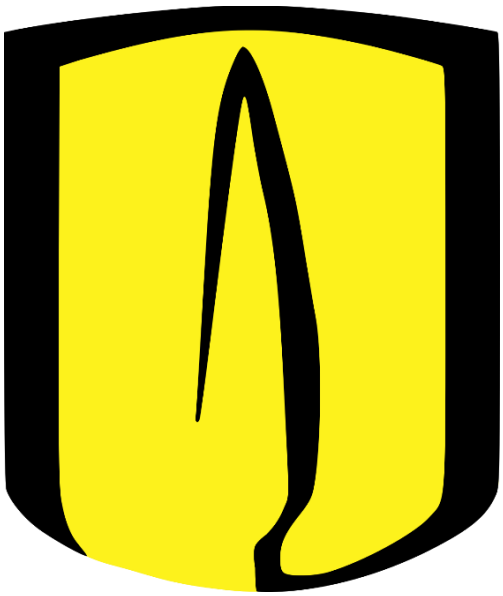
Informe Iteración #1 Proyecto 1 Inteligencia de Negocios

Bogotá DC Universidad de los Andes



Grupo 15 Miembros:

Juan Andrés Bernal Gil – ja.bernalg1@uniandes.edu.co

Juan Felipe Serrano Martínez – j.serrano@uniandes.edu.co

Santiago Diaz – s.diazm1@uniandes.edu.co

Bogotá DC

15 de octubre del 2023

Entendimiento del negocio y enfoque analítico:

Oportunidad de negocio:

La ONU en el año 2015, toma la decisión de crear la agenda 2030, que consiste en el desarrollo de 17 objetivos de desarrollo sostenibles y 169 metas que deberán de cumplirse para esa fecha con el fin de asegurar paz, prosperidad, desarrollo y calidad de vida para todas las personas del mundo.

¿Qué es un ODS?: un ODS es un objetivo de desarrollo sostenible planteado por la ONU para la agenda 2030, el cual se espera que se cumpla para dicha fecha, dentro de los ODS's planteados por la ONU, se encuentran problemas globales tal como el hambre, pobreza, igualdad de género, calidad de vida, desarrollo de infraestructura, educación, y demás temas de vital importancia para la humanidad en los próximos años. Otra finalidad de los ODS es asegurar la prosperidad al momento de proteger el planeta, se espera que todos los países pertenecientes a la ONU trabajen para satisfacer estos objetivos planteados.

Se espera poder crear una herramienta para lograr clasificar el texto cada uno de los ODS asignados a nuestro grupo, para la realización de propuestas acordes a los ODS, se espera desarrollar los algoritmos de aprendizaje automático pertinentes para cumplir con los objetivos planteados por la organización.

Dentro de los objetivos de negocio identificados, tomamos:

Objetivos del negocio

1. Mejorar la Eficiencia del Análisis de Datos Textuales:

Automatizar el proceso de clasificación de textos relacionados con problemáticas locales, eliminando la necesidad de una revisión manual extensiva y permitiendo una respuesta y análisis más rápidos.

2. Facilitar la Interacción con los Resultados del Modelo:

Crear una interfaz de usuario o aplicación amigable que permita a los stakeholders acceder y utilizar los resultados del modelo de manera efectiva, facilitando la toma de decisiones y la formulación de estrategias.

3. Contribuir al Cumplimiento de los ODS en el Contexto Local:

A través de la información y análisis proporcionados por el modelo, orientar la creación e implementación de políticas y estrategias que aborden de manera efectiva las problemáticas locales, contribuyendo al cumplimiento de los ODS.

Criterios de éxito

1. Precisión del Modelo:

Alcanzar o superar un umbral establecido de precisión en la clasificación de textos, asegurando que la información analizada y categorizada es fiable y puede ser utilizada para fundamentar decisiones.

