural, el análisis de los datos resulta ser más interesante pero complejo a la vez. Sin embargo, el análisis de los mismos resulta en información valiosa además de ayudar en la toma de decisiones y la mejora del seguimiento de la salud pública con base a la experiencia colectiva. [3]

Algunos de los enfoques de análisis que se le han dado a estos datos son los siguientes.

1. **Clasificación.** Predecir la condición de lx paciente con base en la reseña.
2. **Regresión.** Predecir la calificación otorgada al medicamento con base a la reseña.
3. **Análisis de sentimiento.** Identificar los elementos de la reseña que hacen que sea de utilidad para otras personas, así como lxs pacientes que tienden a dejar reseñas negativas. Además, determinar si la reseña es positiva, neutral o negativa.
4. **Visualización de los datos.** Identificar tanto los medicamentos como padecimientos que se encuentran en el dataset.

Un aspecto importante a tener en consideración para que los modelos de predicción funcionen correctamente es que las bases de datos deben de ser lo más robustas posibles. [2] Entre más datos

haya para entrenar y evaluar el modelo, mejor serán los resultados arrojados por el mismo. Asimismo, el modelo no es capaz de predecir o identificar correctamente la relación entre distintas condiciones o variables si no se encuentra presente en los datos. Cabe mencionar que el uso de ML dentro del área de salud no sirve como un reemplazo de los diagnósticos y procedimientos llevados a cabo en la actualidad sino como una herramienta que aumenta y mejora su desempeño. [2]

2. CONCEPTOS PREVIOS

Se habló del *Machine Learning* de manera general previamente, pero dentro de este existen dos medios de aprendizaje: supervisado y no supervisado. Anteriormente, se abordó la parte del aprendizaje supervisado pero ahora toca abordar el aprendizaje no supervisado. De igual manera, existen distintos algoritmos propios del aprendizaje no supervisado en *machine learning*.

2.1. Clustering

En el aprendizaje no supervisado, los conjuntos de datos no se encuentran clasificados o etiquetados por lo que tampoco se tiene un resultado conocido. Al implementar este tipo de aprendizaje, se deben deducir las estructuras presentes en los datos por medio de procesos matemáticos. [13] Dentro del aprendizaje no supervisado, se encuentran los algoritmos de agrupamiento, los cuales particionan los datos en k grupos de acuerdo a alguna función de similitud o distancia. [11]

Los algoritmos de agrupamiento a su vez pueden ser catalogados como paramétricos o no paramétricos; en este reporte se estarán abordando los algoritmos no paramétricos. Los algoritmos de agrupamiento no paramétricos se dividen en tres grupos básicos: jerárquicos, particionales y basados en densidad. [11] Los algoritmos jerárquicos van diviendo el conjunto de datos por niveles; si es aglomerativo, se unen dos grupos del nivel anterior y si es divisivo, se separan. [11]

Los algoritmos particionales dividen los datos en grupos desde un inicio y se van moviendo de un grupo a otro, según se optimice la función objetivo establecida. Por último, los algoritmos de densidad se basan en la distribución de densidad de los puntos para dividir los mismos en grupos. De esta manera, cada grupo tiene una alta densidad de puntos dentro de sí y entre los grupos aparecen zonas de baja densidad. [11]

A continuación, se explicarán los distintos algoritmos de agrupamiento a utilizarse para realizar los procesos correspondientes del proyecto.

2.1.1. ***K-means***.

El algoritmo de k-medias, o k-promedios, es un algoritmo de agrupamiento particional y es considerado como uno de los más simples y reconocidos. Este algoritmo divide las instancias en k grupos, cuya k debe ser definida previo a la implementación del algoritmo. Para la división de los grupos, el algoritmo se basa en minimizar el cálculo de las distancias entre los puntos y un k centroide definido.[11]

El funcionamiento de este algoritmo se puede enumerar de la siguiente forma.

1. Se define la cantidad de k grupos a realizar y, por ende, se definen los k centroides a tener.
2. Se calcula la distancia entre cada punto y cada k centroide.
3. Se clasifican los puntos al k grupo con base a la menor distancia al k centroide.
4. Se recalculan los k centroides a través del promedio de los puntos pertenecientes a ese k grupo.
5. Una vez establecido el nuevo k centroide, se vuelven a calcular las distancias entre los puntos y el k centroide.
6. En caso de ser necesario, se reasignan los puntos al k grupo con base a la menor distancia al k centroide.
7. Este proceso se realiza de manera iterativa hasta que el cálculo de los k centroides no cambie; es decir, permanezca igual de una iteración a la siguiente.

