# 

# **Análisis de Comentarios de Pacientes en una Clínica en Maryland utilizando Técnicas de Aprendizaje No Supervisado**

**Mauricio Gonzalez Caro**

**Eva Karina Diaz Gavalo**

**Juan Felipe Padilla Sepúlveda**

**Andrés Eduardo Quiñones Ortiz**

**Universidad de los Andes**

**Maestría en Inteligencia Analítica de Datos**

**Aprendizaje No Supervisado**

**2024**

## **Resumen**

El objetivo de este proyecto es mejorar la evaluación de las encuestas de satisfacción que CCI Health Services envía a sus pacientes después de cada atención. Con 12 ubicaciones en Maryland, Estados Unidos, y aproximadamente 30.000 pacientes por año, este hospital comunitario brinda servicios médicos, dentales, pediátricos y de salud mental. A través de una breve encuesta anónima, el CCI recopila datos demográficos de los pacientes, así como niveles de satisfacción y otros comentarios, que suelen ser extensos y están disponibles en idiomas como el inglés y el español.

Actualmente, las respuestas de las encuestas son recogidas en medios digitales, sin embargo su revisión, análisis y derivación a las áreas especializadas de atención se hace de manera manual, lo que lo convierte en un proceso lento y laborioso. Este enfoque impide la identificación de posibles acciones o resultados a partir de la retroalimentación y limita la capacidad de la clínica para implementar mejoras más profundas y específicas. Si bien es fácil abordar las respuestas estructuradas, las preguntas de respuesta abierta están infrautilizadas a pesar de ser una fuente de retroalimentación valiosa.

Es debido a esto que el proyecto propone el uso de herramientas de procesamiento del lenguaje natural (NLP por sus siglas en inglés) y técnicas de aprendizaje no supervisadas, en particular algoritmos de agrupamiento (Clustering), para clasificar las respuestas de los formularios. Esto identificará temas recurrentes y acercará a las emociones expresadas por los pacientes. Al utilizar estas técnicas, se espera que los datos capturados se transformen en información procesable que ayude a mejorar la práctica clínica, mejorar la experiencia del paciente y fomente las buenas relaciones con los pacientes entre la comunidad médica.

## 

## 

## **Introducción**

En el ámbito de la salud, comprender la satisfacción y las inquietudes de los pacientes es esencial para elevar la calidad del servicio y la atención brindada. Este enfoque permite identificar áreas de oportunidad y mejorar aspectos específicos del cuidado que se ofrece. CCI Health Services, una de las más grandes clínicas comunitarias en el estado de Maryland, Estados Unidos, atiende a aproximadamente 30 mil pacientes al año. Con 12 sedes , ofrece servicios médicos, dentales y de salud mental, apoyados por un equipo de 125 proveedores.

Cada vez que un paciente es atendido, se le envía una breve encuesta diseñada para medir la satisfacción con el servicio recibido. Esta encuesta, que es anónima, permite identificar la ubicación del servicio, el proveedor que lo prestó, y recopila información demográfica básica de los pacientes, como sexo, etnia, raza y lenguaje. Además, incluye preguntas de tipo sí/no para evaluar la satisfacción general, así como una pregunta abierta que invita al paciente a compartir cualquier comentario adicional que considere relevante.

El propósito de esta encuesta es tomar decisiones informadas basadas en la retroalimentación obtenida, con el fin de mejorar la experiencia del paciente y fortalecer la relación entre ellos y la clínica, asegurándoles que sus opiniones son valoradas y consideradas.

Actualmente, el análisis de estas encuestas se enfoca en las respuestas estructuradas, realizar el análisis de los comentarios es un trabajo laborioso, ya que demanda mucho tiempo la lectura de cada comentario para poder determinar cual requiere de una acción específica. Además, no se puede establecer tendencias ni encontrar patrones de comportamiento con este análisis manual. Las respuestas varían en longitud y pueden estar escritas tanto en inglés como en español.