2. Usabilidad de la Aplicación:

Desarrollar una plataforma que sea intuitiva y accesible para los usuarios, permitiendo que interactúen y extraigan valor de los datos procesados sin barreras tecnológicas significativas.

3. Impacto en la Toma de Decisiones:

Que los insights y análisis proporcionados por el modelo sean incorporados en el proceso de toma de decisiones y formulación de políticas, evidenciando un impacto directo en las estrategias implementadas.

4. Alineación con los ODS:

Lograr que las políticas y estrategias derivadas del análisis de los datos estén alineadas con los ODS y contribuyan de manera tangible a la mejora de las

ODS involucrados en el proyecto:

En nuestra temática asignada, nos correspondió trabajar con los ODS 3,4 y 5, que corresponden respectivamente a:

ODS 3: Salud y bienestar

Objetivo: Garantizar una vida sana y promover el bienestar para todos en todas las edades.

Aspectos clave: Este objetivo se centra en la reducción de la mortalidad materna e infantil, el combate contra enfermedades como el VIH/SIDA, la malaria y otras enfermedades tropicales, y la promoción de la salud mental y el bienestar. También aborda la necesidad de fortalecer la prevención y el tratamiento del consumo de sustancias adictivas, la cobertura sanitaria universal y la reducción de los riesgos de accidentes de tráfico, entre otros.

ODS 4: Educación de calidad

Objetivo: Garantizar una educación inclusiva, equitativa y de calidad y promover oportunidades de aprendizaje durante toda la vida para todos.

Aspectos clave: Este objetivo hace énfasis en garantizar que todos los niños completen la educación primaria y secundaria, que tengan acceso a un desarrollo inicial y cuidado preescolar de calidad, y que se asegure el acceso igualitario a una formación técnica, profesional y de educación superior. También busca incrementar el número de jóvenes y adultos con habilidades relevantes para el

empleo, el trabajo decente y el emprendimiento, y promover la educación para el desarrollo sostenible y la cultura de paz.

ODS 5: Igualdad de género

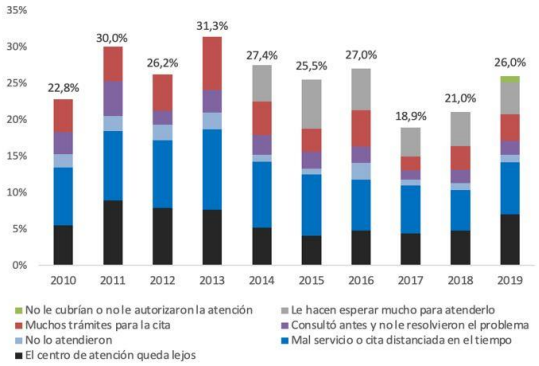
Objetivo: Lograr la igualdad entre los géneros y empoderar a todas las mujeres y niñas.

Aspectos clave: Se enfoca en poner fin a todas las formas de discriminación, violencia y prácticas nocivas contra las mujeres y niñas, garantizar su plena participación y oportunidades en el liderazgo y la toma de decisiones en todos los ámbitos de la vida, asegurar el acceso universal a la salud sexual y reproductiva, y reconocer y valorar el trabajo de cuidados no remunerado realizado por las mujeres, entre otros temas. (ONU, 2015)

Impacto de estos ODS en Colombia:

Colombia se ha caracterizado por ser un país con bastantes carencias en el ámbito de sostenibilidad y desarrollo de la población. De acuerdo con (Lizarazo, 2021), alrededor del 26% de las personas que solicitaron un servicio médico, no recibieron su servicio o atención solicitado, lo que indica una gran carencia en el sistema de salud colombiano.

Gráfica 1. Porcentaje de personas con problemas de salud que no solicita o no recibe atención médica debido a barreras de oferta. Colombia, 2010-2019.



Fuente: Encuesta de Calidad de Vida (DANE). Cálculos: CPC.

