**Proyecto 1 - Entrega 1-** **Inteligencia de Negocios**

**Estudiantes:**

* Josue Mauricio Rivera Ladino – 201914138
* Alejandro Hernández Ordoñez - 202111716
* Santiago Andrés Vela Aguilera - 202026767202026767

**Asistente de Estadística:**

* Juan David Niño Pallares



14 de octubre, 2023

Inteligencia de Negocios

, Universidad de los Andes

Tabla de contenido

[*1. Entendimiento del negocio y enfoque analítico.* 2](#_Toc148291597)

[*1.1.* *Descripción del enfoque analítico para alcanzar los objetivos del negocio.* 3](#_Toc148291598)

[1.1.2. Recopilación y Análisis de Datos 3](#_Toc148291599)

[1.1.3. Evaluación y Retroalimentación. 3](#_Toc148291600)

[1.1.4. Contacto con experto externo al proyecto 4](#_Toc148291601)

[***2.*** ***Descarga y entendimiento de datos.*** 4](#_Toc148291602)

[*3.* *Preparación de los datos.* 5](#_Toc148291603)

[*4.* *Modelamiento, validación y visualización.* 7](#_Toc148291604)

[***I.*** ***Modelo de clasificación de árbol de decisión:*** 7](#_Toc148291605)

[***II.*** ***Modelo Random forest*** 8](#_Toc148291606)

[*III.* *Modelo de clasificación K-NN* 9](#_Toc148291607)

[***5.*** ***Resultados.*** 10](#_Toc148291608)

[***6.*** ***Mapa de actores relacionado con un producto de datos creado con el modelo analítico construido.*** 10](#_Toc148291609)

[7. Trabajo en equipo. 11](#_Toc148291610)

[*8.* *Referencias* 12](#_Toc148291611)

## 1. Entendimiento del negocio y *enfoque analítico.*

**Objetivo 6: Agua Limpia y Saneamiento**

El objetivo de desarrollo sostenible 6 se centra en garantizar la disponibilidad y la gestión sostenible del agua y el saneamiento para todos. Su objetivo es asegurar que las personas tengan acceso a agua potable asequible y a instalaciones de saneamiento adecuadas.

En el contexto de un proyecto empresarial, el criterio de éxito podría incluir el suministro de agua potable de calidad a comunidades locales, la implementación de sistemas de gestión sostenible del agua y el saneamiento en las operaciones de la empresa, y la promoción de prácticas responsables en el uso del agua.

En cuanto al impacto que tendría sobre Colombia Abordar el ODS 6 puede tener un impacto significativo en Colombia al mejorar la calidad de vida de las comunidades, reducir las enfermedades relacionadas con el agua, promover la sostenibilidad ambiental y fomentar el desarrollo económico a través de una gestión adecuada de los recursos hídricos.

**Objetivo 7: Energía Asequible y no Contaminante**

El ODS 7 busca garantizar el acceso a una energía asequible, fiable, sostenible y moderna para todos. Esto implica promover fuentes de energía limpias y tecnologías más eficientes.

En cuanto a los criterios de éxito podrían incluir la adopción de fuentes de energía renovable, la mejora de la eficiencia energética en las operaciones de la empresa y el acceso a la energía para comunidades que carecen de él. Abordar el ODS 7 puede ayudar a Colombia a reducir su dependencia de fuentes de energía no renovable, disminuir las emisiones de carbono y brindar oportunidades económicas a través de la inversión en tecnologías limpias y el acceso a energía para poblaciones rurales.

**Objetivo 16: Paz, Justicia e Instituciones Sólidas**

El ODS 16 busca promover sociedades pacíficas e inclusivas para el desarrollo sostenible, proporcionar acceso a la justicia y construir instituciones eficaces y responsables.

Los criterios de éxito pueden incluir la promoción de la gobernanza responsable, la lucha contra la corrupción, la promoción de la paz y la justicia en las operaciones de la empresa y la colaboración con instituciones locales para fortalecer la administración de justicia.

