Universidad de los Andes Colombia **Universidad de los Andes** **Inteligencia de negocios**

# Proyecto 1

## Etapa 1

**Samuel Freire – 202111460**

**Juan Felipe Garcia – 202014961**

**Lucciano Franco Márquez – 202111458**

**Contenido**

## 1. Entendimiento del negocio y enfoque analítico

| | |
|---|---|
| **Oportunidad/problema Negocio** | Antes de empezar a hablar de algo relacionado con código o transformaciones, es necesario hablar de cuál es el alcance que se espera en esta construcción. Ahora bien, más adelante se van a tratar temas más específicos, con respecto a la parte programática. Sin embargo, no se puede empezar a realizar un análisis, ni transformaciones para el proyecto, sin antes definir las bases y límites de este. En este sentido, tiene relevancia hablar de cuáles serían los objetivos del proyecto. Para este caso, se espera que el proyecto, desde el punto de vista del negocio sea capaz de realizar la automatización, no en esta etapa, de lograr identificar opiniones o textos y clasificarlos en los objetivos de desarrollo sostenible. Luego, otros objetivos seria que el proyecto sea útil para realizar y entender estrategias de negocio para mejorar de acuerdo con estos objetivos de negocio. Además, de el objetivo de poder utilizar e implementar el machine learning en este rubro, de tal manera, que se vuelva un proceso mucho más rápido, y efectivo, que el analizarlo a mano. Ahora bien, hasta ahora hemos definido objetivos directamente relacionado scon el uso analítico del proyecto, es decir, con la parte de la empresa. Sin embargo, no hay que olvidar que este es un proyecto de las naciones unidas, por ende, se espera tenga un sentido social. En este mismo hilo, es necesario decir, que el objetivo más social de este proyecto, es poder escuchar a las personas sobre problemáticas, y poder generar recomendaciones al respecto. Esto quiere decir, que, al analizar las opiniones, con el prototipo a desarrollar, se espera que la retroalimentación de las personas sea más escuchada, y no solo esto, sino que sea más organizado para poder implementar o evaluar propuestas según las recomendaciones y/u opiniones de las personas. Ahora bien, luego de haber realizado este primer limitante o fin, que se espera con el proyecto, parecería lógico pasar a la definición de esos parámetros o métricas que nos van a permitir definir si el proyecto estuviera correcto no. Sin embargo, antes de pasar a eso, sería necesario realizar una pequeña contextualización de que se va a estudiar, de cierta manera. Además de realizar un análisis del impacto que podría llegar a generar el proyecto. En un primer momento, la parte fundamental de la tarea que se fue encargada al proyecto fue la clasificación de las opiniones o textos a unos objetivos de desarrollo sostenible, como se había dicho antes. Sin embargo, hasta ahora no habíamos definido que eran los objetivos de desarrollo sostenible. Estos objetivos hacen parte de una agenda de las naciones unidas (ONU). Los cuales permiten definir un horizonte para reducir o mejorar problemas con respecto a la pobreza, el acceso a la salud y educación, la igualdad de género y los problemas ambientales, entre otros. Dentro de estos objetivos se pueden encontrar unas metas, que no se van a entrar a revisar a profundidad en este estudio. Ahora bien, para el caso específico del grupo, le fueron asignados unos |