(Lizarazo, 2021)

Además, con respecto al tema del cuarto ODS, se evidencia que alrededor del 5.24% de los colombianos era analfabeta (Avila, 2018), además, se evidencia que la zona mas critica era la zona rural, donde pertenece el 70% de la población estudiantil en Colombia. Además, dada la dificultad de llegar a estos territorios, al igual que garantizar una educación de calidad para los estudiantes, alrededor de 500.000 estudiantes abandonan anualmente el sistema educativo, esto corresponde alrededor del 10% de la población estudiantil colombiana. Por lo que nuestro proyecto busca identificar estas propuestas y que sean aplicables de la mejor forma en el contexto colombiano.

Con respecto a nuestro ultimo ODS a trabajar, se ha evidenciado que: Las mujeres entre los 13 y 49 años que alguna vez han estado unidas o que actualmente lo están:

El 23,9 % ha sufrido intimidación de su pareja.

El 39.9 % sufren subvaloración por parte de su pareja.

El 57.9 % han sufrido de acciones de control por parte de su pareja.

El 31.1 % han sido víctimas de violencia económica y patrimonial por parte de su pareja. (UNICEF, 2019)

Estos indicadores, muestran que entre el 25% y 60% de las mujeres ha sufrido algún tipo de violencia de genero por parte de su pareja, lo cual es bastante negativo en la sociedad colombiana.

Organización y rol dentro de ella que se beneficia con la oportunidad:

Teniendo en cuenta los datos y evidencias presentadas previamente, evidenciamos que la sociedad en general y el gobierno son nuestras organizaciones a trabajar, los cuales pueden estar fuertemente beneficiados por nuestro proyecto, gracias a la identificación de sus necesidades especificas y conocimiento de como abordar cada una de sus problemáticas particulares. Entrando un poco mas en detalle, los beneficiados principalmente serán las personas de bajos recursos, que han sufrido problemas en el sistema de salud, en el sistema educativo y violencia de género, esto con el fin de identificar sus problemáticas y mejorar su calidad de vida. Como fue mencionado anteriormente un gran caso de aplicación de nuestra propuesta es la sociedad colombiana y similares que están en vía de desarrollo

Tabla de planeación de proyecto:

Elementos	Respuesta
Oportunidad/problema Negocio	Desarrollar un modelo de aprendizaje automático capaz de clasificar automáticamente textos diversos (como documentos y comentarios) en categorías que correspondan a los 17 Objetivos de Desarrollo Sostenible (ODS) de la ONU, facilitando así el análisis y monitoreo de información textual relevante para la evaluación y seguimiento de políticas públicas y estrategias relacionadas con la Agenda 2030, sin tener que hacer el proceso expenduoso habitual de revision a mano
Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático) incluya las técnicas y algoritmos que propone	Objetivo: clasificar un texto en una categoría correspondiente a los Objetivos de Desarrollo Sostenible (ODS) propuestos por la ONU, basándose en su contenido. Categoria: Clasificacion. Tipo de Aprendizaje: Supervisado. Tareas: Clasificar. Algoritmos: KNN(K-Nearest Neighbors), Árboles de Decisión, Máquinas de Soporte Vectorial (SVM)

utilizar	
Organización y rol dentro de ella que se beneficia con la oportunidad definida	Organizacion Beneficiada: Fondo de Población de las Naciones Unidas (UNFPA). Rol que se Beneficia: Analistas de Políticas y Especialistas en Desarrollo Sostenible
Contacto con experto externo al proyecto	a.guijo@uniandes.edu.co - Antonio Xue Aguijo - Llamada por Discord - 15/10/2023

Análisis Perfilamiento de los Datos:

El conjunto de datos proporcionado para el proyecto de clasificación de textos en referencia a los Objetivos de Desarrollo Sostenible (ODS o "sdg" por sus siglas en inglés) se compone de 3000 registros distribuidos en dos columnas: Textos_espanol y sdg. La primera alberga textos en español y presenta un tipo de dato objeto, mientras que la segunda, que sirve como etiqueta clasificatoria para los textos y es de tipo entero, pese a que podría considerarse como categórica dado su contexto y naturaleza.