El objetivo principal de este algoritmo es minimizar una función de error cuadrático, lo que garantiza encontrar un mínimo local de la función que a su vez depende de los parámetros iniciales del algoritmo. [14] No obstante, cabe mencionar que este algoritmo no necesariamente encuentra el mejor conjunto de agrupaciones. Esto debido a que la selección inicial de los k grupos afecta significativamente el resultado.

$$J = \Sigma_{j=1}^{k} \Sigma_{i=1}^{n} ||x_i^{(i)} - c_j||^2 \tag{1}$$

Para el cálculo de la distancia entre los puntos y los k centroides, suele utilizarse la distancia de Minkowski; en particular, su expresión Euclidiana.

$$D = \left(\Sigma_{i=1}^{n} |p_i - q_i|^p)\right)^{\frac{1}{p}} \tag{2}$$

donde,

$$p = \begin{cases} 1, \ distancia\ Manhattan \\ 2, \ distancia\ Euclidiana \\ otro, \ distancia\ Minkowski \end{cases} \tag{3}$$

Como se menciona con anterioridad, los datos se clasifican dentro de un grupo k de acuerdo a la distancia mínima que haya, y no pueden pertenecer a más de uno a la vez. [8] Una vez se clasifiquen todos los datos, el cálculo del centroide k se realiza de acuerdo a un promedio.

$$c_k = \Sigma_{i=1}^{n} \frac{(x_i, y_i)}{n} \tag{4}$$

donde n es la cantidad de instancias (x_i, y_i) asignadas al centroide c_k.

Para la programación del algoritmo en sí, utilizando la librería de *scikitlearn*, este se establece de la siguiente manera y contiene los hiperparámetros mostrados.

Kmeans(*n_clusters, init, n_init, max_iter, tol, verbose, random_state, copy_x, algorithm*)

n_clusters → cantidad tanto de agrupaciones como de centroides a realizar; default = 8
init → método de inicialización; default = k-means++
n_init → cantidad de veces a correr el algoritmo k-means con distintos centroides iniciales; default = 10
max_iter → iteraciones máximas del algoritmo por corrida; default = 300
tol → tolerancia con respecto a la norma de Frobenius para declarar convergencia; default = 0.0001
verbose → modo de verbosidad; default = 0
random_state → determina números aleatorios para la inicialización del centroide, si se establece un número se hace determinista; default = None
copy_x → no se modifican los datos originales al ir calculando distancias; default = True
algorithm → algoritmo k-means a utilizar; default = lloyd

Los hiperparámetros pueden especificarse con la finalidad de mejorar las métricas de desempeño del modelo. En caso de no ser especificados, se toman los valores *default*.

2.1.2. *Agglomerative Hierarchical Clustering*.

El clustering aglomerativo jerárquico es el tipo más común de agrupación jerárquica utilizada para agrupar objetos en clusters en función de su similitud. También se conoce como AGNES (Agglomerative Nesting). El algoritmo comienza tratando cada objeto como un grupo único. A continuación, los pares de grupos se fusionan sucesivamente hasta que todos los grupos se han fusionado en un gran grupo que contiene todos los objetos. [8] El resultado es una representación en forma de árbol de los objetos, denominada *dendrograma*.

El funcionamiento en sí de este algoritmo se puede enumerar de la siguiente manera.

1. Se inicia calculando la información de similitud (distancia) entre cada par de objetos en el conjunto de datos; suele ser una matriz de distancias.
2. Se usa de la función de vinculación para agrupar objetos en un árbol jerárquico en función de la información de distancia generada en el paso anterior. Esta función de vinculación es usualmente el mínimo de las distancias.
3. Se repiten el paso 1 y 2 hasta que se hayan agrupado todos los elementos o hasta donde se determine según un criterio de decisión.
4. Se clasifica cada uno de los puntos en los clusters.

En cuanto a la preparación de los datos, por ser un algoritmo basado en distancias se recomienda estandarizar las variables antes de realizar análisis posteriores. La estandarización hace que las variables sean comparables cuando se miden en diferentes escalas. [8]

Para el cálculo de distancias, la elección de una métrica adecuada influirá en la forma en que se formen los grupos, ya que algunos elementos pueden estar cerca de uno de otro según un tipo de distancia y más lejanos según otro. Algunas métricas de uso común para la agrupación jerárquica son:

$$D = \left(\Sigma_{i=1}^{n} |p_i - q_i|^p\right)^{\frac{1}{p}} \tag{5}$$

donde,

$$p = \begin{cases} 1, \ distancia\ Manhattan \\ 2, \ distancia\ Euclidiana \end{cases} \tag{6}$$

Para programar el algoritmo utilizando la librería de *scikitlearn*, este se establece de la siguiente manera y contiene los hiperparámetros mostrados.

AgglomerativeClustering(*n_clusters, affinity, memory, connectivity, compute_full_tree, linkage, distance_threshold, compute_distances*)

n_clusters → cantidad de agrupaciones a realizar; default = 2
affinity → métrica utilizada para calcular la relación entre los datos; default = euclidean
memory → almacenamiento en caché de la salida del cálculo del árbol; default = None
connectivity → matriz de conectividad, define los vecinos de cada muestra de acuerdo a una estructura dada; default = None
compute_full_tree → detiene la elaboración del árbol en *n_clusters*; default = auto
linkage → criterio relacional a utilizar, determina qué distancia se emplea entre las instancias para establecer las agrupaciones; default = ward
distance_threshold → umbral de distancia a partir del cual no se fusionan las agrupaciones; default = None
compute_distances → calcula las distancias entre agrupaciones, sirve para la visualización del dendograma; default = False

Los hiperparámetros pueden especificarse con la finalidad de mejorar las métricas de desempeño del modelo. En caso de no ser especificados, se toman los valores *default*.