| | |
|---|---|
| | objetivos específicos para estudiar. Dentro de estos objetivos encontramos los objetivos 3,4,5 de la agenda de la ONU. Específicamente estos objetivos representan de cierta manera unas problemáticas a solucionar. En el caso del objetivo 3 se encuentra la garantía de un bienestar óptimo para todos. Luego, para el caso del objetivo 4 representa la garantía de una educación de calidad en los diversos niveles. Y, por último, el objetivo 5 representa la problemática o la garantía de la igualdad de género. Ahora, luego de haber definido todos los límites, de haber definido a un nivel de profundidad mayor los objetivos específicos asignados al grupo será necesario definir el impacto se espera tenga este proyecto en la solución o análisis de los problemas de la ONU, y, además, específicamente en Colombia. Para este punto, como ya se había mencionado antes, más allá de la parte programática, el objetivo de este proyecto es que las voces de las personas sean escuchadas, y no solo escuchadas, sino que se puede llegar a analizar hasta un nivel que permite de la misma manera generara estrategias pares a lograr apoyar esas problemáticas mencionadas. Para el caso de Colombia, país que cumple o sufre de varias de esta problemática, se espera se genere un apoyo para que la actuación o la entrada en solución de alguna de estas problemáticas, primero sea lo antes posible. Además, de que se espera ayude a que sea lo más lógica posible, ya que se van a escuchar y analizar las opiniones de las personas que sufren o que han visto alguna de estas problemáticas. Para finalizar esta primera introducción al problema se van a definir esas métricas que van a ser que el proyecto sea correcto no. En este punto, no parece necesario hablar de métricas, tan específicas como lo podría ser una recall o una precisión, sino, más bien, hablar de que se espera que haga el proyecto para que su función o su ejecución sea satisfactoria. En este punto, se cree que el pilar de decisión del proyecto es si puede o clasificar los textos en los objetivos de desarrollo sostenible debidos. Siendo, este el punto de inflexión de definir si el proyecto en un 80% logra cumplir con su clasificación, se consideraría que el proyecto logro a cabalidad el objetivo, y luego ya sería genera las propuestas acordes a los análisis realizados. |
| **Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático) e incluya las técnicas y algoritmos que propone utilizar.** | Para este caso al hablar del enfoque analítico, que va a conllevar este estudio, va a ser necesario revisar los conceptos relevantes, para lograr definir esta etapa. En un primer lugar, resulta relevante hablar de que tipo de aprendizaje se definió para el estudio. En este caso, se definió como un tipo de aprendizaje supervisado. Esto se debe a que existen etiquetas que definen las características que se quieren definir como los objetivos de desarrollo sostenible al cual va asignado cada texto. Por esta razón, irse por el lado supervisado resulta lo más conveniente. Luego, para hablar de la tarea de aprendizaje, que se va a utilizar, va muy ligado al tipo que se mencionó antes. En este caso, se espera que la tarea de aprendizaje sea de clasificación. Esto se debe a que la clasificación va a permitir, |

| | |
|---|---|
| | dividir o partir los datos entre los tres objetivos de desarrollo sostenible que se asignaron. Ahora luego de haber definido ambas características principales para lograr entender la parte teórica del problema y, específicamente, de su solución, vamos ahora a hablar de los algoritmos a utilizar. Para este caso, se van a utilizar tres algoritmos. En un primer momento, se quiere utilizar el algoritmo de árbol de decisión, específicamente, con el algoritmo de random forest. Esto se deba que es un algoritmo, de fácil implementación. También, permite generar una visión más clara del modelo para poder explicarlo de cierta manera al negocio. Luego, es un modelo que permite entender y clasificar de manera muy metódica los datos que se tienen. Por otro lado, también se quiere utilizar el algoritmo Regresión logística. Esto por razones muy parecidas a las anteriores. Esto quiere decir, que es un algoritmo fácil de implementar, que no requiere de mayor profundidad para su ejecución. Sin embargo, se escogió por su capacidad de clasificación diferente a la anterior, esto quiere decir, que este algoritmo es capaz de realizar clasificaciones un poco más atípicas, si se quiere decir entre los datos. Además, de lograr reconocer patrones no lineales en su ejecución. Por lo que sería una gran comparación con respecto al otro algoritmo para ver si se llega a lo mismo o, si por el contrario genera algo destinito. Por último, el otro algoritmo que se escogió fue el de SMV. Esto porque, se quería una forma totalmente distinta de hacer los cálculos o las clasificaciones. Este al ser un método basado en probabilidad, puede llegar a ser interesante, ver un punto diferente, de realizar los cálculos y poder analizar cuan diferente es el resultado de los algoritmos. Además, de su fácil implementación y la posibilidad que existe de poder realizarlo con las librerías que se utilizan para los otros dos algoritmos. Ahora bien, luego de haber definido los detalles de la propuesta realizar, es necesario definir por qué esta propuesta es útil, para la empresa. En este caso, se ofrece esta propuesta con la idea de que se va a proponer un modelo que va a permitir con un alto porcentaje clasificar a las opiniones en los objetivos requeridos. Además, se considera esta propuesta ventajas, ya que le va a permitir a la organización entender de manera muy visual lo propuesto. |
| **Organización y rol dentro de ella que se beneficia con la oportunidad definida** | Para este punto, se entiende que al solo existir o más bien, al tener como fuente una información proporcionada por la ONU. Además, de estar explicita la contratación y propuesta del proyecto hacia el respectivo grupo. Se entiende que la organización beneficiada con este proyecto, en un primer momento, sería la ONU. Sin embargo, otra organización, que se podría beneficiar, en un segundo nivel, serían las personas, específicamente de Colombia, y las cuales tengan que ver con las problemáticas estudiadas. Ahora bien, al hablar del caso específico, la UNFPA sería la organización directamente beneficiada con este proyecto, dado que es la contratante de este. Además, de que a partir de este puede genera |