Abordar el ODS 16 es fundamental para consolidar la paz y la estabilidad en Colombia, promoviendo la inversión y el desarrollo económico, así como garantizando un entorno de negocios más transparente y justo.

## *Descripción del enfoque analítico para alcanzar los objetivos del negocio.*

* + 1. **Identificación de Problemas Relevantes**

Según el diario La República, en Colombia, aproximadamente 3.2 millones de personas carecen de acceso a agua potable. Esta realidad puede parecer contradictoria, dado que Colombia ostenta el título de ser el segundo país más biodiverso de la región y uno de los más ricos en recursos naturales.

Adicionalmente, una investigación realizada por el periódico El Tiempo revela que regiones como el Pacífico Nariñense y el Pacífico Chocoano se encuentran entre las más afectadas por la falta de fuentes de energía limpias. Incluso en pleno siglo 21, algunas comunidades dependen de velas como única fuente de iluminación nocturna.

Por último, a lo largo del siglo XX y XXI, Colombia ha enfrentado altos niveles de violencia y desplazamientos forzados, principalmente debido a la prolongada lucha contra el narcotráfico y conflictos políticos. Lamentablemente, el país ha visto perder la vida de muchas personas en este proceso, y las cuestiones de justicia y repatriación de las víctimas aún ocupan una posición central en la agenda colombiana.Estos desafíos son un claro llamado a fortalecer los esfuerzos en relación con los Objetivos de Desarrollo Sostenible (ODS) en Colombia.

## Recopilación y Análisis de Datos

A raíz de una serie de entrevistas preseleccionadas, nuestro objetivo como organización es establecer un modelo destinado a la recopilación y almacenamiento de información escrita. Dicha información será sometida a un análisis profundo basado en su contenido. Las palabras y expresiones contenidas en las entrevistas(escritos) adquirirán un peso significativo, ya que determinarán en qué medida contribuyen al contenido general del proceso. Es importante destacar que algunos datos pueden no ser pertinentes a nuestros objetivos, por lo que serán excluidos.

Además, recibiremos información adicional a través de archivos en formato "xlsx", que deberá evaluarse y clasificarse según la información recopilada. Es relevante subrayar que, dependiendo del contenido de las palabras presentes en los nuevos archivos, se llevará a cabo una clasificación basada en los Objetivos de Desarrollo Sostenible (ODS) a los que corresponda.

## Evaluación y Retroalimentación.

En preparación para futuros textos, debemos desarrollar unos modelos de aprendizaje automático, como el primero basado en árboles para la clasificación. La evaluación continua es fundamental en este proceso. Inicialmente, al analizar y recopilar datos de prueba siguiendo los criterios previamente definidos, se buscarán las palabras y elementos que verdaderamente contribuyan al contenido, estableciendo categorías y patrones de clasificación sólidos.

Cuando los modelos funcionen, es esencial realizar evaluaciones regulares con datos no examinados, que serán proporcionados por la empresa contratista. Garantizando precisión y eficacia en la clasificación. La retroalimentación resultante de estas evaluaciones desempeña un papel crucial al identificar áreas de mejora y ajustar el modelo en consecuencia. La retroalimentación constante contribuye a perfeccionar los modelos a medida que enfrenta nuevos desafíos en la clasificación de textos, impulsando su evolución y mejora continua. comparar

El proceso de evaluación y retroalimentación es un componente esencial en la creación de modelos de aprendizaje automáticos. A medida que recibimos y procesamos nuevas entrevistas clasificadas, el modelo se adapta a las tendencias y variaciones emergentes en el lenguaje natural. La retroalimentación resultante de los resultados de clasificación nos permite realizar ajustes y refinamientos que mejoran su capacidad para comprender y categorizar textos de manera más precisa, considerando métricas como la **exactitud** (Accuracy), **precisión** (Precisión), **puntuación** F1 (F1 Score) y el área bajo la curva **ROC** (AUC-ROC).

Y mediante esta evaluación podemos comparar nuestros modelos creados y decidir cual es que se ajusta mejor al negocio y muestra los mejores resultados.