2.2. Análisis de sentimentos

El análisis de sentimientos es un proceso automatizado textual y visual que clasifica la información detectando, extrayendo y clasificando opiniones, según la polaridad de los datos (positivo, negativo y neutral), pero también según los sentimientos y emociones (enojado, feliz, triste, etc.) , urgencia (urgente, no urgente) e incluso intenciones (interesado vs. no interesado). Se puede realizar en diferentes niveles, tales como: nivel de documento, capturando los sentimientos generales expresados en el texto; nivel de oración, clasificando la polaridad de cada oración en el texto; nivel

de característica, analizando la polaridad de opiniones sobre características/atributos del objeto y nivel de aspecto, encontrando y agregando sentimientos sobre entidades mencionadas dentro de los documentos o aspectos de ellos.

Un proceso de análisis de sentimientos se puede estructurar en cinco procedimientos principales [6]:

1. Extracción de datos
2. Procesamiento previo
3. Detección de sentimientos
4. Clasificación de sentimiento
5. Informe de polaridad, que muestra los resultados de un análisis de sentimientos de varias maneras posibles.

Existen distintos métodos para el análisis de sentimientos de forma no supervisada, los dos acercamientos principales son los siguientes:

Métodos basados en diccionarios.

Los métodos basados en diccionarios (o lexicones, del inglés lexicon) son aquellos donde la tarea de clasificación se realiza utilizando métodos no supervisados de base semántica. El enfoque, también conocido como diccionario o basado-en-el-conocimiento, calcula la orientación sentimental de un documento a partir de la semántica de las palabras o frases que lo componen. Como es un enfoque no supervisado, no requiere un entrenamiento inicial de datos, sino que utiliza una lista predefinida de palabras, a las que se asocia un sentimiento específico. El sentimiento general del texto se calcula en función del recuento de palabras positivas y negativas presentes en él. El análisis de sentimientos utilizando enfoques basados en diccionarios tiene las ventajas de una gran simplicidad de comprensión, implementación y eficiencia, tanto en el uso de recursos computacionales como en la capacidad de predicción. Sin embargo, este enfoque tiene una limitación en términos de generalización, ya que cada solución se crea de acuerdo con el contexto en el que se aplicará. Este enfoque no requiere datos etiquetados, sino que implica la construcción de un diccionario léxico, lo que constituye uno de sus principales desafíos, ya que la necesidad de considerar los contextos de su aplicación dificulta el uso de un diccionario único para cualquier situación. A esta dificultad se suma el gran volumen de datos que normalmente hay que analizar, así como la variabilidad del lenguaje utilizado en distintas regiones, contextos y situaciones.

Métodos basados en relaciones lingüísticas.

Además de los métodos basados en diccionarios, en los sistemas de clasificación no supervisada existen modelos basados en relaciones lingüísticas. Estos métodos buscan las secuencias en los textos que puedan expresar ciertas opiniones y sentimientos con una mayor probabilidad, extrayendo las palabras que las forman para luego ser usadas en la categorización del texto global. Para ello, se obtiene la categoría gramatical de las palabras, conocida en inglés como *parts-of-speech*, y se determina si dichos patrones expresan una opinión positiva, negativa, o de algún otro sentimiento. Por último, el sentimiento global del texto se calcula mediante algún tipo de función matemática [15].

2.2.1. ***Clustering en análisis de sentimientos.***

Técnicas comunes en análisis de sentimientos incluyen el aprendizaje supervisado, que utilizan algoritmos de clasificación como SVM, árboles de decisión y Naive Bayes. Todos estos algoritmos requieren datos de entrenamiento preetiquetados para identificar un modelo que mejor se ajuste a la relación entre el conjunto de atributos y la etiqueta de clase de los datos de entrada. Sin embargo, el etiquetar los datos cuando la base no tiene esta información puede conllevar muchas veces un proceso arduo manual en que un supervisor humano etiquete primero los datos.

El clustering es una técnica que divide datos en grupos (clusters) que son significativos, útiles o ambas cosas. El proceso de agrupamiento tiene como objetivo descubrir clusters, y por lo tanto presenta una visión general de las clases en una colección de objetos, en nuestro caso, palabras. Agrupar objetos basados simplemente en las propiedades internas significa que no hay información adicional o supervisada requerida, por lo que presenta ventajas ante los métodos de aprendizaje supervisado o incluso sobre los métodos de diccionario.

