Noviembre 2022
Universidad de Los Andes
Jesús David Barrios 201921887
Sergio Peñuela 201922873
Jhoan Diaz 201819861

Proyecto 1 Inteligencia de negocios – Etapa 2

1. **Proceso de automatización del proceso de preparación de datos, construcción del modelo, persistencia del modelo y acceso por medio de API**

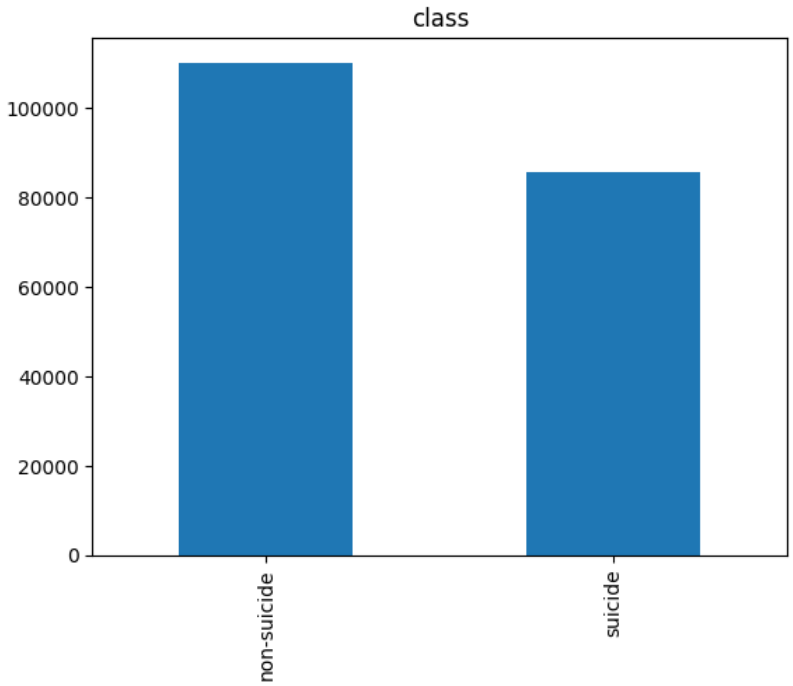
Entendimiento y Preparación de los datos

El proceso de entendimiento y preparación de datos se encuentra implementado y explicado en el notebook entendimiento_procesamientos_datos.ipynb. A continuación, se hace un resumen del procedimiento llevado a cabo.

Inicialmente se descargaron los datos entregados y se obtuvo información básica de estos. A continuación, se muestra la cantidad de datos y un ejemplo de ellos:

```
Número de filas:  195700
Número de columnas:  3
```

	Unnamed: 0	text	class
21738	135566	AAAA I'm literally the stupidest person in exi...	non-suicide
26627	17383	Is it just me or is the sound of rain One of t...	non-suicide
135198	36707	How close i amam tired of living, tired of bei...	suicide
140848	186151	Guess who kissed a girl? Not me.\n\n&#x200...	non-suicide
79670	255320	Should I delete my Reddit account? I mean shou...	non-suicide



Se observa que hay cerca de 200 mil registros, de los cuales alrededor de un 40% hacen referencia a intentos de suicidio. Posteriormente, se midieron dimensiones de calidad en los datos:

- Completitud: Se observa el número de valores nulos en los datos:

```
Número de filas con valores nulos:  0
Número de columnas con valores nulos:  0

Porcentaje de completitud de las columnas: 100.00%
```

  En este caso no fue necesario eliminar ningún dato pues no había valores nulos en el archivo .csv que recibimos.

- Unicidad: Se ve el número de valores repetidos en la base

```
Número de filas duplicadas:  0
Número de filas con indice duplicado:  0
```

  Tampoco fue necesario eliminar datos por unicidad pues no hubo ni un solo dato duplicado.

- 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.

Luego, se realizó la preparación de los datos. Para poder llevar a cabo la ejecución de los diferentes modelos primero tenemos que procesar los textos para que estén limpios y puedan

ser tokenizados facilmente. Lo primero que se hizo fue eliminar todos los caracteres que no pertenecían al alfabeto (e.g. @, %, $, etc.). Después se pasaron todos los textos a minúsculas pues las letras mayúsculas no aportan ninguna información útil al modelo y queremos que todas las palabras estén iguales y no se vayan a diferenciar por una letra en mayúscula.

Posteriormente se hizo la tokenización de los textos y se utilizaron funciones para modificar los tokens de la siguiente manera: se quitaron aquellas palabras que no contuvieran ninguna vocal o "y", pues sin ninguna de estas letras la palabra se considera un typo. Posteriormente se eliminaron las "stopwords". Estas son palabras que no aportan significado por si solas y que se utilizan para cumplir con la sintaxis del lenguaje. Nosotros solo necesitamos palabras que nos aporten valor con respecto al tema que estamos buscando. Pronombres o conectores no aportan respecto al tema del suicidio. Así mismo, se eliminaron los tokens vacíos y aquellos que tuvieran una longitud menor a 2. Además, se utilizaron métodos para lematizar las palabras para obtenerlas de la manera en que se verían en un diccionario, y se eliminaron los prefijos y sufijos (stemming). Todos se pasaron a singular, los verbos se pasaron a infinitivo, etc.

Finalmente se volvieron a unir todos los tokens en un string para poder llevar a cabo el proceso de vectorización para cada modelo. La vectorización de datos fue realizada de manera independiente por cada estudiante en preparación de los datos para los respectivos modelos.

2. Desarrollo de la aplicación y justificación

En este caso para la aplicación se decidió utilizar el modelo que obtuvo los mejores resultados en la etapa anterior del proyecto el cual fue Random Forest.

La aplicación creada puede ser utilizada en varias empresas como una herramienta que ayude al área de recursos humanos para que en caso de que haya sospechas sobre el deterioro psicológico de un trabajador y que se crea que pueda llegar a un posible intento de suicidio, identificarlo cuanto antes y poder tomar las medidas adecuadas. En este caso la aplicación solamente muestra si el mensaje tiene intenciones suicidas o no. Pero también se puede obtener más información que pueda ayudar al psicólogo o aquella persona encargada del bienestar del trabajador ver posibles indicadores y que palabras pueden indicar una mayor posibilidad de un intento de suicidio.

Una aplicación como esta puede ser de gran utilidad para poder identificar problemas que pueda tener la intención de suicidarse. Esto no necesariamente se debe aplicar a los trabajadores de una compañía. Se puede utilizar en colegios, universidades, psicólogos, o cualquier comunidad. Esto puede ayudar a encontrar maneras de lidiar con problemas psicológicos y reducir el número de personas que se suicidan por problemas como estos.

Esto se podría hacer como una aplicación tanto web como móvil. Depende de que tanto alcance se le quiera dar. Se le puede dar acceso a todo el mundo que tenga un móvil y que llenen de información el modelo para obtener mejores predicciones. De esta manera el modelo siempre seguirá creciendo.

3. Resultados

Link del video:

4. Trabajo en equipo

Jesús David Barrios:

El estudiante Jesús Barrios tuvo la labor de ingeniero de software, por lo que se encargó de diseñar e implementar la aplicación. Se encargó de traducir el proceso realizado en la etapa 1 del proyecto en archivos joblib y scripts de Python. Dentro de la aplicación se encargó del API y de la UI.

Sergio Peñuela:

Se encargó de la redacción del documento y tuvo la labor de líder de proyecto.

Jhoan Diaz

Se encargó de la grabación del video y tuvo la labor de ingeniero de datos.

5. Enlace de la wiki