## Contacto con experto externo al proyecto

Nos contactamos con el estudiante de estadística el día 11 de octubre, primeramente, para tener opiniones y consejos sobre los temas que íbamos a tratar y también sobre la forma en la que debíamos abordar este problema, igualmente quedamos en reunirnos la otra semana para mostrarle avances y tener una retroalimentación sobre los modelos construido y que cosas deberíamos cambiar o modificar:

**Nombre**: Juan David Niño Pallares. **Correo**: jd.ninop1@uniandes.edu.co

Finalmente, para resumir toda la parte de entendimiento se realizo una tabla con el contenido más importante:

|  |  |
| --- | --- |
| Oportunidad/problema de Negocio | Abordar el ODS 6 en Colombia (Acceso a Agua Limpia y Saneamiento), el ODS 7 en Colombia (Energía Asequible y no Contaminante) y el ODS 16 en Colombia (Paz, Justicia e Instituciones Sólidas). |
| Enfoque analítico (Descripción del requerimiento desde el punto de vista de Aprendizaje automático) e incluya las técnicas y algoritmos que propone utilizar. | Utilización de técnicas de procesamiento de lenguaje natural (NLP) para analizar entrevistas y documentos relacionados con el acceso al agua, acceso a fuentes de energía en Colombia, con énfasis en energía renovable y datos sobre la violencia, la justicia y la gobernanza en Colombia. Clasificación basada en ODS. Evaluación continua de modelos de aprendizaje automático, como árboles de decisión y demás. Para luego realizar una comparación y definir mejor modelo para la problemática. |
| Organización y rol dentro de ella que se beneficia con la oportunidad definida. | Hay varias organizaciones que se pueden beneficiar como las siguientes:  Comunidades locales en Colombia, empresas enfocadas en servicios de agua y saneamiento, organismos gubernamentales relacionados con la gestión del agua, poblaciones rurales sin acceso a energía, empresas de energía sostenible, entidades gubernamentales relacionadas con la energía y el medio ambiente, Comunidades afectadas por la violencia, instituciones gubernamentales relacionadas con la justicia y la seguridad, empresas comprometidas con la responsabilidad social. |
| Contacto con experto externo al proyecto. | Juan David Niño Pallares, estudiante de estadística, para asesoramiento en el análisis de datos y la construcción de modelos de aprendizaje automático y los resultados obtenidos para poder tener conclusiones claras y correctas. |

### ***Descarga y entendimiento de datos.***

Se descargaron los datos entregados. Inicialmente se obtuvo información básica de los datos. A continuación, se presentan tablas e imágenes de información básica de los datos.

Tabla

Descripción generada automáticamente

Nos podemos dar cuenta que en referencia a nuestros ODS (objetivos de desarrollo sostenible) no tenemos problema de validez, debido a que no hay valores anómalos ya que en teoría los textos que se tienen cumplen con la clasificación "esperada"

**Dimensiones de calidad:**

* Unicidad: Unicidad: Se ven el número de valores repetidos o redundantes y en este caso se puede notar que no hay textos duplicados o redundantes.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

* Distribución: existe distribución uniforme de datos de entrenamiento de acuerdo con los labels, 1000 por cada label.

Texto

Descripción generada automáticamente

* Completitud: Se observa el número de valore nulos en los datos

Texto

Descripción generada automáticamente

* Estadísticas: Se saco conteo, moda, min y Max de los textos y tokens.

Texto, Carta

Descripción generada automáticamente

* Consistencia y validez: La consistencia hace referencia a la integridad de datos entre fuentes y observaciones. Por su parte, la validez mide si los datos hacen sentido para el contexto específico. En este sentido, se considera que en términos generales se cumple con estas métricas.

## **Preparación de los datos.**

Todo el procedimiento se puede ver en el notebook “Entendimiento\_preparacion\_modelamiento\_proyecto.ipynb“.

Para el preprocesamiento de los datos se realizó en siguiente procedimiento, por tres etapas:

* Limpieza de los datos.