Las técnicas clustering se pueden clasificar en agrupamiento jerárquico y particional. Los algoritmos jerárquicos crean una descomposición jerárquica del conjunto de datos, como se explicó anteriormente sobre el algoritmo de *clustering aglomerativo jerárquico*. Si permitimos que los grupos tengan subgrupos, entonces obtenemos un agrupamiento jerárquico, que es un conjunto de grupos anidados, es decir, clústeres que se organizan como un árbol. Por el contrario, un agrupamiento particional, como el de *K-means* simplemente divide el conjunto de objetos de datos en subconjuntos que no se superponen, de modo que cada objeto de datos está exactamente en un solo subconjunto. En la aplicación del análisis de sentimiento binario (*positivo vs negativo*), no hay subclusters anidados tanto en positivo como en negativo, solo pueden ser uno o el otro. Por lo tanto, la técnica de agrupamiento jerárquico no es la más adecuada para el análisis de sentimientos. Por otro lado, el agrupamiento particional es más apropiado para análisis de sentimiento binario porque solo busca dividir el corpus en dos grupos de sentimientos. Sin embargo, en caso de buscar hacer un análisis más complejo que abarque más de dos categorías, se pueden llegar a usar algoritmos jerárquicos.

No obstante, dado que la investigación adopta *k-means* como el algoritmo de agrupamiento por excelencia para análisis binario de sentimientos, un nuevo desafío debe ser enfrentado, es decir, el problema de la inestabilidad de los resultados de la agrupación causados por la selección aleatoria de los centroides iniciales. K-means es extremadamente sensible a la elección de los centroides iniciales. Acercamientos previos a esta problemática lidian con ella dándole un peso previo a las frases para eliminar la incertidumbre de los centroides iniciales [7] o bien, iterando sobre los centroides aleatorios.

3. METODOLOGÍA

Para la resolución del problema en cuestión, se llevó a cabo la metodología CRISP-DM (Cross Industry Standard Process for Data Mining). Esta consta de 6 fases: comprensión del negocio, estudio y comprensión de los datos, análisis de los datos y selección de características, modelado, evaluación y despliegue. A continuación, se detallará el uso de cada fase dentro del reporte. Cabe destacar que la solución para este reto tuvo sustento en el trabajo realizado por Huster, K. [5]

3.1. Comprensión del negocio.

En la industria farmacéutica un área de oportunidad muy grande es la de satisfacción al cliente. Si bien se sabe que este ha sido un problema por un largo tiempo, la introducción de medios digitales, foros de internet y redes sociales, ha creado una nueva dificultad en la regulación de las reseñas de medicamentos. Los últimos años han sido testigos del rápido crecimiento de un servicio de atención médica digital: la consulta médica en línea (OMC por sus siglas en inglés). A pesar de su popularidad, muchas plataformas de OMC han encontrado problemas en la adopción inicial y el uso continuado entre los pacientes, principalmente por una falta de confianza hacia estas plataformas [10]. Este es un fenómeno complejo que involucra consideraciones tanto interpersonales como tecnológicas, sin embargo, filtros tecnológicos que puedan facilitar la interacción del cliente con la herramienta son muy útiles para esta industria. El análisis de opiniones sobre los diversos aspectos de las reseñas de medicamentos puede proporcionar información valiosa, ayudar en la toma de decisiones y mejorar el seguimiento de la salud pública al revelar la experiencia colectiva de los usuarios. Por ejemplo, una recomendación automatizada de medicamentos a los clientes basada en su condición actual o filtros para controlar la desinformación en la sección de comentarios. Esta información compartida en OMCs tiene también gran utilidad para estudios de mercado.

Lo que se propone realizar con estos datos es predecir la connotación de cada reseña hecha en el sitio, por medio del procesamiento de lenguaje natural. Es decir, de acuerdo a las palabras escritas, ¿se puede predecir si la experiencia del usuario fue buena, neutral o mala? Cabe mencionar que se está trabajando con reseñas en inglés, por lo que el análsis está enfocado en este idioma.

3.2. Estudio y comprensión de los datos.

El dataset escogido fue creado como parte de una investigación de análisis de sentimiento. Se utilizaron datos de dos páginas web independientes para recuperar las reseñas de los usuarios y las calificaciones sobre la experiencia con los medicamentos. *Drugs.com* es, según el proveedor, el sitio web de información farmacéutica más grande y más visitado que proporciona información tanto para consumidores como para profesionales de la salud. Proporciona reseñas de usuarios sobre medicamentos específicos junto con la condición relacionada y una calificación del usuario de 10 estrellas que refleja la satisfacción general del usuario [4].

Del mismo modo, *Druglib.com* es un recurso de información sobre medicamentos para consumidores y profesionales de la salud. Contiene un número considerablemente menor de reseñas, pero se proporcionan de una manera más estructurada, pues tienen una sección de Efectividad y Efectos secundarios. Se recopilaron los comentarios y las calificaciones de los usuarios de ambas páginas mediante un *web scraper* automático. El rastreo de estos dominios dio como resultado dos conjuntos de datos que comprenden 215063 reseñas de Drugs.com y 3551 reseñas de Druglib.com. Ambos conjuntos de datos se dividieron en particiones de entrenamiento y prueba de acuerdo con un esquema de muestreo aleatorio estratificado con la proporción de 75 % y 25 %.

