|  |
| --- |
|  |
| Informe Proyecto 1, Entrega 1 Inteligencia de Negocios |
|  |

|  |
| --- |
| Maria Fernanda De la Hoz (202214512), Silvana Sandoval Galvis 202123682, Gabriela Soler 202123744  7 de sept. de 24 |

Contenido

[**Entendimiento del negocio y enfoque crítico** 2](#_Toc176599550)

[**Entendimiento y preparación de los datos** 2](#_Toc176599551)

[**Modelado y evaluación** 4](#_Toc176599552)

[**Resultados** 5](#_Toc176599553)

[**Mapa de los actores** 9](#_Toc176599554)

[**Trabajo en equipo** 9](#_Toc176599555)

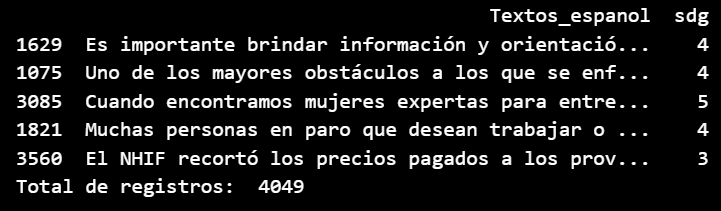
[**Referencias** 10](#_Toc176599556)

# **Entendimiento del negocio y enfoque crítico**

|  |  |
| --- | --- |
| **Oportunidad/Problema del negocio** | Existe la necesidad de un sistema que identifique de manera eficiente los problemas reportados por los ciudadanos y evalúe las soluciones actuales, conectando esta información con los Objetivos de Desarrollo Sostenible (ODS). |
| **Objetivos y criterios de éxito desde el punto de vista**  **del negocio.** | Desarrollar un modelo analítico automatizado para extraer y analizar la información más relevante de opiniones ciudadanas en lenguaje natural, relacionándola con los ODS 3, 4 y 5.  El éxito se medirá por la capacidad del modelo para extraer información clave con precisión, la reducción en costos y tiempo de análisis, y la facilidad de integración y uso de la solución. |
| **Organización y rol dentro de ella que se beneficia con**  **la oportunidad definida** | La organización que se beneficia con la oportunidad definida es el Fondo de Poblaciones de las Naciones Unidas (UNFPA). Dentro de la organización, el rol que se beneficia es el de analista de datos y formuladores de políticas. |
| **Impacto que puede tener en Colombia este**  **proyecto.** | El proyecto podría tener un impacto significativo en Colombia al mejorar la eficiencia en la identificación y evaluación de problemas sociales alineando las acciones con los ODS. Esto apoyaría la toma de decisiones basadas en datos en políticas públicas y puede contribuir a mejorar la calidad de vida de los ciudadanos al abordar de manera más efectiva las necesidades sociales prioritarias. |
| **Enfoque analítico. Descripción de la categoría de**  **análisis (descriptivo, predictivo, etc.), tipo y tarea de**  **aprendizaje e incluya las técnicas y algoritmos que** **propone utilizar.** | La categoría del enfoque analítico será descriptivo y predictivo ya que en este ejercicio se busca predecir el ODS al que pertenece un comentario y, al mismo tiempo, se busca dar un entendimiento de este proceso.  **Tarea**: clasificación de textos para relacionar las opiniones ciudadanas con los ODS específicos. |

# **Entendimiento y preparación de los datos**

Para empezar este proceso de perfilamiento, es importante empezar por el entendimiento de los datos. Para lograr esto, sacamos un sample de 5 registros y obtuvimos el número total de registros:



De esto logramos concluir que los datos con los que vamos a trabajar tienen 4049 registros y 2 columnas. Esto, en términos del problema, se traduce a 4049 opiniones de los ciudadanos y 2 variables con las que vamos a trabajar: la primera es Textos\_espanol que son los comentarios de los ciudadanos, la cual sería nuestra variable independiente X; la segunda, por su parte, sería sgd (o el ODS al que pertenece el comentario) la cual es nuestra variable objetivo. Además, hicimos un test de unicidad, es decir, comprobamos que no hubiera duplicados, lo cual fue el caso. Por otro lado, hicimos test de completitud para asegurarnos de que no hubiera valores nulos. Por último, antes de pasar al preprocesamiento, comprobamos el tipo de datos que contenía este conjunto y estos fueron:

