

Proyecto 1, Etapa 1: Analítica de textos

Federico Melo Barrero, 202021525, f.melo. Sección 1.

Shadith Pérez Rivera, 202014687, s.perezr. Sección 2.

El presente documento describe la manera en la cual se aplicó una metodología de analítica de textos en el contexto del proyecto y expone los resultados obtenidos en cada una de las fases.

Contenido

Entendimiento del negocio y enfoque analítico.....	1
Entendimiento y preparación de los datos	3
Modelado y evaluación	4
Resultados.....	6
Mapa de actores	8
Trabajo en equipo	9
Lista de Referencias	11



Entendimiento del negocio y enfoque analítico

El Fondo de Poblaciones de las Naciones Unidas (UNFPA) es organización internacional comprometida con el cumplimiento de los Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas. Un **ODS** se define como una meta global, establecida para mitigar o solucionar distintos desafíos que actualmente enfrenta el mundo y que afectan la vida de las personas. “Los líderes mundiales adoptaron un conjunto de objetivos globales para erradicar la pobreza, proteger el planeta y asegurar la prosperidad [...] que deben alcanzarse en los próximos 15 años.” [1]

El **objetivo** de este proyecto es diseñar un modelo de aprendizaje automático que, dado un apartado de texto arbitrario, permita al UNFPA clasificarlo como relacionado a alguno de los siguientes tres objetivos de desarrollo sostenible:

6. “Garantizar la disponibilidad de agua y su gestión sostenible y el saneamiento para todos” [2]
7. “Garantizar el acceso a una energía asequible, segura, sostenible y moderna”. [3]
16. “Promover sociedades justas, pacíficas e inclusivas”. [4]



Abordar estos ODS en Colombia tiene un impacto significativo en las condiciones de vida de los colombianos, garantizando agua potable para los 3.2 millones de colombianos que no tienen acceso a este recurso [5], impulsando el desarrollo económico a través de energía

 <div> <div>Universidad de los Andes</div> <div>Colombia</div> </div> <div> <div>Acreditación</div> <div>institución de alta calidad</div> <div>10 años</div> <div>Ministerio de Educación</div> <div>Resolución 1812</div> <div>9 de mayo de 2015</div> </div>	<div>Ingeniería de Sistemas y Computación</div> <div>Pregrado</div> <div>ISIS-3301 – Inteligencia de Negocios</div> <div>Primer Proyecto Semestre: 2023-20</div>	
--	--	---

sostenible e impulsando la paz en un país que cerró el 2022 con 26,1 asesinatos por cada 100,000 habitantes [6].

Se establecen dos **criterios de éxito** para este proyecto: 1. Que el modelo establecido sea capaz de recibir un apartado de texto y clasificarlo como uno de los 3 ODS mencionados. 2. Que el modelo exhiba una probabilidad mayor al 90% de que la clasificación asignada sea correcta.

Oportunidad/problema Negocio	La UNFPA tiene la necesidad de analizar información relacionada con opiniones que representan la voz de los habitantes locales sobre problemáticas de su entorno particular, clasificándola con respecto a su relación con alguno de los ODS de la ONU.
Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático) e incluya las técnicas y algoritmos que propone utilizar.	<p>En el contexto del aprendizaje automático, el proyecto requiere que dado un apartado de texto arbitrario este se clasifique como relacionado con alguno de los 3 ODS mencionados anteriormente. Para esto se hace uso de los siguientes 3 métodos:</p> <ol style="list-style-type: none"> 1. Red Neuronal 2. Regresión Logística 3. Naive Bayes <p>Se eligen estas técnicas porque son complementarias entre sí. A continuación, se justifica la elección de cada uno de los modelos (también se justifican en la sección de Modelado y Evaluación):</p> <ol style="list-style-type: none"> 1. La red neuronal es intrincada y particularmente útil para detectar relaciones complejas y sutiles entre los datos. 2. La regresión logística es especialmente útil en caso de que haya relaciones lineales entre los datos (lo cual no se sabe de antemano) e intenta maximiza la probabilidad de que la clasificación sea correcta. Además, provee dicha probabilidad. 3. Naive Bayes presume de antemano independencia condicional entre los textos (lo cual, de nuevo, no se sabe de antemano si necesariamente es el caso) pero incluso con esa suposición ingenua (de ahí el <i>naive</i>) se caracteriza por arrojar resultados muy buenos, sobretodo en modelos de procesamiento de texto en lenguaje natural, lo cual es sumamente apropiado para este contexto de negocio.