3.3. Preparación de datos. Análisis de los datos y selección de características.

Para realizar la parte exploratoria, se unieron ambos datasets. Además, se realizaron distintas gráficas con la finalidad de comprender el comportamiento de los datos así como la correlación entre los mismos. A continuación, se muestran distintas descripciones de los datos presentes así como las gráficas y apoyos visuales correspondientes.

Descripción general del dataset.

- Número de instancias: 215,063 registros
- Valores faltantes: 1,194, valores
- No. de atributos cuantitativos: 2, siendo estos *rating* y *usefulCount*.
- No. de atributos categóricos: 5, siendo estos *uniqueID*, *drugName*, *condition*, *review* y *date*.

Descripción de atributos

- uniqueID: Identificador de paciente; único para cada entrada por lo que no es realmente de utilidad y se puede reemplazar con el índice.
- drugName: Nombre del medicamento recetado; cuenta con 3,671 registros únicos. Es nuestra variable a predecir.
- condition: Nombre del padecimiento reportado; cuenta con 916 registros únicos. Es de gran utilidad para el proceso de aprendizaje y correlación.
- review: Reseña del medicamento; comentarios acerca de cómo fue su experiencia con el medicamento, por qué fue recetado y sobre el cambio que notaron al haberlo tomado. Es de utilidad para el proceso de aprendizaje no supervisado.
- rating: Calificación otorgada 1-10 al medicamento; esta es dada por los pacientes complementando así el review de lo que observaron al haberlo tomado. Es de gran utilidad para el proceso de aprendizaje.
- date: Fecha en la que se escribió la reseña. No agrega mucha utilidad al modelo, a menos que se decida descartar las entradas de mayor antigüedad.
- usefulCount: Cantidad de usuarios que encontraron útil la reseña, siendo un número basado en sí la reseña les fue útil o no y la frecuencia de los resultados afirmativos. Nos es de utilidad para saber qué reseñas tienen un mayor peso.

3.3.1. Limpieza. En cuanto a la limpieza de los datos, se eliminaron todas las columnas que no necesitamos, pues nuestro análisis será solo en base a las reseñas, por lo que solo mantuvimos las columnas *Review* y *Rating*. Estas no contenían datos nulos ni anómalos.

A00517244 Camila Navarro
A01705747 Enrique García
A01197399 Diana Cadena

Previo a la generación del modelado de aprendizaje automático, fue necesario realizar un procesamiento de los datos de texto. El objetivo de la limpieza era que se pudieran mantener las palabras que mantuvieran un significado para el análisis que se realizaría posteriormente. Existen diferentes técnicas estándar de limpieza para procesamiento de lenguaje.

Para ello se creo una función de limpieza del texto, en el que se eliminaron caracteres especiales, dígitos, signos de puntuación, palabras más comunes en el inglés y se realizó la lematización de palabras. La lematización hace referencia al proceso lingüístico de reducir palabras que vengan de la misma raíz morfológica, es decir, eliminar conjugaciones. Por ejemplo, de «dijeron» a «decir», para que el modelo pueda identificar palabras con diferentes conjugaciones pero un mismo significado. Para esto se hizo uso de la librería NLTK (Natural Language Processing Toolkit), en la que vienen ya integradas funciones de lematización para el idioma inglés [1]. De este modo, al implementar la función se puede obtener un arreglo de listas con el texto de cada reseña, el cual cabe destacar que se separaría posteriormente en palabras.

3.3.2. Vectorización de palabras. A continuación, se inició el modelado de una red neuronal a través de la implementación de Word2Vec de la API Gensim's [12], con el cual se logró establecer un vector para cada palabra. La técnica Word2Vec consiste en una familia de algoritmos que utiliza un modelo de red neuronal para aprender asociaciones de palabras a partir de un gran corpus de texto. De esta forma, se asigna cada palabra distinta con una lista particular de números (vector). Estos vectores están calculados de forma que la similitud coseno entre los vectores indica el nivel de la similitud semántica entre las palabras representada por dichos vectores. Word2vec toma como entrada un corpus de texto y produce un espacio vectorial $n - dimensional$ (en nuestro caso, específicamos 300). Luego asigna cada palabra única en el corpus a un vector correspondiente en el espacio. Los vectores de palabras están colocados en el espacio vectorial de forma que las palabras que comparten contextos comunes en el corpus están localizadas cerca unas de otras en el espacio [9].

3.4. Modelado

Habiendo establecido estos vectores, se le aplicaría el modelo de KMeans con el cual se pudiera establecer los diferentes vectores en diferentes clusters. Representando estos las siguientes categorías:

- neutral: 0
- negativo: 1
- positivo: 2

Con el cual posteriormente, se estableció un resultado de la cercanía que tenía cada vector hacia su centroide y derivado a ello el coeficiente de sentimiento que cada palabra tenía. Lo que para nuestro caso represento que entre más cercano a cero este fuese entonces la palabra no tenía mucho impacto para el análisis, por el contrario la palabra tenía impacto si su coeficiente se acerca a -1.