| | |
|---|---|
| | o evaluar iniciativas que surjan de este mismo. Ahora, al hablar de los roles que se benefician pueden ser las víctimas, de estas problemáticas. Luego, los jefes de propuestas, o jefes de objetivos relacionados, entre otros...,|
| **Contacto con experto externo al proyecto** | Para esta primera etapa, no se plena o no se espera algún tipo de muestra de los resultados o de la parte programática si se quiere pensar. Sin embargo, ya para la primera semana posterior a la entrega, se espera tener una reunión ya pactada con la experta (Maria Paula Paba Torrez, mp.paba@uniandes.edu.co). Esto con la finalidad de poder mostrar los resultados encontrados, el proceso a seguir, y pues claramente poder recibir el feedback e una persona ligada o relacionada con el tema. Esto para en la siguiente etapa poder mejorar y tomar en cuenta estas recomendaciones para modificar de manera óptima el proceso a seguir. |

## 2. Entendimiento del negocio y preparación de datos

Para empezar con el análisis de estos datos, se va a empezar con un análisis menos detallado de los datos. Como podemos ver en el dataframe, resultante del analisis de textos, encontramos muy pocas columnas. En un primer momento, se puede ver el id organizador. Luego, encontramos el pilar de este análisis, que es la columna de opiniones de las personas sobre los objetivos de desarrollo. Por último, encontramos la columna, que define a que objetivo de desarrollo sostenible hace referencia la opinión añadida en la columna anterior. Para el caso de la única columna numérica, realizamos un análisis numérico. Sin embargo, es necesario recordar que esta es una columna numérica, de tipo categórico. Esto quiere decir, que realmente no representa números como tal, sino categorías. Para el caso de proyecto estos números van a ser tomados como los objetivos de desarrollo sostenible asignado a la información. Para el análisis se encontraron 3000 registros. Con una media de 4. Luego una desviación de 0.81. Luego, un valor mínimo de 3 (que es el objetivo de desarrollo 3). Y, por último, un valor máximo de 5 (otro objetivo de desarrollo).

Luego, para seguir con el análisis, se a hablar de cada parámetro de la calidad de datos. Para empezar, se va a decir cómo se está de completitud en los datos. Como podemos ver no existe ningún valor nulo, en los datos que fueron presentados por lo que no debería tomarse ninguna decisión sobre si o si no se deben tomar en cuenta, puesto que no hay variedad. Luego, al hablar de unicidad se encontró lo siguiente. Como se puede ver no existen valores duplicados. Esto es que la información tiene muy buena cantidad de variedad para evaluar. Po resta razón, no es necesario quitar los valores duplicados. Luego, al hablar de consistencia, es necesario decir que va a suceder lo mismo anterior. En este caso, se puede ver que todos los valores cumplen con los rangos dentro de los rangos predefinidos por los datos. Por ende, se dice que están bien definidos los rangos. Luego, hablando del último de parámetros, es del tipo de validez. En este punto, no es necesario analizar, esta opción. Esto se debe a que los datos han cumplidos con todos los formatos. Además, que las columnas no tienen ningún tipo de relación entre sí por lo que no se puede decir que incumplen los valores.