Este proyecto se enfoca en el análisis de estos comentarios de texto libre con el objetivo de clasificarlos en temas y comprender el sentimiento expresado en ellos, utilizando técnicas de aprendizaje no supervisado como un algoritmo de clustering para agrupar los comentarios para identificar patrones y tendencias que permitan responder a la pregunta:  
  
***¿Qué sentimientos y temas recurrentes se expresan en los comentarios de los pacientes, y cómo se pueden utilizar estos para mejorar los servicios de la clínica?***

## 

## 

## **Revisión de la Literatura**

[(PDF) Sentiment Analysis in Healthcare: A Brief Review](https://www.researchgate.net/publication/337653541_Sentiment_Analysis_in_Healthcare_A_Brief_Review): este documento examina cómo las técnicas de análisis de sentimientos se han aplicado en la atención médica para medir la satisfacción de los pacientes y las experiencias de atención recibidas

[Supervised and unsupervised learning models for pharmaceutical drug rating and classification using consumer generated reviews - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S2772442523001557): En este paper se describe cómo técnicas de agrupamiento se pueden usar para detectar patrones y tendencias en grandes volúmenes de datos no estructurados, como los comentarios de los pacientes.

[Patient Satisfaction with Healthcare Services and the Techniques Used for its Assessment: A Systematic Literature Review and a Bibliometric Analysis](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10001171/): Este estudio ofrece una visión integral de los métodos utilizados en la recolección de datos de satisfacción y cómo el análisis de texto no estructurado puede mejorar la evaluación de los servicios de salud.

[Using patient feedback to drive quality improvement in hospitals: A qualitative study](https://www.researchgate.net/publication/346380130_Using_patient_feedback_to_drive_quality_improvement_in_hospitals_a_qualitative_study): este paper analiza cómo los sistemas de retroalimentación automáticos permiten identificar áreas de mejora continua en las clínicas y hospitales

[(PDF) Text Mining In Healthcare](https://www.researchgate.net/publication/338336460_Text_Mining_In_Healthcare): Este estudio presenta técnicas de minería de datos, con un enfoque en la clasificación de minería de texto en salud. Se analizan en detalle los avances y mejoras propuestas en este campo, junto con los desafíos relacionados con el tipo de datos en el sector.

[Development and testing of a text-mining approach to analyse patients’ comments on their experiences of colorectal cancer care | BMJ Quality & Safety](https://qualitysafety.bmj.com/content/25/8/604.short): Este estudio desarrolló y probó un enfoque de minería de texto basado en aprendizaje para facilitar el análisis de las experiencias de los pacientes y crear un modelo explicativo que ilustre su impacto en la HRQoL.