3.5. Evaluación

Así mismo, se realizó una matriz con el peso de valor para cada palabra. De modo que se generé el análisis de sentimientos no solo en la frecuencia de cada una, sino en lo que representa. Esto es dado a través de la medida estadística TF-IDF (term frequency-inverse document frequency). Al estarlo aplicando nosotros desde sklearn, ya trae consigo un banco de palabras con el que estaríamos comparando los clusters obtenidos en el modelo de kmeans.

De tal forma, que se obtendrían la matriz de confusión que se observa en la figura 1 y derivado de esta los resultados de 2. Del cual podemos describir que nuestro modelo acertó en el 84.18 % de las clasificaciones. Sin embargo, para la precisión solo es capaz 13.64 % de no clasificar una muestra negativa como positiva y la exhaustividad solo clasifico correctamente el 2.49 % de los reales.

Confusion Matrix

	0	1
0	180336	5809
1	28196	722

Figura 1: Matriz de confusión

Scores

	scores
accuracy	0.841884
precision	0.110550
recall	0.024967
f1	0.040735

Figura 2: Métricas

4. RESULTADOS

4.1. Análisis descriptivo

En total el dataset contiene 7 columnas y 21,097 filas. De estas, la única columna con valores nulos es *condition* con 1209 entradas nulas.

Las variables categóricas son:

1. uniqueID
2. drugName
3. condition
4. review
5. date

Las cuantitativas son:

1. rating
2. usefulCount

Estas fueron descritas anteriormente.

Además, pudimos observar como del dataset contaba con el 21.9 % del rating con una calificación baja siendo esta menor cuatro. Para la calificación media se tenía un 17.9 % el siendo este un rango entre los que calificaron entre el cuatro al ocho. Por la parte de la calificación alta se tuvo al 60.1 % de los registros siendo estos igual o mayores que ocho.

A continuación, se muestra un histograma del rating.

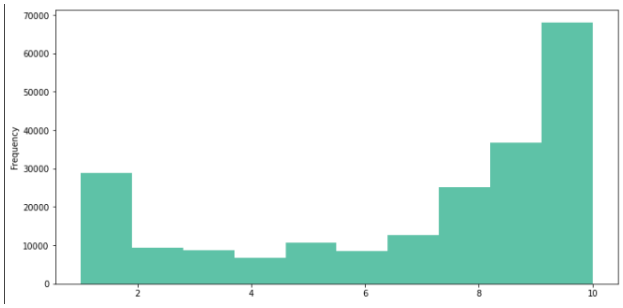

Figura 3: Histograma calificaciones

En cuanto a las reseñas en sí, se buscó encontrar si existía alguna relación en cuanto a su longitud y la reseña dada. No obstante, se encontró lo siguiente.

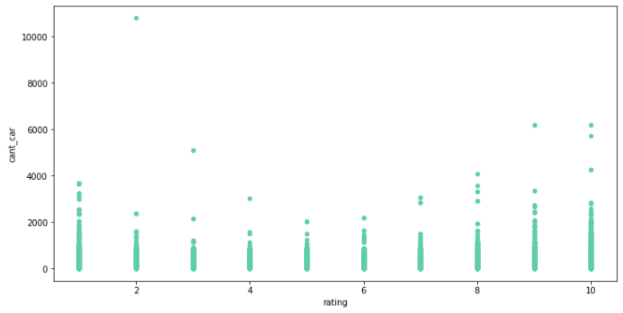

Figura 4: Calificación vs. Longitud

De igual manera, se muestra el histograma de la longitud de las reseñas. Como se observa, la mayoría de las reseñas tenían una longitud de entre aproximadamente 3 y 1,000 caracteres.

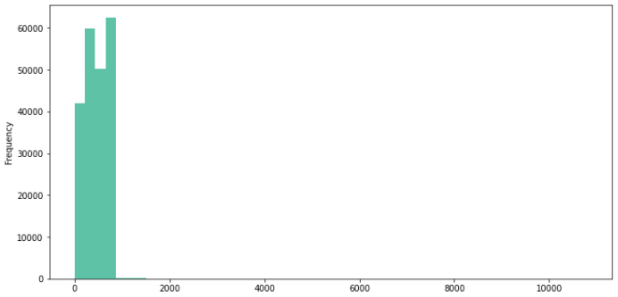

Figura 5: Histograma longitud de reseñas

Asimismo, se muestran los datos estadísticos descriptivos de este mismo apartado.

count	215063.000000
mean	458.620748
std	240.995226
min	3.000000
25%	262.000000
50%	456.000000
75%	690.000000
max	10787.000000

Figura 6: Estadística descriptiva

Además, el modelado de kmeans descrito anteriormente, generé el agrupamiento mostrado en la figura 7, de modo que hay una mínima diferencia entre los comentarios antagónicos, favorable un 0.92 % hacia los comentarios positivos. Sin embargo, el grupo de palabras neutras es notoriamente el mayor siendo así el 82.64 % de todas las analizadas en la reseña.