Ahora al hablar de la preparación de datos, al no tener muchas opciones, relacionada con la calidad de datos, es necesario realizar transformaciones con respecto al objetivo del análisis. En este punto, se definen funciones o transformaciones necesarias para lograr analizar el texto. En un primer momento se va a realizar una transformación relacionada a ASCII. En esta transformación, lo que se busca es quitar esos símbolos o caracteres que no estén en un formato ASCII, o algo conocido a ASCII. Luego, para evitar comparación de las mismas palabras, pero unas con mayúsculas y las otras

en minúsculas, lo que ese va a realizar es una transformación que busca que todo quede en minúsculas, para que sea más fácil el análisis. Luego, lo que se busca es retirar los puntos, comas, punto y comas. Esto para que el texto quede lo más puro posible, si se quiere pensar para que el análisis sea solo del contenido. Luego, se necesita que los números desaparezcan, en este caso se busca que los números sean cambiados por palabra, para que no afecte el análisis. Luego, se trae un término que se habia mencionado antes, que es stopwords. En este caso, lo que se busca es que quitar estas palabras que son recurrentes pero que no aportan en el análisis y lo pueden hacer muy densos. Luego, se realizan esas transformaciones para empezar con el análisis del texto. En este punto, Al hablar de transformaciones, relacionadas con los algoritmos, se van a dividir los datos, de tal manera que sea posible trabajarlos en los algoritmos. En este punto, se van a token izar, y trabajar de diversas maneras los datos, para poder utilizar los algoritmos, que más adelante se van a mencionar. Por lo que, ya se va a pasar a mencionar las decisiones de modelo y evaluaciones a realizar.

## 3. Modelado y evaluación

Para empezar con esta sección se vana realizar diversas evaluaciones, de diversos algoritmos tanto de clasificación como de vectorización. Sin embargo, parece necesario realizar una descripción de todos los algoritmos que se vana a utilizar antes de empezar a realizar su evaluación. Para esto se va a empezar con los algoritmos de vectorización. Y seguido se van a describir los algoritmos de clasificación usados.

El primer algoritmo de vectorización a describir es el algoritmo BOW, o bag of words. Este algoritmo realiza la vectorización, o división del texto, de acuerdo con la ocurrencia de las palabras. Este modelo permite tener una representación de tipo tabla, ligando las palabras con su aparición o no en el texto, y realizando un análisis de frecuencias (IBM, 2023). Esto resultado en que, si la entrada es un texto, la salida corresponde a la matriz de frecuencia, relacionando los textos con la frecuencia de las palabras en ellos. Es decir, como se ve, en el notebook adjunto, devuelve una importancia de palabras dada su frecuencia. El segundo algoritmo que se va a querer describir es el algoritmo llamado TF-IDF. En este caso, resulta un poco mas complejo que el anterior. Para este algoritmo se realiza o realiza por detrás un conteo de palabras y con cuanta frecuencia aparece una palabra en el texto analizado. Luego, relaciona esas palabras y su frecuencia, con las frecuencias que han tenido en otros textos. En sentido, de ahí viene el nombre, frecuencia de termino-frecuencia de documento invertida (Linkedin, 2022). En este sentido, ayuda a definir escritura, de manera que se dé importancia a aquellas palabras que no sean repetitivas en el lenguaje, sino que busca darle o identificar mayormente esas palabras únicas para ese rubro o texto. Es decir, que la entrada va a ser un texto, su salida seria la matriz que relaciona los textos con las palabras del vocabulario y su intersección serían los puntajes so importancia de estos. El tercer algoritmo para analizar es el algoritmo Doc2Vec. Este algoritmo, resulta un poco en la idea de similitud entre textos. En sentido, el algoritmo realiza es una extracción o vectorización más de significado en las palabras. Esto quiere decir, que se va a lograr relacionar o vectorizar textos similares de la misma manera, generando así una definición del texto entrada, en textos similares. Esto sirve, para definir si un texto es bueno, será parecido de cierta manera un texto que es bueno, realizando este proceso para malos también y para cosas en común (Nayak, 2019). Para el proyecto, por ejemplo, podría ser que textos de un objetivo 3 deben ser parecidos a textos del objetivo 3. En este sentido, si la entrada es un textos o grupo de textos, la salida sería los vectores e los mismos, agrupados o ubicados cercanos a otros similares con el contexto. Ahora bien, el ultimo algoritmo de vectorización utilizado fue el algoritmo fue el algoritmo GloVe. Este algoritmo utiliza la idea de recuento global de las palabras para lograr realizar una similitud de textos. La idea de palabras que aparecen en textos similares puede resultar en que su ocurrencia similar significa que mantiene una relación estrecha en su uso (Villegas, 2008). Por lo que se podría considerar que es muy parecido al anterior, algoritmo, sin embargo, difieren como se dijo antes, en que glove tiene un contexto mucho