En una inspección más detallada de la columna sdg, se observa que, aunque los datos son numéricos (int64), se comportan más como una variable categórica, ya que representan categorías (ODS) más que valores cuantitativos. Los datos revelan que la moda es 3, y al indagar en los valores únicos y sus cuentas, es evidente que esta variable solo toma los valores 3, 4 y 5, todos con 1000 instancias respectivamente. Esta uniformidad en la distribución es fundamental y estratégicamente valiosa, dado que se pueden preservar estas proporciones cuando se divida el conjunto de datos para entrenamiento y evaluación del modelo, evitando así sesgos y proporcionando robustez a nuestro futuro modelo predictivo.

La columna Textos_espanol, por otro lado, plantea desafíos y consideraciones propias del procesamiento de lenguaje natural (NLP). Antes de ser usado en cualquier modelo analítico, este texto requerirá un cuidadoso y probablemente complejo preprocesamiento para transformar el contenido textual en una forma que pueda ser interpretada y utilizada por los algoritmos de aprendizaje automático. Este proceso puede incluir una serie de pasos como la limpieza del texto, la tokenización, la eliminación de palabras vacías y la vectorización, todos ellos esenciales para construir representaciones numéricas del texto que mantengan su semántica original.

Análisis Calidad de Datos:

Completitud:

El conjunto de datos mostró una completitud del 100% en todas las columnas, no presentando ni un solo valor nulo tanto en la columna Textos_espanol como en sdg. Además, se verificó que no hay cadenas vacías presentes en las transcripciones.

Unicidad:

Los datos superaron las pruebas de unicidad, estableciendo que no hay filas repetidas en el conjunto de datos. En lo que respecta a la unicidad, es importante mencionar que incluso después de dividir los textos en tokens, si hay coincidencias entre ellos, estas no se ven como duplicados no deseados. De hecho, estas coincidencias se perciben como relevantes y válidas para el modelo que se desarrollará.

Validez:

La validez de las transcripciones de texto fue un punto de revisión interesante, aplicándose un criterio práctico de requerir al menos 25 palabras o la presencia de dos oraciones aproximadamente como una medida básica de validez. En este contexto, las 3000 transcripciones cumplieron con este requisito de longitud y se consideraron válidas. En cuanto a los valores en la columna sdg, se confirmó que todos los valores son 3, 4 o 5, lo cual está en línea con las expectativas y por lo tanto se consideran válidos.

Consistencia:

En términos de consistencia, se encontraron desafíos específicos en la columna de texto. Se detectó que una abrumadora mayoría de las filas, 2996 para ser exactos, contenían caracteres no ASCII. Esta inclusión de caracteres especiales necesita una revisión detallada para decidir sobre la necesidad de limpieza, reemplazo o manejo especial durante la tokenización para asegurar una representación coherente en el modelado del lenguaje.

Reporte preparación de los Datos:

Con respecto a la preparación de los datos, se tomaron las siguientes decisiones:

- Se eliminaron los valores ASCII de las cadenas que no pertenecen y no tienen sentido alguno con el español, los cuales son: œ, ˜, , –, —, ’, “, ”, …, €, ™, ¡, ¢, £, ª, «, ®, °, º, ». Simplemente serán eliminados estos caracteres.
- Se identifican que valores fueron mal codificados y tienen que ver con tildes y símbolos en español, se puede evidenciar nuestro reemplazo de la siguiente forma: 'Ãº': 'ú', 'Ã³': 'ó','Ã±': 'ñ', 'Ã¡': 'á','Ã': 'í', 'í©':'é'. Teniendo esto, ya aseguramos el correcto manejo de las tildes. Los demás símbolos fueron eliminados.