Data type test image

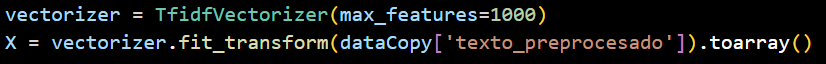
Lo anterior nos permite inferir que estos datos son de muy buena calidad ya que no existen duplicados, ni valores nulos, lo cual permite que su procesamiento sea más rápido y oportuno ya que no hay que eliminar datos. Por otro lado, esto será beneficioso a la hora de crear los modelos porque al tener una mayor cantidad de datos, permitimos que haya una menor probabilidad de sesgos.

**Preprocesamiento de los datos**

Para empezar el preprocesamiento, hicimos uso de la librería spaCy que permite dividir las cadenas de texto en listas donde cada palabra corresponde a un índice de esta. Después de esto, creamos tokens a partir de las palabras para asegurarnos de tener cada palabra de forma individual. Por último, hicimos la lematización para obtener la forma base de las palabras. Esto nos permitió llevar a cabo la remoción de las *stopwords*, es decir, las palabras que, sin contexto no aportan ninguna información. Por último, filtramos todos los caracteres especiales para asegurarnos de que estábamos trabajando con caracteres alfabéticos. Después, obtuvimos un listado de todas las entidades existentes junto al su *label*. Además, clasificamos las palabras por el tipo de sintagma (nominal, verbal, adjetival, etc.) al que pertenecen. Es decir, clasificamos las palabras para saber si son adjetivos, adverbios, pronombres, sustantivos, verbos, entre otros. Por último, volvimos a articular el texto procesado a partir de los espacios para poder guardar el resultado en una tabla de Excel. Lo anterior nos permitirá crear de forma más sencilla los vectores que se usarán para entrenar los modelos.

**Transformación de los datos para el modelado**

Para poder usar los datos para el modelado de los algoritmos, recurrimos a 3 técnicas diferentes. La primera, y la más importante, fue convertir cada línea de texto en un vector TFIDF. Esto permite que los algoritmos puedan que solo toman atributos numéricos, puedan aproximar de forma más exitosa la variable objetivo. Esto lo hicimos por medio de la siguiente función:



Luego, proseguimos a hacer un count vectorizer el cual nos permite pasar los textos a una matriz que contabiliza los tokens usados en los mismos. Esto lo realizamos con la siguiente función:

Count Vectorize

Por último, creamos datos dummies a partir del texto preprocesado. Esto lo hicimos para tener una tercera opción para entrenar a los modelos y lo hicimos con la siguiente función:

Dummy

# **Modelado y evaluación**

El primer algoritmo que elegimos para la construcción del modelo fue árbol de decisión. Este es un algoritmo que sirve para clasificación que consiste en crear nodos (que serían los test de atributo para iniciar la clasificación) y hojas que representan los resultados la final de toda la clasificación (IBM, s.f.). Este se escogió debido a que nos permitiría crear un modelo que fácilmente clasificara los comentarios una vez estos pasaran por una transformación TF-IDF. El segundo algoritmo que elegimos fue K Vecinos más Cercanos. Esto lo hicimos ya que este es un algoritmo de clasificación que se encarga de asignar la clase de nuevos datos dependiendo de la proximidad que tienen con los ya existentes (IBM, s.f.), lo cual nos permitía tener otra aproximación para resolver el problema y así poder evaluar el mejor modelo. Por último, elegimos Bayes Ingenuo Multinominal porque este es un algoritmo que sirve específicamente para clasificación de texto, específicamente, lengua natural (IBM, s.f.). Este fue sumamente útil para este proyecto ya que los comentarios son lenguaje natural y, después de sufrir sus determinadas transformaciones, permitió tener una muy buena aproximación para escoger el modelo adecuado.

