PROYECTO 1: ANALISIS DE ODS DE TEXTOS PARA LAS NACIONES UNIDAS

Grupo 11

Integrantes

Nicolas Londoño Cuellar Tales Alejandro Losada Aristizábal Juan Diego Castellanos Bonilla

CONTENIDO

INTRODUCCIÓN

El proyecto es de analítica de ODS para la ONU, usando información textual (recopilaciones), donde se quiere clasificar de los 17 objetivos de desarrollo sostenible de la ONU los que son de la categoría: salud y bienestar (3), Educación de calidad (4) e igualdad de género (5). De manera que con el modelo en construcción se pueda facilitar la clasificación de testimonios. Para poder hacer un análisis automatizado de opiniones que representan la voz de los habitantes sobre problemáticas de su entorno particular.

Este modelo podría llegar a tener un beneficio en las naciones unidas y en el país. Pues la ventaja está en que rápidamente se podría llegar a una idea de cómo son las problemáticas y cuáles son las problemáticas de su entorno para tomar decisiones sobre como se puede abordar este problema en el entorno colombiano.

(10%) ENTENDIMIENTO DEL NEGOCIO Y ENFOQUE ANALÍTICO.

Oportunidad/problema Negocio	usando información textual (recopilaciones), donde se quiere clasificar de los 17 objetivos de desarrollo sostenible de la ONU los que son de la categoría: salud y bienestar (3), Educación de calidad (4) e igualdad de género (5). Construir un modelo que clasifique la información textual.
Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático) e incluya las técnicas y algoritmos que propone utilizar.	Teniendo en cuenta que los valores dados eran (información textual y un valor numérico el cual decía el tipo de ODS que quiere mejorar la ONU) y que el enunciado dice que se debe clasificar el texto. Se decidió implementar técnicas de clasificación supervisada. Con el fin de tener un análisis correcto se decidió que cada estudiante implementara dos algoritmos, dando un total de 6 entre estos esta: SVM, Arboles de decisión, Regresión logística, KNN, Pipelines y Naive Bayes.
Organización y rol dentro de ella que se beneficia con la oportunidad definida	La organización que se beneficia con la oportunidad definida: Construcción de un modelo que clasifique la información textual. Es Las naciones unidas (ONU) y el rol que dentro de ella que se beneficia es la UNFPA.
Contacto con experto externo al proyecto	En cuanto al contacto con un experto al proyecto se tiene un estudiante de estadística. En el caso del grupo 11 es: Laura Jimena Gama Duque y su correo de contacto es l.gama@uniandes.edu.co

(20%) ENTENDIMIENTO Y PREPARACIÓN DE LOS DATOS.

Entendimiento de datos

- 3000 datos
- 1 columna categórica (Textos_español)
- 1 columna numérica (sdg)

Completitud

- Todos los datos están completos

Unicidad

- Todos los datos de Textos_español son únicos
- Los valores de sdg se repiten los cuales son: 3,4,5

Nulos

- No hay datos nulos

Validez

- Todos los datos son validos

Limpieza y tratamiento de datos:

Limpieza de datos:

1. Se eliminó los datos duplicados

Preparación de datos

1. **Se eliminaron todos los caracteres no ASCII**
   el motivo de eliminar estos caracteres no ASCII es que no pertenecen al alfabeto latino.
2. **Todo carácter pasó a estar en minúscula**
   esto se realizó para evitar problemas al momento de trabajar con palabras y que sean diferentes.
3. **Se elimino toda puntuación (./,/;/ (/))**
   no aporta mucha información al momento de usar la raíz de la palabra
4. **Se remplazó todo número a palabras con num2words**
   como se tienen que pasar todos los valores a numérico para los modelos se remplazan todos los numero a texto para evitar errores.
5. **Se cambiaron las palabras vacías (comunes) con stopwords con la librería nltk a espacios.**

quitar las palabras comunes es importante, ya que estas no aportan nada al texto, ni al análisis.

6. **Con countVectorize y TfidfVectorizer se construyó una matriz que mirara cuantas, si aparece una palabra en especifico**
   Esto se realizo con el fin de poder pasar las palabras a una representación numérica y de esta forma poder clasificarlos.

Técnicas y algoritmos

- SVM (máquinas de vectores de soporte)
- Arboles de decisión
- Regresión logística
- KNN
- Pipelines
- Naive Bayes

Técnica: Clasificación

(25%) MODELADO Y EVALUACIÓN.

SVM (máquinas de vectores de soporte)

El algoritmo de máquinas de vectores de soporte es un algoritmo de aprendizaje supervisado, el cual se usó para la clasificación binaria y poder separar de la mejor forma posible dos clases diferentes de puntos de datos.