mas abierto que lo podría ser el de Doc2Vec. En este sentido, si la entrada es un grupo de textos, su salida seria vectores de palabras, cercanos entre si según el contexto.

Luego, de haber descrito los algoritmos de vectorización, se va a pasar describir los algoritmos clasificación, que fueron tres los utilizados. El primer algoritmo utilizado fue Random Forest. Este algoritmo, como su nombre lo indica, menciona un bosque aleatorio. Esto va muy ligado a su paso a paso, ya que es basada en los arboles de decisión. Esto quiere decir, que el random forest, genera una variedad de árboles decisión, todos sin una parte de la información base. Luego, realiza una validación cruzada entre todos los árboles que género (IBM, 2023). Para de esta manera llegar al árbol final o al árbol resultado que es compuesto por la validación de todos los anteriores. Por ende, su salida va a resultar en las hojas, y un recorrido de caracterización de cada una de su clasificación. El segundo algoritmo utilizado fue la regresión logística. Específicamente el tipo que se va a utilizar es el tipo nominal. En este caso, el algoritmo, se basa en probabilidades y ecuaciones, para lograr calcular la probabilidad en la que un dato está en grupo o en otro (AWS, 2023). Identificando la relación entre conjunto matemática, y probabilística, para que, a la hora de clasificar un nuevo entrante, se base en la colinealidad entere si para lograr desarrollar un conjunto correcto. Por último, el tercer algoritmo de clasificación utilizado fue SVM. Sus siglas hacen referencia a máquina de vectores de soporte. Su principal característica, es que busca generan de alguna manera un esquema que mejor segrega las clases. Esto significa que busca que los cercanos estén muy cerca y los alejados lo más alejados posible. Es decir, busca el hiperplano, es la dimensión que mejor divida las clases, generando así la clasificación de mejor manera (aprendeia, 2023).

Ahora bien, luego de haber descrito cada uno de los algoritmos, tanto de clasificación como de vectorización que se utilizaron va a ser necesario realizar la presentación de los resultados de cada uno de los modelos que se generaron, para presentar el modelo y su justificación. Para este punto, se realizaron por cada uno de los cuatro algoritmos de vectorización, se realizaron para cada uno los tres algoritmos de clasificación. Para cada uno de los 12 modelos generados se calculó, para la parte analítica, su precisión, el recall y el f1 score.