 Universidad de los Andes Colombia	Ingeniería de Sistemas y Computación Pregrado ISIS-3301 – Inteligencia de Negocios Primer Proyecto Semestre: 2023-20	
Organización y rol dentro de ella que se beneficia con la oportunidad definida	A partir de la oportunidad definida existen beneficios en términos de negocio para la UNFPA y para la ONU. Automatizar la clasificación de los textos permite ahorrar tiempo y capital humano en saber qué información puede utilizarse para cuáles causas, y permite proveer a cada colaborador información relevante con su causa. Por ejemplo, a aquellos enfocados en colaborar con el acceso a agua potable, rápidamente se les brindan los insumos, descripciones y opiniones necesarias para su trabajo de forma automática, que son los clasificados como relacionados con el ODS 6, lo cual los hará más efectivos en su contribución a cumplir el ODS. Nótese que, a causa del contexto en el que se enmarcan los ODS, realmente su cumplimiento beneficia a todos los seres humanos.	
Contacto con experto externo al proyecto	Se tuvo contacto con la estudiante de estadística Melany Buenaños Sanmiguel, m.buenanos@uniandes.edu.co , con quien se tendrán reuniones constantes de manera presencial a partir de la primera reunión el jueves 19 de octubre a las 16:00 horas, en la cual se le informará del proceso del proyecto y de los resultados de esta primera parte, y se tendrá una comunicación fluida usando como canal el grupo de WhatsApp.	

Entendimiento y preparación de los datos

Se realizó el entendimiento y preparación de los datos. En el perfilamiento de los datos, se identificó que se trata un archivo con 3000 filas y 2 columnas. La primera columna corresponde a descripciones de problemáticas relacionadas con los ODS de la ONU en algún contexto particular, mientras que la segunda corresponde a la clasificación correcta de las problemáticas según su relación con el ODS 6, 7 o 16 determinada por un experto.

Se evidenció que los 3000 registros son todos únicos y que no cuentan con datos faltantes. Sumado a eso, cada registro está clasificado en exactamente una de las tres categorías y hay 1000 registros para cada categoría, por lo que cada una está representada uniformemente. Respecto a estadísticas sobre los textos, se evidenció que cuentan con una longitud promedio de 770 caracteres, con un mínimo de 143 y un máximo de 1616; también que las palabras más comunes en los textos son palabras de parada (por lo cual más adelante serán removidas), siendo la más común en general la palabra “de”; adicionalmente, la palabra más larga de cada texto tiene un promedio de 16 caracteres; y, por último, la palabra más corta de cada texto tiene en promedio 1 carácter y es también una palabra de parada.

Con base en el perfilamiento se analizó la calidad de los datos teniendo en cuenta las dimensiones de unicidad, consistencia, validez, y completitud. En cuanto a unicidad, debido a que no se encontraron registros duplicados, no fue necesario eliminar ningún registro. Adicionalmente, se encontró que los datos están completos: ninguna fila tenía un registro

 <p>Universidad de los Andes Colombia</p> <p>Acreditación institución de alta calidad 10 años Ministerio de Educación Resolución 2812 de 2015</p>	<p>Ingeniería de Sistemas y Computación Pregrado</p> <p>ISIS-3301 – Inteligencia de Negocios Primer Proyecto Semestre: 2023-20</p>	 <p>ABET Engineering Accreditation Commission</p>
--	--	--



faltante y a cada uno de los registros de texto le correspondía exactamente una clasificación como ODS. Más aún, se determinó que los datos son válidos debido a que todos los textos son cadenas de caracteres de más de 10 caracteres y a qué todas las clasificaciones son un número entero en el conjunto {6, 7, 16}. Por último, respecto a la validez, se identificó que alrededor del 0.1% de los datos estaban en un idioma foráneo. Se tomó la decisión de eliminar dichos datos, debido a que eran muy pocos de la muestra total y no se desea introducir un factor de propagación del error intentando traducir dichos términos.