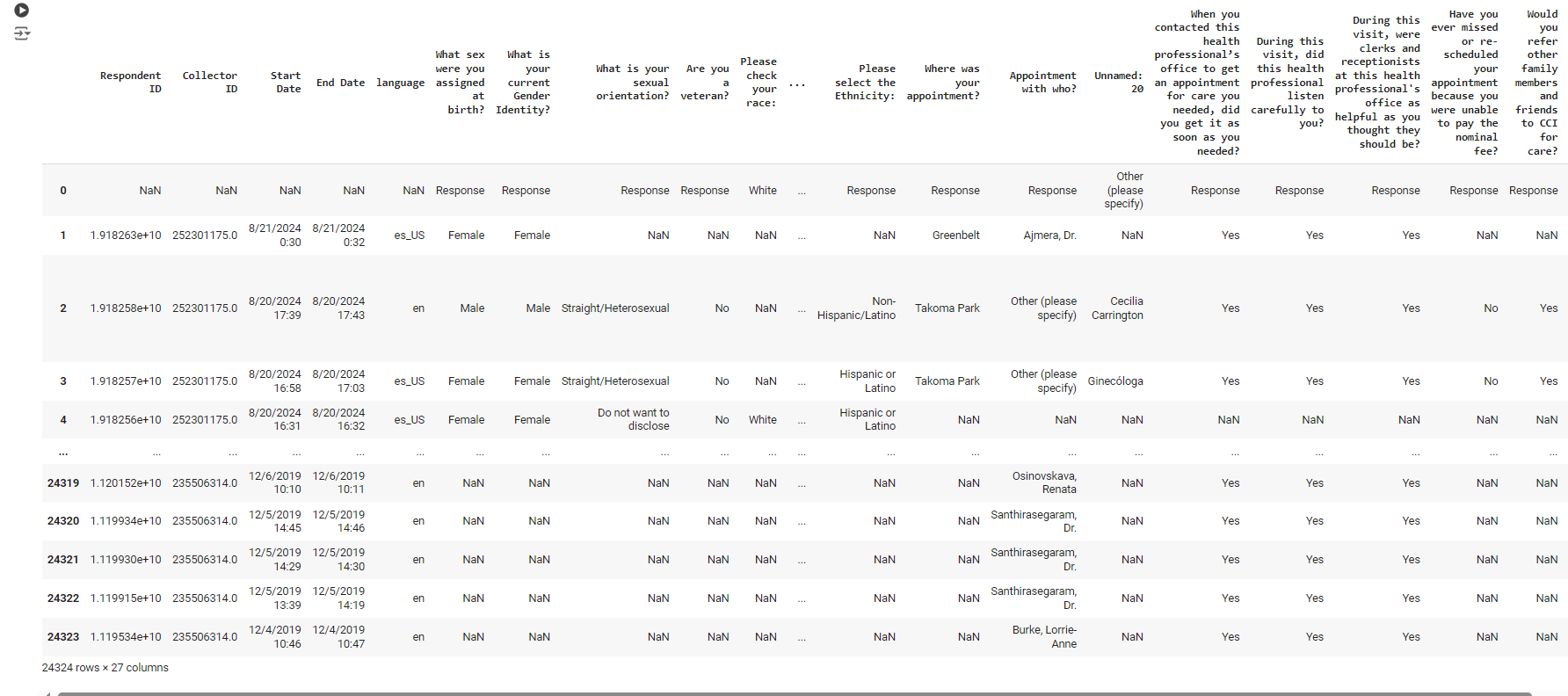
**Descripción de los datos**

La base de datos analizada en este proyecto comprende un total de 24,324 observaciones y 27 variables, las cuales recogen información de las encuestas realizadas a pacientes tras recibir servicios en CCI Health Services desde el 4 de diciembre del 2019 hasta el dia que se descargaron los datos Agosto 21, 2024.

Estas encuestas son enviadas a los pacientes después de cada cita médica y están diseñadas para captar varias respuestas que incluyen tanto información estructurada como comentarios abiertos.

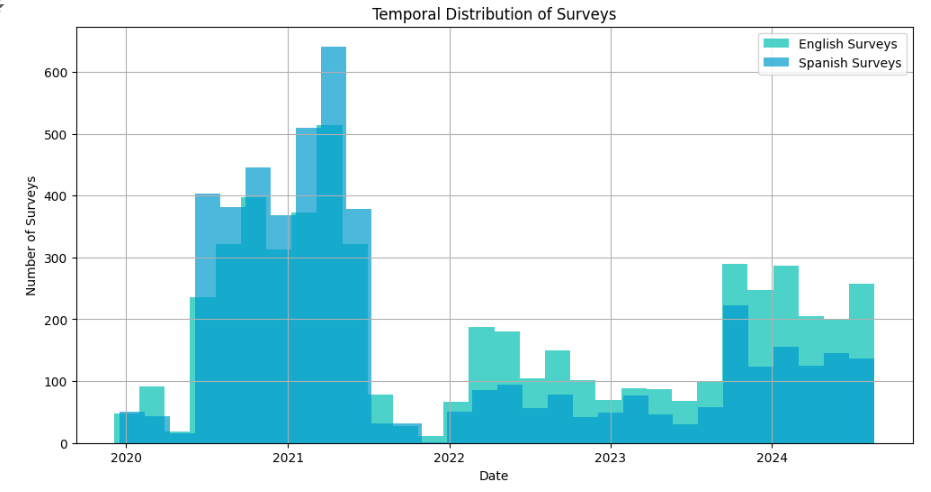
Con el objetivo de centrar el análisis en aspectos directamente relacionados con los comentarios de los pacientes y su contexto, se seleccionaron las siguientes variables: **Respondent ID, End Date, Language, y Do you have any other comments or concerns?.**

El Respondent ID es un identificador único en formato float que permite individualizar cada respuesta. La variable End Date recoge la fecha en la que la encuesta fue completada, y se convertirá a formato de fecha para facilitar su análisis temporal. Language indica el idioma en el que se completó la encuesta, y la variable Do you have any other comments or concerns? contiene los comentarios abiertos proporcionados por los pacientes.