Esta función es útil para eliminar caracteres no ASCII de unas listas de palabras tokenizadas en español. Esto permitirá limpiar los textos de caracteres especiales o caracteres no latinos que no son relevantes para el análisis de texto, en este caso debido a que importamos la librería vamos a solo trabajar con palabras en español

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

Esta función nos posibilitará la conversión de todas las palabras a minúsculas, una práctica común en el procesamiento de texto en español con el fin de garantizar que las palabras en mayúsculas y minúsculas sean tratadas de manera uniforme. Esta normalización facilita la comparación y el análisis de manera más efectiva.

Imagen que contiene Logotipo

Descripción generada automáticamente

Eliminar la puntuación de una lista de palabras es útil para eliminar caracteres no alfabéticos o especiales que no aportan información semántica al texto en español. como puntos y comas.

Texto

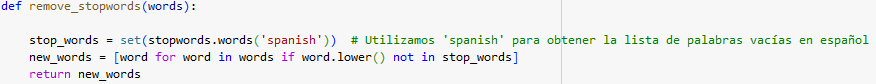
Descripción generada automáticamente

Reemplazamos números en una lista de palabras con su representación textual nos ayudara bastante debido que las ODS que tenemos son más de números y de cifras entonces al convertir el numero en palabra será más fácil darle un valor.

Texto

Descripción generada automáticamente

Igualmente, la eliminación de palabras vacías (stop words) para las entrevistas es esencial para eliminar palabras comunes que no aportan información significativa al análisis de texto, como "el", "la", "de", etc.



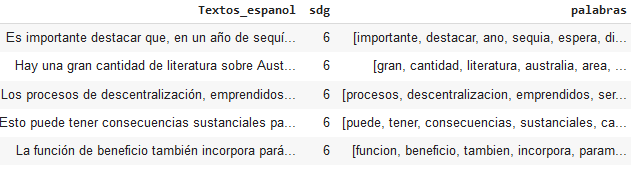
Esta función combina las funciones anteriores en un solo paso de preprocesamiento. Realiza la conversión a minúsculas, reemplaza números con su representación textual, elimina la puntuación, caracteres no ASCII y palabras vacías en español.

Texto

Descripción generada automáticamente

* Tokenización:

En primer lugar, realiza la tokenización de palabras, dividiendo el texto en unidades más manejables, lo que es fundamental para el procesamiento de texto. Luego, aplica una función de preprocesamiento que elimina el ruido, como signos de puntuación, números y stopwords, lo que mejora significativamente la calidad de los datos. Además, asegura la uniformidad en el formato de los datos, lo que facilita el análisis textual posterior al tratar las palabras de manera uniforme, independientemente de su capitalización o la presencia de elementos no deseados.



Podemos darnos cuenta como quedan los textos en los cuales ya no se aprecian palabras que generan ruido, de la misma forma, nos podemos dar cuenta que las palabras se han reducido esto con el fin de no leer toda la palabra si no que al saber una abreviación ya podemos saber el significado de la palabra, igualmente, nos permite mejorar la dimensionalidad en representaciones vectoriales del texto

* Normalización:

De igual manera usaremos dos técnicas fundamentales: stemming y lematización, aplicadas en el contexto del idioma español. El SnowballStemmer se encargará de llevar a las palabras a su forma raíz, lo que simplifica y normaliza el texto, facilitando la comparación de palabras relacionadas. Por otro lado, el **WordNetLemmatizer** realiza la lematización, que es un proceso más sofisticado que el stemming y permite reducir las palabras a su forma base o lema.

Combinando ambas técnicas, este código crea una función que transforma las palabras de entrada en una lista de sus formas stem y lema. Esto es valioso para tareas de análisis de texto, procesamiento de información y modelado de lenguaje, ya que ayuda a comprender y representar el texto de manera más efectiva, mejorando la calidad de los datos y su utilidad en diversas aplicaciones de NLP.

Texto

Descripción generada automáticamente

Aplicamos la lematización y Eliminación de Prefijos y Sufijos. Luego separamos la variable predictora y los textos a utilizar. Este caso como nuestro objetivo principal es usar unos modelos de clasificación de tendremos que separar las variables objetivo de las otras variables, algo bueno que tenemos es que la variable objetivo ya está en números, y no tenemos que preocuparnos por hacer algún cambio debido a que los números que hay son los números que necesitamos.