Con respecto a las transformaciones realizadas a los datos, se tomaron las siguientes acciones:

1. **Preprocesamiento (y remoción de las palabras de parada).** Esta etapa consistió de remover los caracteres que no hacen parte de la codificación ASCII, remover todos los que no son alfabéticos, convertir todos los caracteres de mayúscula a minúscula, reemplazar los caracteres numéricos por su equivalente en palabras (“2” pasa a “dos”) y remover las palabras de parada (que en español suelen ser artículos, preposiciones, conjunciones, determinantes, entre otros).
2. **Tokenización.** Se realiza una tokenización con granularidad de palabras en la que distintas palabras se identifican por un mismo término, dividiendo el texto para aplicar los modelos. Se utiliza como separador el espacio.
3. Normalización mediante **stemming**. Se realiza la normalización de las palabras haciendo uso de stemming. Se utiliza SnowBall Stemming debido a que cuida no reducir al mismo término las palabras con significados diferentes y evita realizar stemming en casos complicados (recordando que en cuestión de stemming es peor hacer un stemming excesivo que hacer menos). No siempre reduce los términos a una palabra existente, pero eso no representa una dificultad en este contexto de negocio.

Con respecto a **porqué se hace uso de stemming y no de lematización**, se realizó una investigación acerca de qué técnica era preferible utilizar teniendo en cuenta que se están trabajando textos en español. Se encontró que “la utilización del estemizado es más efectiva que la del lematizador en lenguas flexivas, como el español.” [7] Adicionalmente, se implementaron ambas técnicas en el notebook y al ejecutar el notebook con cada una de las técnicas, efectivamente se identificó que se tiene mejores métricas de evaluación haciendo uso de stemming

Adicionalmente a lo anterior, se debe realizar el proceso de generación de vectores. En esta oportunidad se elige usar el modelo de TF-IDF (Term Frequency-Inverse Document Frequency), el cual se considera mejor para esta aplicación de negocio debido a que asigna un peso a cada palabra que dependiendo de su frecuencia o rareza la clasifica como más o menos informativo. Más aún, nótese que en TF-IDF no se tienen problemas respecto a la longitud de los textos. Esto es preferible al modelo binario (que simplemente indica si un término está o no, siendo poco útil en este contexto) y a los modelos de frecuencias absoluta y relativa (pues estos últimos pueden dar un alto peso a palabras que simplemente son comunes pero que no resultan ser muy informativas).

 <p>Universidad de los Andes Colombia</p> <p>Acreditación institucional de alta calidad 10 años Ministerio de Educación Resolución 2812 9 de agosto de 2015</p>	<p>Ingeniería de Sistemas y Computación Pregrado</p> <p>ISIS-3301 – Inteligencia de Negocios Primer Proyecto Semestre: 2023-20</p>	
--	--	---

Modelado y Evaluación

Se realiza la aplicación de 3 algoritmos distintos para el aprendizaje automático descrito anteriormente, en el que el modelo debe ser capaz de recibir una entrada de texto arbitrario con una descripción de un problema relacionado a un ODS, y clasificarlo en uno de tres ODS: 6, 7 o 16, de forma precisa.

Se utilizan los siguientes tres métodos:

Método	Descripción y justificación de su uso
Red neuronal	<p>La red neuronal es un modelo basado en el funcionamiento del cerebro de las personas en donde se tienen capas de nodos, análogos a las neuronas, que tienen conexiones ponderadas (es decir con pesos, valores, asociados a ellas) los cuales se ajustan durante el proceso de entrenamiento. Los datos de entrada pasan a través de distintas capas de nodos o neuronas, y se optimizan los pesos de los nodos de forma que se minimice la diferencia entre la clasificación real de cada texto y la clasificación que otorga el modelo.</p> <p>Este es el método más intrincado de los utilizados y precisamente por eso es capaz de detectar patrones complejos en los datos que la regresión logística y el método de Naive Bayes quizás pueden pasar por encima. Se utiliza precisamente porque complementa los otros 2 métodos, enfocándose en características más complejas y profundas de los textos e identificando factores que pueden estar carentes en los otros 2 métodos.</p>
Regresión logística	<p>La regresión logística es un método que ajusta una función logística a los datos, con base en eso determina la probabilidad de que cada dato pertenezca a una categoría y a continuación con base en las probabilidades asigna la clasificación.</p> <p>Es especialmente útil en caso de que haya relaciones lineales entre los datos (lo cual no se sabe de antemano) e intenta maximizar la probabilidad de que la clasificación sea correcta. Además, provee dicha probabilidad. Por ende, se complementa con los demás modelos elegidos.</p>
Naive Bayes	<p>Naive Bayes se basa en el teorema de probabilidad condicional de Bayes. Presume de antemano independencia condicional entre los textos (lo cual, de nuevo, no se sabe de antemano si necesariamente es el caso) pero incluso con esa suposición ingenua (de ahí el <i>naive</i>) se caracteriza por arrojar resultados muy buenos, sobretodo en modelos de procesamiento de texto en lenguaje natural, lo cual es sumamente apropiado para este contexto de negocio.</p>

Respecto a la evaluación cuantitativa de los modelos, se utilizaron los siguientes métodos:

- Precisión (**accuracy**): Indica qué proporción o porcentaje de los registros fueron clasificados de forma correcta.

- **Macro precisión** o precisión por clase: Realiza un promedio de la precisión que tiene el modelo para cada clasificación. Permite evaluar si el modelo es bueno clasificando textos en el ODS 6, en el 7 y en el 16, o si solo es bueno clasificando textos en ODS específico.
- **Recall** o sensibilidad: Mide la tasa de “verdaderos positivos” clasificados correctamente, es decir, qué proporción de los textos fueron asignados su ODS correcta. Se complementa con la precisión, pues puede pasar que un modelo sea muy sensible pero poco preciso (por ejemplo, si clasifica todos los textos como el ODS 6, va a tener una alta sensibilidad a la clase del ODS 6 pero una mala precisión).
- Puntaje **F1**: Teniendo en cuenta que la precisión y la sensibilidad se complementan, el puntaje F1 los contempla ambos, siendo su promedio ponderado.

También se construye para los modelos la matriz de confusión, que permite visualizar los verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos para ver si el modelo está cometiendo algún error específico (e.g. si es propenso a caer en falsos positivos).

Se muestran a continuación los resultados de esas métricas para cada uno de los modelos. En el notebook se pueden evidenciar gráficas con estas métricas.

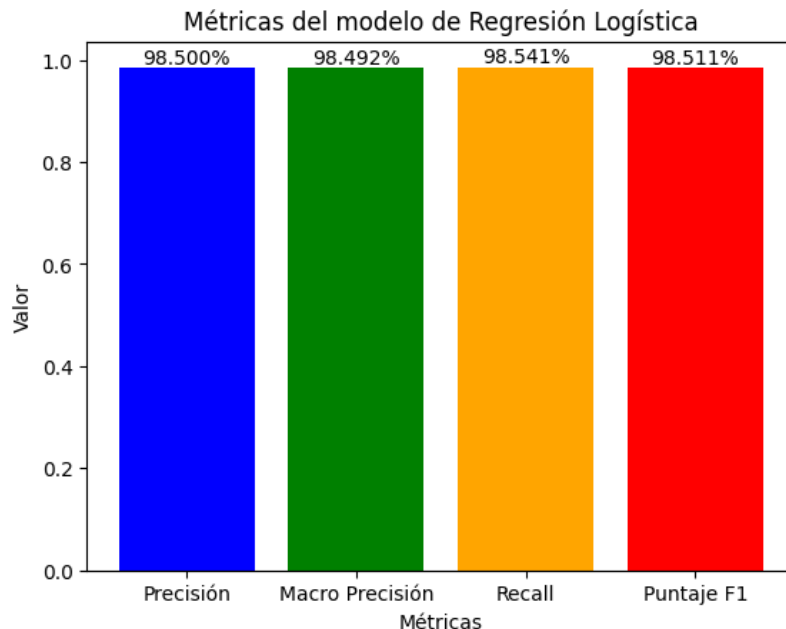
Método	Evaluación cuantitativa
Red neuronal	Precisión: 98.167% Macroprecisión: 98.182% Recall: 98.228% F1: 98.203%
Regresión logística	Precisión: 98.500% Macroprecisión: 98.492% Recall: 98.541% F1: 98.511%
Naive Bayes	Precisión: 98.167% Macroprecisión: 98.206% Recall: 98.206% F1: 98.205%

Se evidencia entonces que la regresión logística es el modelo con mejores indicadores de evaluación cuantitativa es por eso que en el siguiente apartado de este documento se recomienda hacer uso de ese modelo al negocio.