Luego, de haber aclarado esto, se va a empezar con la descripción de cada uno de los modelos generados. El primer modelo generado fue el de Bow con random Forest. En sentido presento una precisión de 96%, un recall de 96.88% y un f1 de 96.88%. Esto quiere decir, que el error es menor a 4%, por lo que es un modelo que clasifica correctamente, con muy bajo porcentaje de mala clasificación. El segundo modelo, que utiliza la técnica Bag of Words (BoW) combinada con regresión logística, demostró un rendimiento superior. En concreto, consigue una precisión, recuperación y puntuación F1 del 97%. Esto significa que el error es inferior al 3%, lo que significa que el modelo de clasificación es muy preciso y tiene una tasa de clasificación errónea muy baja. El tercer modelo, que utiliza la técnica Bag of Words (BoW) combinada con SVM. En concreto, consigue una precisión, recuperación y puntuación F1 del 96.88%. Esto significa que el error es inferior al 4%, lo que significa que el modelo de clasificación es muy preciso y tiene una tasa de clasificación errónea muy baja. Sin embargo, no llega a superar al anterior. El cuarto modelo generado fue el de TF-IDF con random Forest. En sentido presento una precisión de 97.4%, un recall de 97.4% y un f1 de 97.4%. Esto quiere decir, que el error es menor a 3%, por lo que es un modelo que clasifica correctamente, con muy bajo porcentaje de mala clasificación. Este cuarto modelo, hasta ahora resulta ser el que mayor porcentaje de acierto tiene. El quinto modelo, que utiliza la técnica TF-IDF combinada con regresión logística, demostró un rendimiento superior. En concreto, consigue una precisión, recuperación y puntuación F1 del 97.55%. Esto significa que el error es inferior al 3%, lo que significa que el modelo de clasificación es muy preciso y tiene una tasa de clasificación errónea muy baja. En este sentido, se convierte en el mejor modelo probado hasta ahora. El sexto modelo generado fue el de TD-IDF con SVM. En sentido presento una precisión de 97.5%, un recall de 97.5% y un f1 de 97.5%. Esto quiere decir, que el error

es menor a 3%, por lo que es un modelo que clasifica correctamente, con muy bajo porcentaje de mala clasificación. Presentado un resultado igual al anterior. El séptimo modelo, que utiliza la técnica Doc2Vec combinada con Random Forest. En concreto, consigue una precisión, recuperación y puntuación F1 del 91%. Esto significa que el error es inferior al 9%, lo que significa que el modelo de clasificación es muy preciso y tiene una tasa de clasificación errónea muy baja. Siendo este modelo, el, pero hasta ahora. El octavo modelo generado fue el de Doc2Vec con regresión logística. En sentido presento una precisión de 92%, un recall de 92% y un f1 de 92%. Esto quiere decir, que el error es menor a 8%, por lo que es un modelo que clasifica correctamente, con muy bajo porcentaje de mala clasificación. El noveno modelo, que utiliza la técnica Doc2Vec combinada con SVM. En concreto, consigue una precisión, recuperación y puntuación F1 del 91%. Esto significa que el error es inferior al 9%, lo que significa que el modelo de clasificación es muy preciso y tiene una tasa de clasificación errónea muy baja. El décimo modelo generado fue el de GloVe con Random Forest. En sentido presento una precisión de 81%, un recall de 81% y un f1 de 81%. Esto quiere decir, que el error es menor a 8%, por lo que es un modelo que clasifica correctamente, con muy bajo porcentaje de mala clasificación. Siendo este modelo él, pero modelo hasta ahora, por sus bajos estándares. El undécimo modelo, que utiliza la técnica GloVe combinada con regresión logística. En concreto, consigue una precisión, recuperación y puntuación F1 del 96%. Esto significa que el error es inferior al 4%, lo que significa que el modelo de clasificación es muy preciso y tiene una tasa de clasificación errónea muy baja. El doceavo modelo generado fue el de GloVe con SVM. En sentido presento una precisión de 95%, un recall de 95% y un f1 de 95%. Esto quiere decir, que el error es menor a 5%, por lo que es un modelo que clasifica correctamente, con muy bajo porcentaje de mala clasificación. Siendo este modelo él, pero modelo hasta ahora, por sus bajos estándares. Por ende, luego de haber realizado la comparación de resultados, presentada anteriormente, donde los porcentajes de precisión representan el porcentaje de los que clasifico correctamente frente a los que no. Luego el recall, representa el porcentaje de los que clasifico bien dentro de los que clasifico en las clases. Por último, el f1 representa la unión entre la precisión y el recall. Luego, e haber aclarado esto se puede decir que el modelo que se usó el algoritmo de vectorización TD-IDF, y el algoritmo de clasificación SVM. Esto puede deberse, que al ambos tener bases de similitud entre los grupos que generen, pueden que al ambos tener el objetivo de agrupación más claro, resulte en una clasificación con un porcentaje de error tan bajo.