En este caso tenemos podemos usar dos técnicas de vectorización de texto comunes en el procesamiento de lenguaje natural. Ambas se pueden usar para convertir los textos en datos numérico Sin embargo decidimos usar la librería **TfidfVectorizer** ya que Asigna valores más altos a las palabras que son frecuentes en un documento, pero raras en otros, lo que permite identificar palabras que son distintivas de un documento en particular, de la misma manera, Puede manejar automáticamente palabras vacías y realizar normalización, lo que simplifica el preprocesamiento del texto.

En la mayoría de las aplicaciones de procesamiento de lenguaje natural, **TfidfVectorizer** suele ser la elección preferida debido a su capacidad para identificar palabras clave y su capacidad para lidiar con palabras vacías y realizar normalización automáticamente.

## *Modelamiento, validación y visualización.*

# **Modelo de clasificación de árbol de decisión:**

Primero creamos un modelo de clasificación de árbol de decisión con una profundidad máxima de 4 y utilizaremos la entropía como criterio de división. Este modelo se almacena en la variable árbol y se puede usar para entrenar y predecir con datos de clasificación. Luego de entrenarlo, particionar y evaluar, Ahora para esta nueva prueba aplicaremos Hiper parámetros con nuestro modelo para evaluar si es correcto o no.

Fijaremos un valor de max\_depth en 8, construyamos el modelo y probemos sobre validación. Utilicemos la exactitud como métrica, Ahora configuraremos validación cruzada K-Fold para evaluar el rendimiento de un modelo de aprendizaje automático. Con los parámetros n\_splits=10, lo que significa que estaremos dividiendo tus datos en 10 "pliegues" o subconjuntos. Y con la opción shuffle=True indicaremos que se realizará una mezcla aleatoria de los datos antes de la división, lo que ayuda a evitar sesgos en la partición.

Ahora realizaremos una búsqueda sistemática de hiperparámetros para el modelo de árbol de decisión y luego entrena un nuevo modelo con los hiperparámetros óptimos encontrados. Esto te permitirá mejorar el rendimiento del modelo al encontrar la mejor combinación de parámetros que se adapta a nuestros textos.

***Diagrama

Descripción generada automáticamente***

Tabla

Descripción generada automáticamente

los nuevos resultados indican un rendimiento excepcionalmente alto para el algoritmo. Pasando de una Exactitud de **0.90 a 0.99,** pasando de Recall de 0**.903 a 0.95**, y un puntuaje de F1 score de **0.90 a 0.95.**

Nos podemos dar cuenta que en la nueva matriz que hicimos la cantidad de acertar a los valores aumenta, respecto al ODS 16 son más la cantidad de registros que podemos predecir exactos, por otra parte, si bien es cierto que el número de falso positivo aumento esto no implica que mi algoritmo no haya mejorado