- Se tokeniza el texto, es decir, se obtiene palabra por palabra para cada una de las frases almacenadas. Además, se toma la decision de trabajar con los Lemas de las palabras, para así darle un poco de flexibilidad al modelo y no limitarlo a las palabras directamente sino a las que se derivan de una palabra madre.

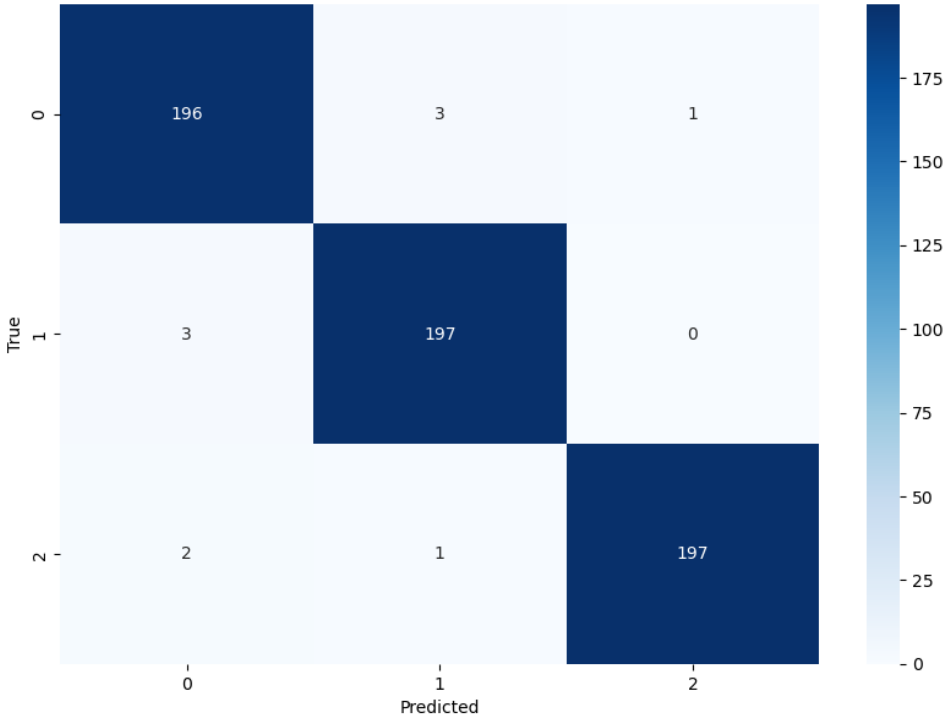
- Adicionalmente a este paso, se decide implementar una normalización para la bolsa de palabras (Dada la gran cantidad que poseemos de esta), el método para esta normalización fue tf-idf, dado que consideramos que hay una cantidad de palabras que se repiten considerablemente más que otras y pueden ser determinantes en la clasificación.

- Finalmente, separamos nuestros datos en variable dependiente (Y) que hace referencia a la etiqueta y variable independiente (X) que se refiere a la frase.

Con base a estos datos, se crea un entrenamiento para todos los algoritmos del 20% con sus mejores hiperparametros.

Resultados de los algoritmos implementados:

KNN:

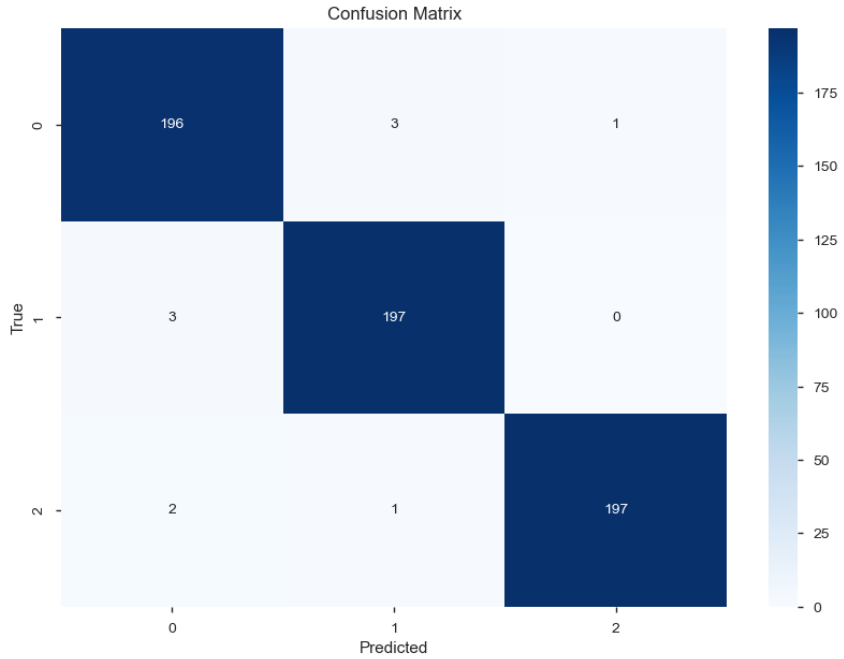


La implementación del algoritmo KNN (K-Nearest Neighbors) con K=26 ha demostrado ser notablemente efectiva, logrando una precisión general del 98%, según se refleja en los resultados proporcionados. La precisión, recall y f1-score

para cada clase (3, 4 y 5) se mantienen consistentemente altas, evidenciando una habilidad robusta del modelo para clasificar correctamente las instancias de todas las categorías presentes en el conjunto de datos.

Al observar más detenidamente las métricas de cada clase, todas ellas están en un rango cercano al 98-99%, lo que indica un rendimiento uniformemente alto a través de las diferentes clases, lo que es crucial para asegurar que el modelo no solo es preciso en general, sino también equitativamente preciso para cada clase individual. Además, la consistencia entre precisión, recall y f1-score sugiere que el modelo ha logrado un equilibrio saludable entre precisión y exhaustividad, y que no está siendo excesivamente penalizado por falsos positivos o falsos negativos en una proporción significativa

Máquinas de Soporte Vectorial:



En el proceso de desarrollo, se empleó una búsqueda de hiperparámetros mediante la metodología de búsqueda en cuadrícula. Se exploró distintas combinaciones de C, kernel, gamma, class_weight, y degree, todo esto con el objetivo de optimizar la métrica F1.

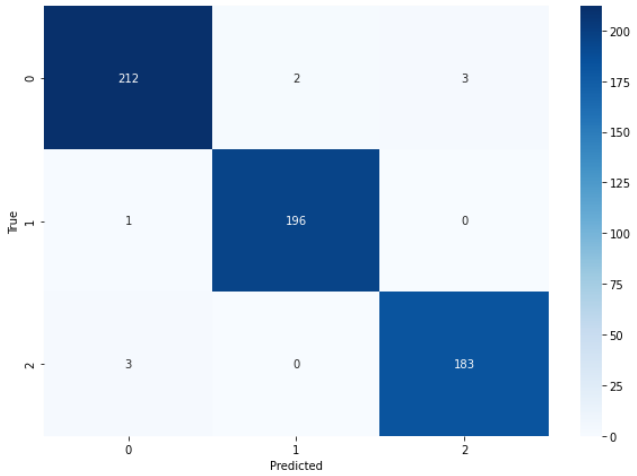
El parámetro C controla el equilibrio entre una clasificación precisa y la suavidad del margen de decisión, mientras que kernel especifica la función utilizada para transformar los datos en un espacio de características alternativo, con rbf refiriéndose a un kernel gaussiano. El parámetro gamma regula cuánto influye una única instancia de entrenamiento, con una influencia que varía de ‘lejana’ a ‘local’

dependiendo si el valor es bajo o alto, respectivamente. Por otro lado, class_weight ajusta las ponderaciones para las clases y degree detalla el grado del polinomio en el kernel polinómico, aunque no afecta si el kernel es de tipo rbf.