****

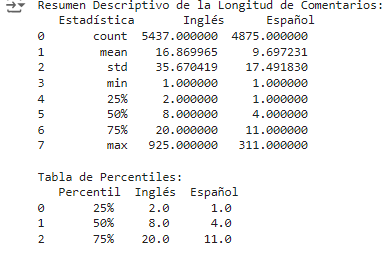
Debido a la naturaleza bilingüe de la base de datos, se consideró esencial dividirla en dos subconjuntos: uno para los comentarios en inglés y otro para los comentarios en español. Esta división es relevante, ya que las diferencias pueden influir en los resultados del análisis.

De las 11,638 filas presentes en la base de datos surveys\_english, se identificaron 6,201 registros con comentarios nulos, lo que representa aproximadamente el 53.27% del total de las observaciones en inglés. En cuanto a la base de datos surveys\_spanish, que contiene 12,680 filas, se detectaron 7,805 registros con comentarios nulos, lo que equivale a un 61.54% de las observaciones en español. Recordemos que esta pregunta es opcional, preguntando a los pacientes si tienen algun otro comentario adicional a las preguntas realizadas.

****

Una vez eliminados los comentarios nulos de las bases de datos en inglés y español, se procedió graficar la distribución temporal de las encuestas, en el que se observa un pico considerable a mediados de 2020 y 2021. Este aumento podría estar relacionado a la pandemia de COVID.

Aunque durante la pandemia las encuestas en español fueron significativamente numerosas, en la mayoría de los casos las encuestas en inglés parecen ser más frecuentes. En 2024, la cantidad de encuestas en español y en inglés se muestra más equilibrada, aunque las encuestas en inglés siguen dominando en cantidad.



Se procedió a analizar la longitud promedio de los comentarios en la base de datos en inglés, encontrando que el promedio de palabras por comentario es de 16.87, con una desviación estándar de 35.67. Esto indica una considerable variabilidad en la cantidad de texto proporcionado por los pacientes. El comentario más corto consta de una sola palabra, mientras que el más largo alcanza las 925 palabras. Los percentiles muestran que el 25% de los comentarios contienen 2 palabras o menos, el 50% tienen 8 palabras o menos, y el 75% tienen 20 palabras o menos. A pesar de la presencia de algunos comentarios extensos, la mayoría son relativamente breves, lo que sugiere una distribución desequilibrada en la longitud de los comentarios. Este desequilibrio podría introducir sesgos en el análisis, ya que los comentarios extremadamente cortos, como aquellos de 2 palabras, podrían estar duplicados o no aportar valor significativo al análisis. Por tanto, en la siguiente etapa se aplicarán técnicas de preprocesamiento de lenguaje natural para limpiar y normalizar estos datos.

Por otro lado, en la base de datos en español, se analizaron 4,875 comentarios. La longitud promedio es de 9.70 palabras, con una desviación estándar de 17.49 palabras. Al igual que en los comentarios en inglés, existe una amplia variabilidad, pero en menor grado. El comentario más largo tiene 311 palabras. Los percentiles indican que el 25% de los comentarios tienen 1 palabra o menos, el 50% tienen 4 palabras o menos, y el 75% tienen 11 palabras o menos. Estos resultados muestran que los comentarios en español tienden a ser más cortos que los comentarios en inglés..

**Metodología**

Los datos utilizados en este proyecto consisten en comentarios escritos por pacientes, recopilados a través de encuestas de satisfacción. Después de realizar un análisis exploratorio de los datos, se describen las etapas metodológicas empleadas para clasificar estos comentarios en diferentes tópicos y sentimientos utilizando técnicas de procesamiento de lenguaje natural y Aprendizaje No Supervisado.