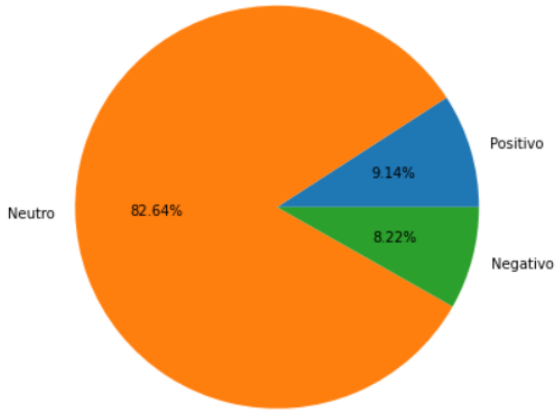

Figura 7: Resultados de agrupamiento de sentimiento

En cuanto a la visualización propia de las palabras más utilizadas en las reseñas, se tienen las siguientes nubes de palabras.
General:

A00517244 Camila Navarro
A01705747 Enrique García
A01197399 Diana Cadena

Figura 8: Nube de palabras más utilizadas en total

Reseñas específicas:

Figura 9: Nube de reseñas más comunes

Reseñas positivas:

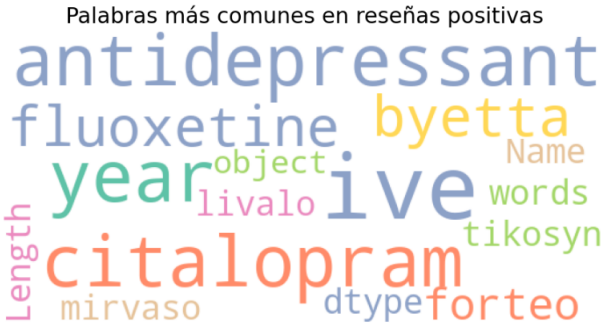

Figura 10: Nube de palabras más utilizadas en reseñas positivas

Reseñas negativas:

Figura 11: Nube de palabras más utilizadas en reseñas negativas

Reseñas neutras:

Figura 12: Nube de palabras más utilizadas en reseñas neutras

Estas últimas tres visualizaciones se obtuvieron después de la configuración y evaluación de los algoritmos utilizados referentes al análisis de sentimientos.

5. CONCLUSIONES Y REFLEXIONES INDIVIDUALES

Con base en los resultados obtenidos, podemos observar que el modelo obtenido nos genera un análisis de sentimientos neutros con mínima tendencia a positivos. Es importante destacar que a pesar de tener un buen resultado para la exactitud, el modelo presenta resultados bajos en las métricas de precisión y exhaustividad. Por lo mismo, sus limitantes sería la predicción correcta de los valores verdaderos. Por lo mismo queda como área de oportunidad poder dedicar más tiempo al procesamiento de los datos, previos a ser aplicados en el modelo.

Con el proyecto, también se puede observar que este es solo el comienzo del análisis de sentimientos. Puesto que la función de describir si las reseñas son positivas o negativas es una de las muchas formas de agrupamiento, se podría aplicar también la relación que tiene con las demás palabras, así como la estructura de las oraciones y el momento en que se menciona cada cosa, funcionan para poder obtener resultados más profundos sobre el sentimiento que quisiera transmitir cada una de las reseñas.

5.1. Navarro, C.

Si bien el aprendizaje no supervisado es un área que tiene mucho por explorar, este fue un buen primer acercamiento. Nos ayudó mucho escoger un tema de interés para nosotros, como lo es el procesamiento de lenguaje. Este conllevó muchísima investigación ya que era un tema completamente nuevo para mí. Desde la comprensión de la metodología hasta la preparación de los datos, la tarea número uno fue investigar y la segunda, experimentar. La limpieza de los datos fue probablemente la parte que más experimentación llevó y en la que más participé, intenté por ejemplo eliminar primero solo errores en el texto, pero no era suficiente para dar buenos resultados. Poco a poco fui eliminando por ejemplo, links, palabras comunes, etc., pero a decir verdad fue un proceso de prueba y error. En cuanto al resto del modelo, el proceso llevó a cabo mucha lectura, y sobre todo documentación acerca de las librerías usadas. Entender el funcionamiento de herramientas como NLTK y Word2Vec, que fueron las dos nuevas librerías que no habíamos usado antes, tuvo un papel esencial en nuestro proceso. En Word2Vec también hubo un proceso de mucha prueba y error. Si bien se podría haber buscado una técnica preestablecida para ajustar estos hiperparámetros, creo que el ajustarlo con prueba y error me ayudó más a comprender su funcionamiento.