Ahora bien, para escoger el modelo que se aproximaba mejor, nos enfocamos en las siguientes estadísticas:

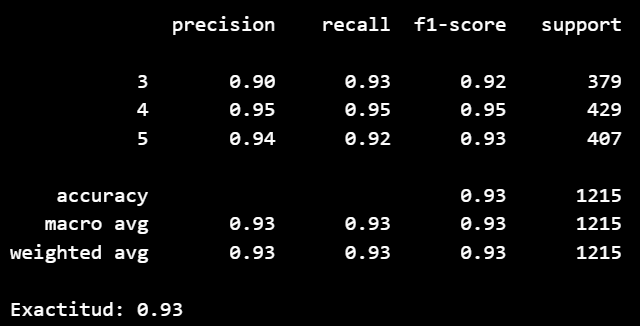


Ilustración 1. Estadísticas árbol de decisión

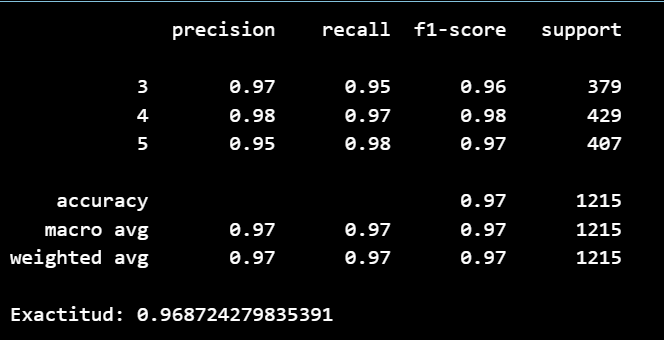


Ilustración 2. Estadísticas Bayes Ingenuo

Pantalla de celular con aplicaciones

Descripción generada automáticamente con confianza media

Ilustración 3. Estadísticas K Vecinos

A partir de estas métricas, decidimos que el modelo que debíamos implementar para el trabajo de clasificación de los comentarios que aún no tenían un ODS asociado era Bayes Ingenuo Multinominal. Esto lo decidimos ya que es el modelo con mayor exactitud con un porcentaje de aproximadamente el 97%. Además, de los tres, es el que en general tienen el mejor recall en todas las clases, es decir, el que mejor clasifica los ODS sobre la clase. También sobre sale en su precisión en todas las clases donde la más baja es el ODS 5 con 95%, lo cual significa que solo un 5% de los datos clasificados como ODS 5 son de otras categorías. Finalmente, los F-Scores de este también son los mejores de los tres modelos con el menor siendo de 0.96, el cual está asociado con el ODS 3 que es el tiende a identificar como otros en el 5% de las veces.

# **Resultados**

**Resultados obtenidos**

Los resultados de los modelos son los siguientes:

**Gráfico, Gráfico de rectángulos

Descripción generada automáticamente**

Ilustración 4. Matriz de Confusión Árbol de decisión

Como se puede observar, la clase que menos precisión tiene es el ODS 3, lo cual implica que esta es la clase cuyos datos se tiende a catalogar erróneamente. Sin embargo, el recall de esta clase es el más alto de todas, lo cual implica que lo que se categoriza dentro de la clase tiene una mayor precisión. Es decir, que aquello que el algoritmo, dentro del ODS 3, categoriza como ODS 3 y el resto lo hace con mayor precisión es de 0.97, lo cual implica que en el 97% de los casos los comentarios que son de ODS . Por otra parte, para hacer referencia al F-Score, el de las 3 clases es bastante alto, lo cual implica que tanto la precisión y el recall son buenos como se mostró en las métricas anteriores. Esto aporta a la consecución de los objetivos del negocio de manera que permite tener un entendimiento de que el peor modelo que ellos podrían usar igualmente les permite tener una muy buena exactitud al momento de predecir los ODS de los comentarios.