Gráfico, Gráfico de dispersión, Gráfico de rectángulos

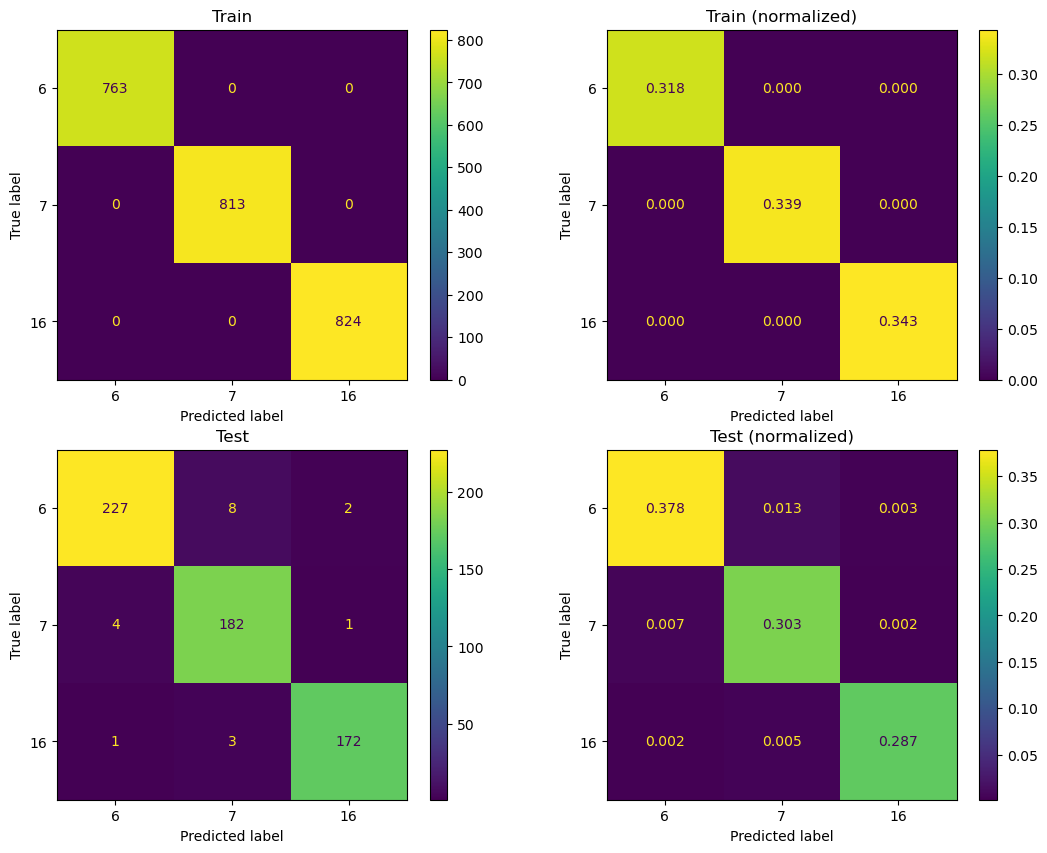
Descripción generada automáticamente

# ***Modelo Random forest***

Para empezar, se crea un modelo de clasificación Random Forest con una semilla aleatoria establecida en 1, de esta forma se puede garantizar la reproducibilidad de los resultados, posteriormente se entrena el modelo para que sea capaz de predecir las etiquetas de clase en base a las características de los datos usados para el entrenamiento y se generan predicciones a parir del modelo entrenado.

Se generaron 4 matrices de confusión, 2 para los datos de entrenamiento y 2 para los datos de prueba, a su vez, se tienen los datos sin alterar, y luego los datos normalizados. Se puede ver en la imagen que, en el caso de los datos de entrenamiento, se tienen 0 falsos positivos, lo que asegura la validez del entrenamiento realizado. En cuanto a los datos de prueba se puede decir que, aunque se presentan falsos positivos, se tiene un número aceptable.