Así que, luego de haber revisado todos los porcentajes, y haber encontrado el mejor cuantitativamente de los modelos. Se puede decir que el modelo escogió fue el de TD-IDF, como algoritmo de clasificación, y SVM como el algoritmo de clasificación. Esto se debe, mirándolo desde un apartado netamente cuantitativo, porque es el modelo que mayor, f1, precisión y recall tiene. Esto quiere decir, que es el modelo que logra clasificar en cada clase, los que son de esa clase, con menor porcentaje de erro de clasificarlos en otras. Además, de que tiene una partición equitativa e este porcentaje, dentro del resto de clases. Por ende, el f1 va a ser muy alto por sus medidas altas de precisión y recall. Lo que nos indica que netamente cuantitativa el modelo que se debe escoger es este, que tiene un porcentaje de erro r muy bajo menor al 3%.

## 4. Resultados

Una posible estrategia sería implementar este modelo en tareas de clasificación que requieran una alta precisión, como la detección de fraudes, la categorización de contenido o cualquier aplicación donde la exactitud sea esencial. Esto se traduciría en una toma de decisiones más sólida y una mejora significativa en la calidad de las predicciones.

Además de la implementación en casos de alta precisión, la organización podría considerar la expansión de este modelo para automatizar tareas de clasificación en otros departamentos o áreas, como la atención al cliente, la gestión de inventario o la optimización de procesos internos. Esto permitiría aprovechar la capacidad predictiva del modelo en múltiples aspectos del negocio.

Otra estrategia importante es la creación de un equipo interdisciplinario que incluya tanto a científicos de datos como a expertos en dominios específicos de la organización. Este equipo podría colaborar en la mejora continua del modelo y en la identificación de nuevas oportunidades de aplicación, maximizando así su valor.

Además, consideramos fundamental establecer un umbral de confianza para las predicciones y realizar un monitoreo continuo del modelo para detectar posibles desviaciones. Asimismo, la capacitación y educación del personal en el uso del modelo es esencial para garantizar su correcta implementación. La experimentación continua con diferentes configuraciones puede ayudar a optimizar aún más el rendimiento del modelo.

Por otro lado, los resultados de los otros enfoques de vectorización y algoritmos de clasificación también son valiosos para la organización. Cada uno de ellos tiene su lugar en situaciones específicas donde se busque un equilibrio entre precisión y eficiencia. La elección del enfoque y algoritmo adecuados puede marcar la diferencia en la eficiencia y efectividad de las aplicaciones y sistemas de la organización, lo que se traduce en un mejor cumplimiento de los objetivos del negocio.

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

| **Rol dentro de la empresa** | **Tipo de actor** | **Beneficio** | **Riesgo** |
|---|---|---|---|
| UNFPA | Cliente y usuario | Generación de iniciativas coherentes y efectivas sobre las problemáticas encontradas | Puede generar estrategias que no tengan un sentido lógico, puede estar clasificando problemas en áreas o en objetivos que no son. |
| Personas (victimas) | Beneficiado | Recibe soluciones a tiempo, adaptadas a su contexto | Pérdida de confianza en la institución, o en cualquier organismo publico |
| ONU | Cliente y usuario | Apoyar labores sociales preventivas de problemáticas | Puede generar una confusión, en cuanto, a como está el estado actual, como van las iniciativas generadas… |

| | | | |
|---|---|---|---|
| Departamento de servicios de tecnologías de UNFPA | Proveedor | Garantiza la generación de un sistema apto para la toma de decisión y evaluación de iniciativas | Generación incorrecta de análisis y resultados erróneos. Además, de la afectación sobre los datos |
| Departamento de analítica UNFPA | Financiador | Automatización de la toma de decisiones, con base en opiniones de las víctimas. | Puede ser un proyecto fallido, en el cual se pierde dinero y recursos para no generar el resultado esperado. |
| Entidades publicas | Cliente y usuario | Generación de pautas para generar soluciones coherentes y efectivas | Pueden no estar al tanto del manejo de información, amenaza de datos, y falta de entendimiento con respecto al formato del proyecto y su objetivo |