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Ilustración 5. Matriz de Confusión Bayes Ingenuo

Como se puede observar, la clase que menos precisión tiene es el ODS 3, lo cual implica que esta es la clase cuyos datos se tiende a catalogar erróneamente. De igual manera, el recall de esta clase es el más bajo de todos, lo cual puede explicar porque ocurre dicha tendencia. Por el contrario, su precisión es del 97%, lo cual implica que en el 97% de las veces que categoriza un comentario como ODS 3, lo es. Por otra parte, para hacer referencia al F-Score, el de las 3 clases es bastante alto, lo cual implica que tanto la precisión y el recall son buenos como se mostró en las métricas anteriores. Por último, este modelo tiene una exactitud del 97%, la mayor de todos. Esto aporta a la consecución de los objetivos del negocio de manera que permite predecir con bastante exactitud nuevos comentarios en sus ODS correspondientes para entender los temas de discusión más frecuentes de los ciudadanos.

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

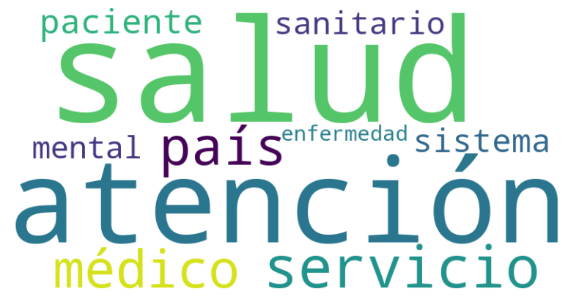
Ilustración 6. Matriz de Confusión K Vecinos

Como se puede observar, la clase que menos precisión tiene es el ODS 5, lo cual implica que esta es la clase cuyos datos se tiende a catalogar erróneamente. De igual manera, el recall de esta clase es el más bajo de todos con un 0.94, lo cual puede explicar porque ocurre dicha tendencia. Por el contrario, su precisión es del 97%, lo cual implica que en el 97% de las veces que categoriza un comentario como ODS 5, lo es. Por otra parte, para hacer referencia al F-Score, el de las 3 clases es bastante alto, lo cual implica que tanto la precisión y el recall son buenos como se mostró en las métricas anteriores. Por último, este modelo tiene una exactitud del 96%, la segunda mejor de todas. Esto aporta a la consecución de los objetivos del negocio de manera que permite tener un modelo de respaldo que permite clasificar con cierta exactitud los comentarios en caso de que el otro modelo falle o no esté disponible.

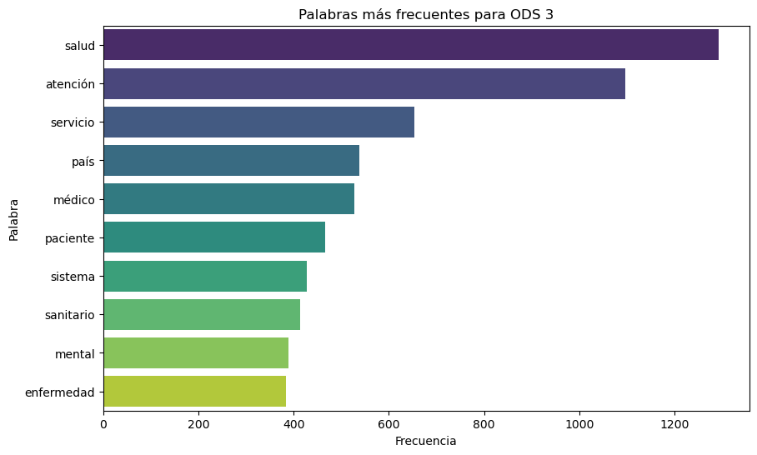
Todo lo anterior le permite tener un mejor entendimiento de las distribuciones de los datos del negocio y como, para este, hay una abundancia de modelos que podrían servir para la tarea de clasificación, unos con mejor alcance que otros.