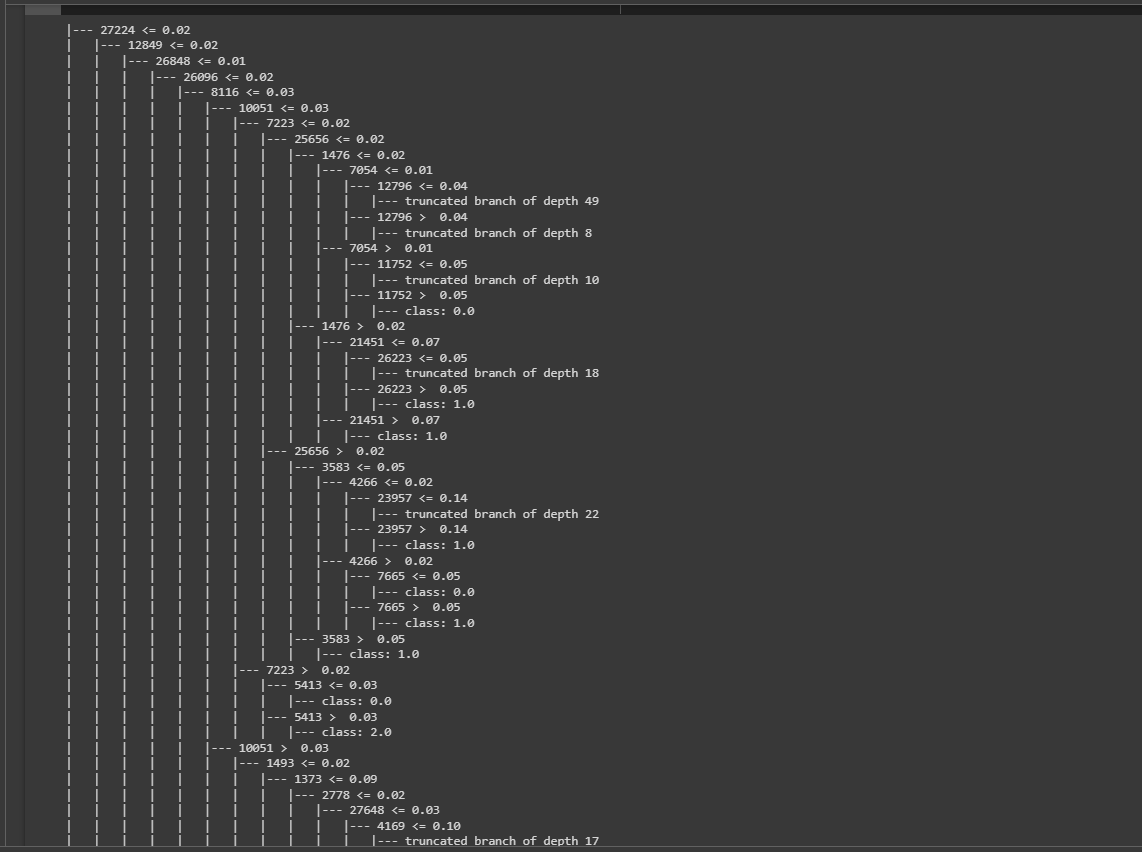
Para concluir con la validez del modelo se tienen las siguientes medidas que fueron calculadas para los datos de prueba y para los datos de entrenamiento:

******

Tabla

Descripción generada automáticamente

Se generó una visualización del primer árbol usado ya que, si bien Random Forest está compuesto por múltiples árboles de decisión, en este caso es posible visualizar uno de estos árboles formateado en texto como se ve a continuación.



## *Modelo de clasificación K-NN*

Con los datos se decidió crear un modelo K-NN de manera que se pudiera hacer una clasificación debido a su fácil interpretación, robustez ante ruido en los datos y adaptación a datos no lineales.

El proceso fue simple se entrenó con los datos de entrenamiento y se probó su de manera que se creó unas matrices de confusión normalizadas para examinar y un reporte de métricas de precisión.

***Gráfico, Gráfico de rectángulos

Descripción generada automáticamente***

Texto, Tabla

Descripción generada automáticamente

Como se puede ver los resultados obtenidos muestran en general un promedio del 94%, con su mayor dificultad en casos de falsos positivos, pero por lo demás se puede establecer que es un buen modelo.

Finalmente se realizó una visualización de manera que se pudiera ver como se clasifican los tokens textuales de acuerdo con sus valores vectorizados para ver cómo se agrupaban e identificar anomalías.

***Gráfico, Gráfico de dispersión

Descripción generada automáticamente***

# ***Resultados.***

Finalmente, luego de realizar las ejecuciones, entrenamientos y pruebas de los modelos, mediante las evaluaciones cualitativas y cuantitativas definimos que para definir el mejor modelo el modelo que tenga el mejor F1\_Score, y luego de compararlos el mejor modelo de acuerdo con el F1\_Score e incluso el promedio de los demás calificadores es el modelo RandomForest.