1. **Preprocesamiento de la Base de Datos**

El preprocesamiento es esencial para preparar los datos textuales para el análisis, asegurando que sean consistentes y adecuados para las técnicas de aprendizaje no supervisado que se aplicará posteriormente.

**Conversión a minúsculas:** Evitando que palabras idénticas sean tratadas como distintas.

**Eliminación de Signos de Puntuación y Caracteres Especiales:** Eliminar ruido en los datos.

**Remoción de Stopwords:**Centrar el análisis en términos más relevantes.

**Eliminación de Números:** Los números suelen no aportar valor en el análisis de sentimientos y tópicos.

**Comentarios Duplicados:**Detectar y eliminar duplicados para evitar sesgos en el análisis.

**Lematización:**Reducimos las palabras a su forma base.

**Análisis de N-gramas:**Descubriremos frases y combinaciones de palabras que aparecen con frecuencia, proporcionando más contexto que las palabras individuales.

**2.Vectorización**

Una vez que el texto ha sido limpiado y preprocesado, es necesario convertirlo en una representación numérica que pueda ser procesada por algoritmos de aprendizaje automático.Seutilizará el vectorizador TF-IDF de scikit-learn para transformar el corpus preprocesado.

**3. Reducción de la Dimensión**

En esta etapa procederemos a la reducción de dimensionalidad para ayudar a simplificar la base de datos de comentarios manteniendo la mayor cantidad de información posible. Se hara prueba con dos técnicas:

**Análisis de Componentes Principales (PCA):** Identificamos los componentes principales en la que los datos varían más y proyectaremos los datos en un espacio de menor dimensión. Normalizamos la Base de datos y seleccionamos k de manera que se explique al menos un 95% de la varianza original.

**Descomposición en Valores Singulares (SVD):** Utilizar la función TruncatedSVD de scikit-learn para calcular sólo los primeros k componentes necesarios. Analizaremos la varianza explicada para determinar el número óptimo de componentes (matrices esparzas libro - leí algo sobre esto pero quiero buscar la justificación para ver si elegimos uno o el otro o probamos ambos no se)

**4. Clustering**

Una vez que los datos han sido vectorizados y reducidos en su dimensionalidad, procederemos a aplicar diferentes técnicas de clustering con el fin de agrupar los comentarios en clusters que representan tópicos o sentimientos similares. Los algoritmos de clustering que se utilizarán son:

**K-Means:** Este algoritmo se seleccionará inicialmente debido a su simplicidad y eficacia en la partición de datos en clusters. Se determinará el número óptimo de clusters utilizando el método del codo, el cual nos permitirá identificar el punto donde la mejora en la varianza explicada por cada cluster adicional comienza a disminuir significativamente. Cada cluster resultante será analizado para entender su composición y relevancia en la clasificación de los comentarios.

**K-Medoids:** Se usará este algoritmo para verificar la estabilidad de los clusters obtenidos y compararlos con los resultados de K-Means. Se evaluará si este método ofrece clusters que sean más coherentes y útiles para la interpretación de los comentarios.

**Clustering Jerárquico:** Este enfoque no requiere un número predefinido de clusters y permite visualizar la relación entre los datos a través de un dendrograma. Se utilizará clustering jerárquico para explorar posibles relaciones jerárquicas entre los comentarios, que podrían revelar sub-temas dentro de los tópicos principales identificados con otros métodos.

**DBSCAN:** Será implementado para evaluar si existen patrones en los datos que no pueden ser capturados por métodos basados en la distancia.

Finalmente, los resultados de estos métodos serán analizados para determinar la calidad y utilidad de los clusters obtenidos.

**Para Análisis de Sentimientos**: Usaremos k = 3 para poder dividir en 3 sentimientos.

**Para Agrupación de Temas**: Conseguiremos el K óptimo para agrupar nuestros comentarios.

**Bibliografía**

Clinical Natural Language Processing in languages other than English: opportunities and challenges: <https://link.springer.com/article/10.1186/s13326-018-0179-8>

# 