Resultados

En esta primera etapa del proyecto, el resultado obtenido es la construcción de un modelo que permite clasificar apartados de texto en una de las 3 posibles categorías de ODS: 6. “Garantizar la disponibilidad de agua y su gestión sostenible y el saneamiento para todos” [2], 7. “Garantizar el acceso a una energía asequible, segura, sostenible y moderna” [3], y 16. “Promover sociedades justas, pacíficas e inclusivas”. [4].



De los modelos implementados, se recomienda usar el modelo de regresión logística esto se debe a que presenta los siguientes indicadores:



Los cuales son muy positivos para el negocio, pues indican que haciendo uso de este modelo es altamente probable que dado un apartado de texto arbitrario este sea clasificado de forma correcta en una de las 3 clasificaciones: como el ODS 6, 7 o 16. En particular la precisión refleja que el 98.500% de los textos son clasificados correctamente; la macro precisión de magnitud similar refleja que la precisión mencionada anteriormente es aproximadamente igual para la clasificación de un texto en cualquiera de las categorías de ODS, no clasifica mejor en una categoría que en las otras; el *recall* de 95.541% indica que el modelo es hábil clasificando acertadamente los verdaderos positivos, los datos que efectivamente debería estar en la categoría que se les asigna; y por último, un puntaje F1 de 98,511%, quizás el más importante muestra que el modelo es robusto y equilibrado entre precisión y sensibilidad, siendo capaz de clasificar acertadamente la gran mayoría de textos que les son dados.

Esto resulta en un aporte significativo a favor de los objetivos de negocio, recordando que la motivación del proyecto era un modelo de aprendizaje automático que, dado un apartado de texto arbitrario, permitiese al UNFPA clasificarlo como relacionado a alguno de los tres ODS mencionados. Esto a su vez permitirá a la UNFPA utilizar la información que recopila de manera eficiente para apoyar u supervisar el cumplimiento de los ODS.




A partir de los resultados obtenidos, una posible estrategia que puede utilizar la organización es recolectar las opiniones y descripciones de las problemáticas relacionadas con los ODS, utilizar el modelo de clasificación automática para asignarles una categoría como ODS y a partir de eso entregar la información a las organizaciones que corresponde y a los agentes que actúan en favor de alguna de las causas. Por ejemplo, la información clasificada con el ODS 6 debe ser entregada a entes y organizaciones que se ocupan de facilitar el acceso al agua

 <div> <div>Universidad de los Andes</div> <div>Colombia</div> </div> <div> <div>Acreditación</div> <div>institución de alta calidad</div> <div>10 años</div> <div>Ministerio de Educación</div> <div>Resolución 2812</div> <div>9 de agosto del 2015</div> </div>	<div>Ingeniería de Sistemas y Computación</div> <div>Pregrado</div> <div>ISIS-3301 – Inteligencia de Negocios</div> <div>Primer Proyecto Semestre: 2023-20</div>	
---	--	---

potable. Este proceso automático de clasificación facilita entonces que los entes que tienen la posibilidad de ayudar tengan disponible la información pertinente para su trabajo y esto se realiza sin invertir capital humano excesivo. Se enfatiza en esta estrategia a través del mapa de actores expuesto en la siguiente sección.