**Frecuencia de palabras por ODS**

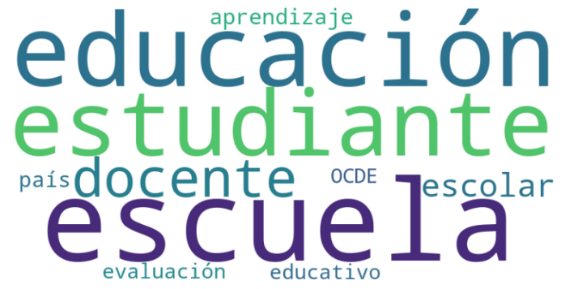
**ODS 3**



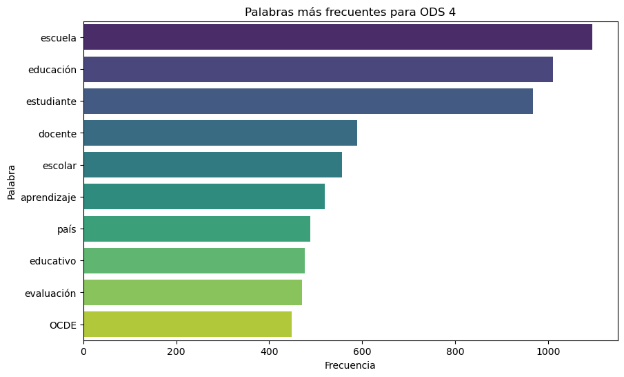
La imagen anterior nos permite apreciar que las palabras más usadas en los comentarios sobre el ODS tres giran alrededor del campo semántico de la salud, es decir, las palabras que con más frecuencia se usan giran en torno al área de la salud. Esto corresponde perfectamente con este ODS ya que se enfoca en la salud y el bienestar de la población (Naciones Unidas, s.f.). De igual manera, esto se puede ver claramente en el siguiente gráfico de barras donde la palabra "salud" corresponde a la barra más larga:



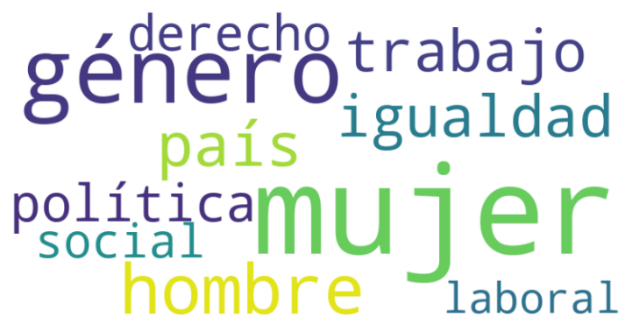
**ODS 4**



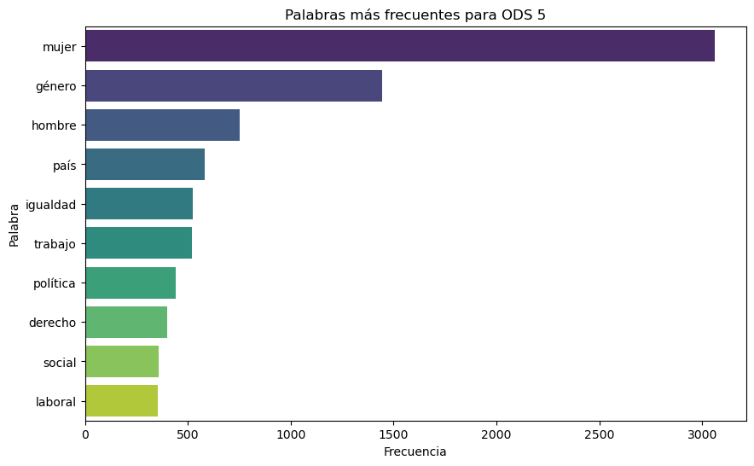
La imagen anterior nos permite apreciar que las palabras más usadas en los comentarios sobre el ODS cuatro giran alrededor del campo semántico de la educación, es decir, las palabras que con más frecuencia se usan giran en torno al sector educativo. Esto corresponde perfectamente con este ODS ya que se enfoca en que la población tenga acceso a educación de calidad (Naciones Unidas, s.f.). De igual manera, esto se puede ver claramente en el siguiente gráfico de barras donde las palabras "escuela" y "educación" corresponden a las barras más largas:



**ODS 5**



La imagen anterior nos permite apreciar que las palabras más usadas en los comentarios sobre el ODS cinco giran alrededor del campo semántico de la igualdad de género, es decir, las palabras que con más frecuencia se usan giran en torno a la igualdad de género. Esto corresponde perfectamente con este ODS ya que se enfoca en conseguir la igualdad de género entre los diferentes sectores que conforman la población (Naciones Unidas, s.f.). De igual manera, esto se puede ver claramente en el siguiente gráfico de barras donde las palabras "mujer" y "género" corresponden a las barras más largas:



# **Mapa de los actores**

|  |  |  |  |
| --- | --- | --- | --- |
| **Rol dentro de la empresa** | **Tipo de actor** | **Beneficio** | **Riesgo** |
| Gobierno local | Financiador-Cliente | Priorización eficiente de recursos para proyectos alineados con los ODS según las necesidades ciudadanas. | Malinterpretación de resultados, lo que podría llevar a decisiones equivocadas. |
| Analista de datos | Usuario | |  | | --- | |  | | Automatización del análisis de grandes volúmenes de datos, reduciendo tiempo y recursos. | | | Sobrecarga de datos que pueden no estar completamente validados o filtrados. |
| Ciudadanos | Beneficiado | Mayor influencia en políticas públicas relacionadas con sus preocupaciones expresadas. | Si las opiniones son mal clasificadas, puede generar frustración al sentir que sus necesidades no fueron atendidas. |
| ONG de desarrollo sostenible | Proveedor | Acceso a datos clasificados para impulsar sus proyectos y enfoques en los ODS más relevantes. | Riesgo de falta de neutralidad en la clasificación, afectando la veracidad de los datos. |

# **Trabajo en equipo**

1. María Fernanda: Ella se encargó de crear el repositorio en github, al igual que de asignar las funciones y verificar que aquellos que el resto del grupo añadiera al repositorio fuera lo indicada. Ella hizo el perfilamiento de los datos, el algoritmo K vecinos y las nubes de palabras. Por último, ella hizo el entendimiento y enfoque crítico del negocio. **Roles:** *Líder del proyecto y datos.*
2. Gabriela: Ella se encargo de hacer las transformaciones de los datos para que pudieran ser usadas para el entrenamiento de los modelos. Además, hizo el algoritmo Bayes Ingenuo y la presentación que será usada para el video. Por último, ella hizo el código que clasificó los comentarios faltantes a partir de su algoritmo. Por último, ella hizo el mapa de actores. **Roles:** *Líder analítica.*
3. Silvana: Ella se encargó de hacer el pipeline que se usará más adelante. Además, hizo el árbol de decisión y se encargó de darle formato a todo el notebook para que fuera de fácil lectura. También, hizo la wiki. Por último, redactó *entendimiento y preparación de datos*, *modelado y evaluación,* y *Resultados* del presente documento. **Roles:** *Líder de negocio.*

# **Referencias**

IBM. (s.f.). *¿Qué son los clasificadores Naïve Bayes?* Obtenido de IBM: https://www.ibm.com/mx-es/topics/naive-bayes

IBM. (s.f.). *Algoritmo de k vecinos más cercanos*. Obtenido de IBM: https://www.ibm.com/es-es/topics/knn

IBM. (s.f.). *Árboles de decisión*. Obtenido de IBM: https://www.ibm.com/es-es/topics/decision-trees

Naciones Unidas. (s.f.). *¿Sabes cuáles son los 17 objetivos de desarrollo sostenible?* Obtenido de Objetivos de desarrollo sostenible: https://www.un.org/sustainabledevelopment/es/2018/08/sabes-cuales-son-los-17-objetivos-de-desarrollo-sostenible/