En cuanto a la evaluación de la calidad del modelo con los hiperparámetros seleccionados, este demostró un rendimiento significativamente alto, con una precisión del 98.33%. Las métricas adicionales de precisión, recall y f1-score para cada clase, y los promedios ponderados y macro, también destacaron la capacidad del modelo para clasificar efectivamente las instancias en las tres clases, demostrando así la eficacia y fiabilidad del modelo para la clasificación propuesta con la configuración de hiperparámetros encontrada.

Regresión logística:

Se obtuvo un porcentaje de accuracy del 98.5% en la implementación del algoritmo, evidenciado en la siguiente matriz de confusión: Esto indica que el algoritmo fue correctamente implementado y funcional con base a nuestra preparación de datos y estudio de negocio.



Se evidencia un modelo de regresión logística bastante bueno con métricas de alrededor del 97.67% de asertividad, además de un coeficiente f1 de 98%. El modelo es capaz de acertar casi todos los casos, lo cual es bastante positivo para el negocio apara a la identificación de los ODS a estudiar. Hay que aclarar que para una regresión logística hay otros parámetros que pueden determinar tales como la importancia de los betas y significancia individual, con base a cada palabra. Dado que es un modelo de texto y se manejan alrededor de 10.000 palabras, es bastante completo determinar la significancia individual de cada beta. Por lo que, es más conveniente estudiar principalmente las métricas del modelo

Mapa de actores relacionados con un producto de datos.

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
Gobierno nacional	Tomador de decisión	Tomar las mejores decisiones para el desarrollo sostenible de la sociedad	Si el modelo no funciona correctamente, no se podrá clasificar correctamente las necesidades de la población.
Entidades no gubernamentales (ONG’s y demás)	Usuario - Cliente	Obtienen acceso a la información y enfoque dado por el gobierno para orientar sus intervenciones y proyectos en las áreas relacionadas	De no tener clara la clasificación y enfoque propuesto, se pueden desperdiciar recursos en cuanto a la implementación de los ODS
Instituciones educativas	Beneficiado	Reciben los recursos por parte del gobierno para mejorar su infraestructura	No reciben los recursos correctamente, desperdicio de dinero y no satisfacción del ODS asociado
Entidades prestadoras de servicios de salud	Beneficiado	Se satisface las necesidades de salud y calidad de vida de las personas	En caso de un fallo en el modelo, se puede empeorar la calidad de vida de las personas al igual que, llegar al punto de la muerte.
Ministerio de la igualdad	Usuario - Cliente	Garantizar las demandas y problemas asociados con la cultura y respeto de la igualdad de genero	En el caso de que el modelo falle, se quedan en impunidad y desconocimiento
Personas de la sociedad	Beneficiados	Se garantizara la satisfacción de los ODS para el mejoramiento de su calidad de vida y desarrollo de autorrealización	En caso de que el modelo no funcione correctamente, se ve comprometida la calidad de vida de las personas y no satisfacción de los ODS

Bibliografía

Avila, L. G. (2018). *ANEIA - Uniandes*. Obtenido de ANEIA - Uniandes: https://aneia.uniandes.edu.co/2019/04/la-educacion-un-grave-problema-de-la-ruralidad-colombiana/#:~:text=La%20falta%20de%20instituciones%20educativas,la%20escolaridad%20en%20varias%20regiones.

Lizarazo, L. (Febrero de 2021). *Compite*. Obtenido de Compite: https://compite.com.co/blog_cpc/algunos-desafios-del-sistema-de-salud-en-colombia/

ONU. (2015). *ONU*. Obtenido de ONU : https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/

UNICEF. (2019). *UNICEF*. Obtenido de UNICEF: https://www.unicef.org/colombia/genero