5.2. García, E.

Me parece que es un área de mucha aplicación, a través de este trabajo considero que he desarrollado habilidades de agilidad de investigación, por el corto tiempo en que se realizó el trabajo, sin emabrgo la gran cantidad de información que se tenía que entender para poder aplicar un análisis de sentimientos. Me queda claro, que si bien tiene mucha oportunidad de mejora sobre el resultado obtenido del modelo de kmeans, es bastante bueno que se haya experimentado sobre la forma en que se hizo. Ya que se esta aplicando una relación entre los modelos no supervisados para cumplir distintas tareas en vez de generarlos para su comparación, como fue el caso de la vectorización de palabras a través de la red neuronal y la agrupación de estos vectores por medio de kmeans. Me quedo con mucho interés en el área y espero poder seguir aprendiendo más sobre el tema.

5.3. Cadena, D.

Dentro del *machine learning*, existe una gran variedad de algoritmos y maneras de entrenarlos lo que a su vez conlleva a que se puedan aplicar estos mismos a cualquier área, prácticamente. Hasta ahorita, hemos estado más familiarizadxs con el aprendizaje supervisado; hasta cierto punto, se puede decir que hasta nosotrxs mismxs lo tenemos automatizado. Por otro lado, hemos tocado muy superficialmente el aprendizaje no supervisado pero con este proyecto se logró adentrarse un poco más en él. Además, el haber trabajado con el mismo dataset que en el reporte anterior, ayuda a visualizar los distintos enfoques que se le puede dar a los mismos datos. En particular, la parte de análisis de sentimientos; si el aprendizaje no supervisado fuera un charco de agua no explorado, el análisis de sentimientos es el mar. Definitivamente fue algo retador en un inicio el comprender cómo se programa, pues se requieren de distintas librerías y parámetros que ayuden a la normalización del texto para poder configurar los algoritmos correctamente. A pesar de haber sido un primer acercamiento, la métrica de precisión resultó ser bastante buena (> 0,80) aunque las demás quedan pendientes de mejora. En cuanto a mi contribución en el reporte, realicé la descripción de los algoritmos de agrupación y las visualizaciones dentro del código.

REFERENCIAS

[1] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* .O'Reilly Media, Inc.".

[2] Arwinder Dhillon and Ashima Singh. 2019. Machine learning in healthcare data analysis: a survey. *Journal of Biology and Today's World* 8, 6 (2019), 1–10.

[3] Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. 2018. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In *Proceedings of the 2018 International Conference on Digital Health* (Lyon, France) *(DH '18)*. Association for Computing Machinery, New York, NY, USA, 121–125. https://doi.org/10.1145/3194658.3194677

[4] Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. 2018. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In *Proceedings of the 2018 International Conference on Digital Health* (Lyon, France) *(DH '18)*. Association for Computing Machinery, New York, NY, USA, 121–125. https://doi.org/10.1145/3194658.3194677

[5] K. Huster. 2020. UCI ML Drug Review dataset. (2020). https://github.com/katya/Udacity-_capstone/blob/master/UCI_ML_Drug_Review_dataset.ipynb

[6] Sandra Jardim and Carlos Mora. 2022. Customer reviews sentiment-based analysis and clustering for market-oriented tourism services and products development or positioning. *Procedia Computer Science* 196 (2022), 199–206. https://doi.org/10.1016/j.procs.2021.12.006 International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021.

[7] Gang Li and Fei Liu. 2012. Application of a clustering method on sentiment analysis. *Journal of Information Science* 38, 2 (2012), 127–139. https://doi.org/10.1177/0165551511432670

[8] María Isabel Lumbreras Herrera. 2020. Evaluación de análisis de clustering jerárquico en datos moleculares de alta dimensión. (2020).

[9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. https://doi.org/10.48550/ARXIV.1301.3781

[10] Melody Kiang Fangyun Yuan Ming Yang, Jinglu Jiang. 2021. Re-Examining the Impact of Multidimensional Trust on Patients' Online Medical Consultation Service Continuance Decision. *Information systems frontiers : a journal of research and innovation* (2021), 1–25. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7932182/#Bib1title

[11] D Pascual, F Pla, and S Sánchez. 2007. Algoritmos de agrupamiento. *Método Informáticos Avanzados* (2007), 164–174.

[12] Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3, 2 (2011).

[13] Claudia Russo, Hugo Ramón, Nicolás Alonso, Benjamin Cicerchia, Leonardo Esnaola, and Juan Pablo Tessore. 2016. Tratamiento masivo de datos utilizando técnicas de Machine Learning. (2016).

[14] Reinaldo Sánchez Álvarez. 2021. Clasificación no supervisada de imágenes médicas y minería de datos. Algoritmo S3 vs K-medias. *Revista Cubana de Investigaciones Biomédicas* 40 (2021).

[15] José Carlos Sobrino Sande. 2018. Análisis de sentimientos en Twitter. (2018).

[16] Jenna Wiens and Erica S Shenoy. 2018. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clinical Infectious Diseases* 66, 1 (2018), 149–153.