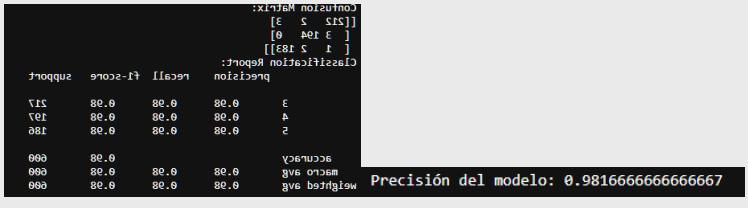
```
Confusion Matrix:
[[213   3   1]
 [  6 191   0]
 [  1   3 182]]
Classification Report:
              precision    recall  f1-score   support

           3       0.97      0.98      0.97       217
           4       0.97      0.97      0.97       197
           5       0.99      0.98      0.99       186

    accuracy                           0.98       600
   macro avg       0.98      0.98      0.98       600
weighted avg       0.98      0.98      0.98       600
```

```
Precisión del modelo: 0.9766666666666667
```

Nuevos vectores:

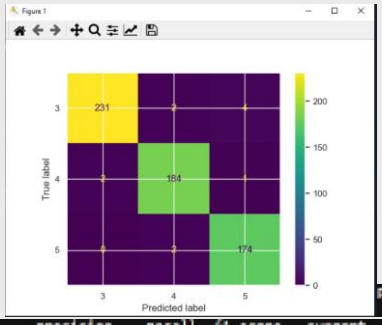


```
Precisión del modelo: 0.9816666666666667
```

Arboles de decisión (se mostrará el mejor resultado)

Los árboles de decisión es un algoritmo de clasificación en un aprendizaje supervisado Los nodos hojas representan los resultados posibles dentro del conjunto de datos, donde hay un numero de

informes de texto y un numero de entropía. Cabe resaltar que entre menor sea dicho número mejor son la clasificación de los textos.




```
Precisión del modelo: 0.9816666666666667
```

	precision	recall	f1-score	support
3	0.99	0.97	0.98	237
4	0.98	0.98	0.98	187
5	0.97	0.99	0.98	176
accuracy			0.98	600
macro avg	0.98	0.98	0.98	600
weighted avg	0.98	0.98	0.98	600

Regresión logística

El modelo de regresión logística para análisis de textos es una técnica de aprendizaje supervisado que se utiliza para predecir la probabilidad de que un texto pertenezca a una o más categorías predefinidas.

```
Precisión: 0.9766666666666667
```

```
Matriz de confusión:
 [[306   4   3]
 [  4 284   3]
 [  3   4 289]]
```

```
Informe de clasificación:
              precision    recall  f1-score   support

           3       0.98      0.98      0.98       313
           4       0.97      0.98      0.97       291
           5       0.98      0.98      0.98       296

    accuracy                           0.98       900
   macro avg       0.98      0.98      0.98       900
weighted avg       0.98      0.98      0.98       900
```

```
print(y_proba_pos[:10])
```

```
[0.84264132 0.92822304 0.79426823 0.10022985 0.01399149 0.99604052
 0.87998217 0.14837287 0.18077024 0.00629769]
```

```
print(y_proba_neg[:10])
```

```
[0.08792974 0.02543949 0.13819742 0.78278829 0.01360102 0.00205898
 0.07461658 0.24324491 0.59857659 0.00777299]
```

KNN

El modelo de KNN para análisis de textos es una técnica de aprendizaje supervisado que se utiliza para clasificar, predecir o buscar elementos similares en función de los vecinos más cercanos.

	precision	recall	f1-score	support
3	0.97	0.98	0.98	217
4	0.98	0.96	0.97	197
5	0.96	0.97	0.96	186
accuracy			0.97	600
macro avg	0.97	0.97	0.97	600
weighted avg	0.97	0.97	0.97	600

```
Precisión del modelo: 0.97
```

Pipelines (SVC)

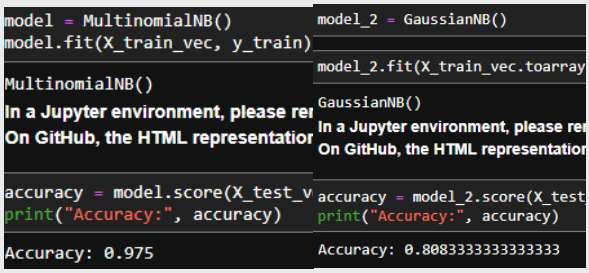
Es un algoritmo de aprendizaje automático utilizado para la clasificación en problemas de aprendizaje supervisado. Sirve para

clasificar de forma Binaria, multiclase, con características no lineales, grandes conjuntos y alta dimensionalidad.

```
Exactitud del modelo: 0.9766666666666667
```

Naive bayes

Naive Bayes es un algoritmo de aprendizaje automático utilizado principalmente para problemas de clasificación en aprendizaje supervisado. Este algoritmo se basa en el teorema de Bayes y asume independencia condicional entre los atributos o características.



Mejor técnica: SVM o Arboles de decisión

Ambos dieron **0.9816**

(15%) RESULTADOS.