Mapa de actores



Rol	Tipo de Actor	Beneficio	Riesgo
Desarrolladores del Modelo de Aprendizaje Automático (autores del presente documento)	Proveedor	Entrega un modelo de aprendizaje automático eficiente que clasifica apartados de texto arbitrarios en uno de tres ODS. Esto permite que la información que llega a la UNFP sea procesada de forma eficiente, clasificada acertadamente y que la información apropiada sea entregada a aquellos que pueden ayudar una determinada causa.	Posibles errores en el modelo podrían afectar la precisión de la clasificación y la confianza en la información proporcionada. Si se clasifica información erróneamente, podría entregarse información útil en un área a colaboradores que trabajan en otra área.
Área de recolección de información de la UNFPA	Usuario-Cliente	Utiliza la clasificación automática para no tener que clasificar a qué ODS corresponde cada registro de forma manual. Con eso, asigna recursos de manera eficiente a cada área relacionada con los ODS y brinda la información apropiada al área que la necesita sin que haya una inversión grande de tiempo y capital humano.	Debe comprenderse que este modelo en particular solo clasifica en 3 ODS, por lo que información que no corresponda con ninguno puede ser mal clasificada. Para el resto de ODS debe invertirse capital humano o solicitarnos un nuevo modelo enfocado a eso.
Equipo de comunicaciones de la UNFPA	Beneficiado	A partir de la clasificación tiene acceso a información sobre cada tema para así poder difundir más fácilmente el impacto de los	Nuevamente es imperativo que el equipo reconozca que el modelo no es la verdad absoluta y solo

 		Ingeniería de Sistemas y Computación Pregrado ISIS-3301 – Inteligencia de Negocios Primer Proyecto Semestre: 2023-20		
		<p>problemáticas y proyectos relacionados con los ODS 6, 7 y 16. Eso ayuda a hacerlos públicos y a una mayor difusión, lo cual a su vez brinda más atención y colaboración a la causa de cumplir los ODS.</p>	<p>clasifica en 3 ODS, por lo que información que no corresponda con ninguno puede estar mal clasificada.</p>	
Área de finanzas	Financiador	<p>Con la clasificación, se puede hacer una asignación eficiente de recursos financieros hacia proyectos específicos relacionados con cada ODS, sabiendo qué problemáticas tienen, sus descripciones y opiniones sobre ellas.</p> <p>Se pueden tomar entonces decisiones informadas de donde y en qué problemáticas invertir.</p>	<p>Si el modelo no clasifica bien, las decisiones y la asignación de fondos podría estar basada en información errónea o engañosa, lo cual es lo opuesto a lo que se desea y no estaría alineada con las necesidades reales.</p>	
Equipo de Evaluación de Impacto	Usuario-Cliente	<p>Utiliza el modelo y las clasificaciones que otorga para evaluar el impacto de las intervenciones en los ODS, por ejemplo pasándole textos de acciones que se han tomado.</p>	<p>Si las acciones no están dirigidas a uno de los ODS que el modelo contempla, se obtendrá una clasificación errónea y engañosa.</p>	

Trabajo en equipo

A continuación, se muestran los roles que adoptó cada uno de los integrantes del grupo:

Integrante	Participación y roles adoptados
Federico Melo Barrero	<p>Roles: Líder de proyecto, Líder de analítica.</p> <p>Acumulado de número de horas dedicadas al proyecto: aproximadamente 12 horas de trabajo neto.</p> <p>Algoritmo(s) trabajados: Red Neuronal y Naive Bayes</p> <p>Retos enfrentados en el proyecto y formas planteadas para resolverlos: Gestionar eficazmente el tiempo y la</p>



 <div> <div>Universidad de los Andes</div> <div>Colombia</div> </div> <div> <div>Acreditación</div> <div>institución de alta calidad</div> <div>10 años</div> <div>Ministerio de Educación</div> <div>Resolución 1812</div> <div>9 de agosto del 2015</div> </div>	<div>Ingeniería de Sistemas y Computación</div> <div>Pregrado</div> <div>ISIS-3301 – Inteligencia de Negocios</div> <div>Primer Proyecto Semestre: 2023-20</div>	
	carga de trabajo, implementé una estructura de gestión del tiempo más efectiva y me rodeé de un entorno de trabajo inspirador. Puntos repartidos: 50/100.	
Shadith Pérez Rivera	Roles: Líder de negocio, Líder de datos. Acumulado de número de horas dedicadas al proyecto: aproximadamente 12 horas de trabajo neto. Algoritmo(s) trabajados: Red Neuronal y Regresión Logística Retos enfrentados en el proyecto y formas planteadas para resolverlos: La gestión del tiempo. Para abordar este reto, implementé una planificación detallada, asignando tareas específicas a intervalos regulares y estableciendo plazos realistas. Puntos repartidos: 50/100.	