## 6. Trabajo en equipo

| **Integrante del Grupo** | **Rol** | **Tareas Realizadas** | **Tiempo Dedicado (horas)** | **Algoritmo Trabajado** | **Retos Enfrentados** | **Formas de Resolver los Retos** |
|---|---|---|---|---|---|---|
| Samuel Freire | Líder de Proyecto | Gestión y coordinación del proyecto. | 15 | Modelos de Clasificación en Aprendizaje Automático<br><br>(Regresión logística) | Coordinación de horarios y actividades del equipo. | Establecimiento de horarios flexibles para reuniones y comunicación regular. |
| | | Definición de fechas de reuniones y entregables. | | | Toma de decisiones en situaciones de desacuerdo. | Facilitar espacios para la expresión de opiniones y consensuar decisiones cruciales. |
| Felipe García | Líder de Datos | Gestión y manipulación de los conjuntos de datos del proyecto. | 20 | Preprocesamiento de Datos y Extracción de Características<br><br>(Random forest o árboles de decisión) | Limpieza y estructuración de datos complejos. | Utilización de técnicas avanzadas de limpieza y transformación de datos. |
| | | Asignación de tareas relacionadas | | | Garantizar la calidad y coherencia | Implementación de validaciones y controles de |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | con la preparación y procesamiento de datos. | | | de los datos para el análisis. | calidad en los procesos de datos. |
| Lucciano Franco | Líder de Analítica | Supervisión de las actividades analíticas del grupo. | 18 | Modelos de Clasificación y Evaluación de Resultados<br><br>(Máquinas de vectores de soporte (SVM)) | Elección del modelo más adecuado para el problema. | Implementación de técnicas de validación cruzada y comparación de modelos. |
| | | Verificación del cumplimiento de estándares de análisis de resultados. | | | Hay que asegurar que los resultados sean interpretables y aplicables. | Comunicación efectiva de hallazgos y recomendaciones basadas en los resultados. |
| | | Selección del mejor modelo considerando las restricciones. | | | | |

**Reuniones:**

- Reunión de Ideación: Definición de la organización/empresa/institución beneficiada y sus roles.
- Reuniones de Seguimiento: Se realizan reuniones semanales o comunicación vía correo para actualizar avances y tareas a través de herramientas como Trello.
- Reunión de Finalización: Se lleva a cabo para consolidar el trabajo final, revisar el desempeño del grupo y analizar mejoras para la siguiente etapa del proyecto.
- Asignación de Puntos: Samuel Freire: 33 puntos Felipe García: 37 puntos Luciano Franco: 32 puntos
- Puntos Para Mejorar: Reforzar la comunicación para garantizar una mayor sincronización entre las áreas de datos y analítica. Establecer métricas más específicas para evaluar el desempeño individual en el proyecto.

## 7. Referencias

IBM documentation. (s. f.). https://www.ibm.com/docs/es/rpa/21.0?topic=classification-text-algorithms

G, A. (s. f.). El algoritmo TF-IDF y las preferencias políticas. *es.linkedin.com*. https://es.linkedin.com/pulse/el-algoritmo-tf-idf-y-las-preferencias-pol%C3%ADticas-alejandro-gregori

Nayak, M. (2021, 10 diciembre). An intuitive introduction to Document Vector(Doc2VEC). *Medium*. https://pub.towardsai.net/an-intuitive-introduction-of-document-vector-doc2vec-42c6205ca5a2

Carlos, L., & Villegas, G. O. (s/f). *UNIVERSIDAD MAYOR DE SAN ANDRÉS*. Umsa.bo. Recuperado el 15 de octubre de 2023, de https://repositorio.umsa.bo/bitstream/handle/123456789/29142/TM-3823.pdf?sequence=1&isAllowed=y

*What is Random Forest?* | *IBM*. (s. f.). https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.

*¿Qué es la regresión logística? - Explicación del modelo de regresión Logística - AWS*. (s. f.). Amazon Web Services, Inc. https://aws.amazon.com/es/what-is/logistic-regression/

Gonzalez, L. (2020). Máquinas vectores de soporte clasificación – teoría. 🤖 *Aprende IA*. https://aprendeia.com/maquinas-vectores-de-soporte-clasificacion-teoria /