En cuanto a los resultados obtenidos entre todos los modelos que se realizaron se tiene en cuenta que todos los algoritmos implementados son de clasificación de texto. De igual manera cabe mencionar que en todos estos se realizo la misma limpieza y el único cambio que se dio en métricas entre el mismo texto fue por el cambio de vectorización con countVectorize y TfidfVectorizer sin embargo al momento de hacer la comparación de la precisión del algoritmo dio como resultado que el modelo SVM y el modelo Arboles de decisión (randomforest) dieron el mismo resultado de precisión el cual fue de 0.9816. Esto nos dice que los datos están en buen estado y que el modelo es bueno para la toma de decisiones, ya que solo tiene un 0,0184 de margen de error. De igual manera al calcular el F1 Score para todos los modelos se pudo observar como los modelos SVM, regresión logística y arboles tuvieron un F1 promedio de 0.98.

Una vez se sabia esto se recomienda a la Organización de las naciones unidas y a la UNFPA el uso de los siguientes modelos: SVM y arboles ya que ambos tuvieron la misma precisión y un F1 Score igual.

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
Agenda de la Organización de las naciones unidas.	Usuario-cliente	Apoya los objetivos de sostenibilidad que se deben desarrollar en los próximos 15 años.	Si el modelo no tiene un buen desempeño, puede estar alertando a los gobiernos, el sector privado y la sociedad civil a hacer todo lo opuesto para cumplir con los ODS 3,4,5
UNFPA	Seguidor	Mecanismo de toma de identificación mediante la evaluación de políticas públicas y su impacto con los datos reconocidos.	En caso de que el modelo no funcione es dinero mal invertido y se pueden identificar de manera incorrecta políticas y/o implementar una solución al problema que este mal clasificado.
Universidad de los andes	Proveedor	Garantiza el cumplimiento de estándares de calidad de los modelos desarrollados, que incluye métricas de los datos utilizados y una explicación para la empresa.	Manejo incorrecto de los datos que lleve a la violación de la privacidad de los datos
UNFPA (gente encargada de los datos)	Beneficiado	Recibe un modelo con los datos clasificados al tiempo que le	Recibir un modelo el cual no sea confiable o no este bien

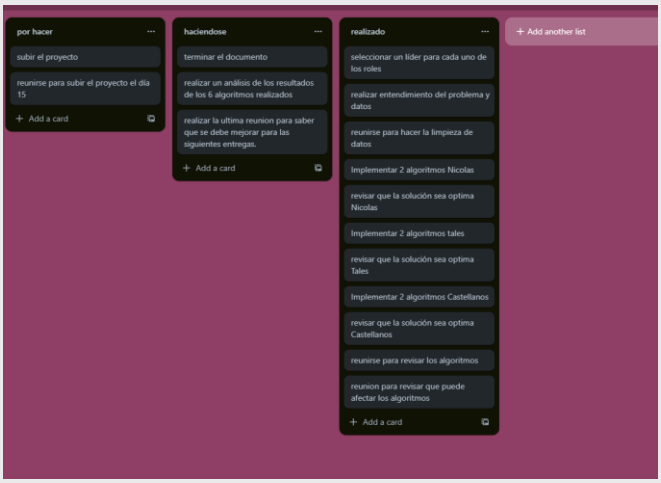
		permite pensar en soluciones para varios problemas en simultaneo para problemas parecidos.	realizado, puede llevar a combinar diferentes ODS los cuales al momento de implementar una solución se encontrara con una peor solución.

(8%) Trabajo en equipo

- **Líder de proyecto: Juan Diego Castellanos Bonilla**
- **Líder de negocio: Nicolas Londoño**
- **Líder de datos: Tales Losada**
- **Líder de analítica: Juan Diego Castellanos Bonilla**

Reuniones:

- **Reunión de lanzamiento y planeación:** se llevó a cabo el lunes 1 de octubre, en el cual se definió el problema y se buscaron algoritmos para la solución del problema.
- **Reunión de ideación:** Se llevo a cabo el lunes 8 de octubre, en la cual se realizó la limpieza de datos de manera grupal, con el fin de que todos lo entendieran y se realizó la introducción del documento, para definir la organización, rol y beneficiario de la solución.
- **Reuniones de seguimiento:** se realizaron 3 reuniones de seguimiento. La primera el día 7 de octubre para saber que todos revisaron los datos y entendieron el problema. El día 11 de octubre para revisar la implementación de los algoritmos y avance del proyecto. La última reunión se hizo el día 13 con el fin de seleccionar y explorar que puede afectar la precisión de cada algoritmo.



- **Reunión de finalización:** Una vez realizada la reunión de terminación del proyecto se verifico el trabajo y se decidió una fecha extra para explicarse entre todos cada uno de los algoritmos y todos conocer lo que se realizo en el trabajo. De igual manera para mejorar se acordó no depender del trabajo de otras personas para continuar el trabajo individual.