Reflexión:

Este proyecto podemos resaltar la importancia y relevancia de cómo la tecnología y el aprendizaje automático pueden ser utilizados para abordar desafíos globales y mejorar la eficiencia en la toma de decisiones en la UNFPA. Se proporciona información relevante a las partes interesadas, lo que puede impulsar acciones y colaboraciones más eficientes en la consecución de los ODS.

La elección de utilizar varios modelos de aprendizaje automático, como la red neuronal, la regresión logística y Naive Bayes, demuestra un enfoque integral para abordar la complejidad de los datos, donde cada modelo aporta su propia capacidad para detectar patrones y relaciones dentro de los textos, lo que mejora la solidez del sistema. Por otro lado, el análisis y preprocesamiento de datos son fundamentales para garantizar la calidad de las entradas al modelo, junto a la normalización mediante stemming y la eliminación de palabras de parada técnicas efectivas que han demostrado su utilidad en este contexto.

Sin embargo, más allá de los aspectos técnicos, en este proyecto también resaltamos la importancia del trabajo en equipo donde experimentamos complicaciones relacionadas con cambios en el equipo, también en la gestión del tiempo y resaltamos la importancia de las habilidades personales, la colaboración en equipo y la adaptabilidad en el logro de nuestros objetivos.

 <p>Universidad de los Andes Colombia</p> <p>Acreditación institucional de alta calidad 10 años Ministerio de Educación Resolución 3522 9 de agosto de 2015</p>	<p>Ingeniería de Sistemas y Computación Pregrado</p> <p>ISIS-3301 – Inteligencia de Negocios Primer Proyecto Semestre: 2023-20</p>	 <p>ABET Engineering Accreditation Commission</p>
--	--	--

Lista de Referencias

- [1] ONU. (2023). *Objetivos de Desarrollo Sostenible* [En línea]. Disponible en: <https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/#>
- [2] ONU. (2023). *Objetivo 6: Garantizar la disponibilidad de agua y su gestión sostenible y el saneamiento para todos* [En línea]. Disponible en: <https://www.un.org/sustainabledevelopment/es/water-and-sanitation/>
- [3] ONU. (2023). *Objetivo 7: Garantizar el acceso a una energía asequible, segura, sostenible y moderna* [En línea]. Disponible en: <https://www.un.org/sustainabledevelopment/es/energy/>
<https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/#>
- [4] ONU. (2023). *Objetivo 16: Promover sociedades justas, pacíficas e inclusivas* [En línea]. Disponible en: <https://www.un.org/sustainabledevelopment/es/peace-justice/>
- [5] La República. (2023). *En Colombia, 3,2 millones de personas no tienen acceso al servicio de agua potable.* [En línea]. Disponible en: <https://www.larepublica.co/economia/en-el-colombia-3-2-millones-de-personas-no-tienen-acceso-al-servicio-de-agua-potable-3576736>
- [6] Statista. (2023). *Número de homicidios cometidos por cada 100.000 habitantes en Colombia de 2014 a 2022.* [En línea]. Disponible en: <https://es.statista.com/estadisticas/1289833/tasa-de-homicidios-colombia/#:~:text=En%202022%2C%20hubo%20aproximadamente%2026,por%20debajo%20de%20los%2025.>
- [7] E. Perdomo, J. Díaz, A. Ojeda y N. Amador. *Análisis de los procesos de lematización y estemizado en lingüística computacional*. Instituto de Literatura y Lingüística. Cuba. 2017. Disponible en: https://www.researchgate.net/profile/Josval-Diaz-Blanco/publication/322364515_ANALISIS_DE_LOS_PROCESOS_DE_LEMATIZACION_Y_ESTEMIZADO_EN_LINGUISTICA_COMPUTACIONAL/links/5a560e1445851547b1be8080/ANALISIS-DE-LOS-PROCESOS-DE-LEMATIZACION-Y-ESTEMIZADO-EN-LINGUEISTICA-COMPUTACIONAL.pdf