# ***Mapa de actores relacionado con un producto de datos creado con el modelo analítico construido.***

A continuación, se presenta una tabla que resume la información proporcionada en el texto en relación con el enfoque analítico para alcanzar los objetivos del negocio en el contexto de los Objetivos de Desarrollo Sostenible (ODS).

|  |  |  |  |
| --- | --- | --- | --- |
| **Rol dentro de la Organización** | **Tipo de Actor** | **Beneficio** | **Riesgo** |
| Identificación de Problemas Relevantes | Investigador interno o externo | Identificación de desafíos y necesidades relacionados con ODS 6, 7 y 16 en Colombia. | Posible falta de acceso a datos relevantes o información inexacta. |
| Recopilación y Análisis de Datos | Analista de datos | Recopilación y análisis de información escrita y datos en formato "xlsx" para evaluar la contribución a los ODS. | Riesgo de recopilar datos irrelevantes o de baja calidad. |
| Evaluación y Retroalimentación | Científico de datos o analista de aprendizaje automático | Desarrollo y ajuste de modelos de aprendizaje automático para clasificación. Evaluación continua de la precisión y eficacia. | Posible falta de precisión en la clasificación de textos y necesidad de ajustes continuos. |

## Trabajo en equipo*.*

**Líder de Proyecto: Josue Mauricio Rivera Ladino**

* Responsable de la gestión general del proyecto.
* Coordinación de los diferentes integrantes y partes interesadas.
* Asegurarse de que se cumplan los plazos y se alcancen los objetivos del proyecto.

**Líder de Negocio: Alejandro Hernández Ordoñez**

* Responsable de comprender y articular las necesidades del negocio y cómo se relacionan con los ODS.
* Identificación de oportunidades de negocio relacionadas con la implementación de los ODS.
* Colaboración con socios comerciales y partes interesadas en la ejecución de estrategias relacionadas con los ODS.

**Líder de Datos: Santiago Andrés Vela Aguilera (Todos los miembros del equipo realizan parte de esto)**

* Encargado de la recopilación, gestión y análisis de datos relevantes para los ODS.
* Desarrollo de enfoques analíticos y técnicas para extraer información valiosa de los datos.
* Garantizar la calidad y la integridad de los datos utilizados en el proyecto.

**Líder de Analítica: Josue Mauricio Rivera Ladino (Todos los miembros del equipo realizan parte de esto)**

* Encargado de aplicar técnicas analíticas avanzadas para obtener conocimientos y generar recomendaciones basadas en los datos recopilados.
* Desarrollo de modelos de aprendizaje automático o estadísticas, si es necesario para cumplir con los objetivos del proyecto.

**Reunión de Lanzamiento y Planeación: Todos los miembros del equipo deben participar en esta reunión.**

* Definición de los objetivos del proyecto y los roles y responsabilidades de cada miembro.
* Establecimiento de un plan de acción y un calendario para el proyecto.

**Reunión de Ideación: Todos los miembros del equipo pueden aportar ideas y soluciones.**

* Generación de ideas creativas y estrategias para abordar los ODS específicos.
* Discusión y evaluación de enfoques innovadores para el proyecto.

**Reuniones de Seguimiento: Todos los miembros del equipo, con un enfoque en los líderes de cada área.**

* Evaluación del progreso del proyecto.
* Identificación de posibles desafíos y ajustes necesarios en tiempo real.

**Reunión de Finalización: Todos los miembros del equipo deben estar presentes.**

* Evaluación final de los resultados del proyecto y su impacto en relación con los ODS.
* Documentación de lecciones aprendidas y recomendaciones para futuros proyectos similares.

## *Referencias*

## *Información acerca de árbol de decisión:*

*¿Qué es un árbol de decisión? | IBM*. (s. f.). <https://www.ibm.com/es-es/topics/decision-trees>

## *Información acerca de Random Forest:*

Team, D. (2022, 1 agosto). *Random Forest: Bosque aleatorio. Definición y funcionamiento*. Formation Data Science | DataScientest.com. <https://datascientest.com/es/random-forest-bosque-aleatorio-definicion-y-funcionamiento>

## *Información acerca de KNN:*

Srivastava, T. (2023). A complete guide to K-Nearest neighbors (Updated 2023). *Analytics Vidhya*. <https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/>