Martin Ubaque
Alejandro Tovar
Gabriela Paez

Proyecto 1 - Salud Mental

Prefacio - Trabajo en grupo

Roles	
Líder de proyecto	Martín Ubaque
Líder de negocio	Diego Alejandro Tovar
Líder de datos	Diego Alejandro Tovar
Líder de analítica	Gabriela Páez

a. Lista de trabajos a realizar

Martin Ubaque
Alejandro Tovar
Gabriela Paez

A. Preparación de los datos
Encargado: Diego Alejandro Tovar
Horas de trabajo: 6.

B. Construcción modelo 1 - Árboles de decisión
Encargado: Martín Ubaque
Horas de trabajo: 6.

C. Construcción modelo 2 - KNN
Encargado: Gabriela Paez
Horas de trabajo: 6.

D. Construcción modelo 3 - Bernoulli NB
Encargado: Diego Alejandro Tovar
Horas de trabajo: 3.

C. Evaluación y resultados
Encargado: Gabriela Paez
Horas de trabajo: 3.

E. Organización del proyecto
Encargado: Martín Ubaque
Horas de trabajo: 3.

1. Comprensión del negocio y enfoque analítico

a. Marco teórico:

La depresión es considerada un trastorno mental, el cual es caracterizado por sentimientos de tristeza y un muy bajo estado de ánimo (Depresión). De acuerdo con la Organización Mundial de la Salud, alrededor del 5% de los adultos en el mundo sufren de este trastorno, lo cual equivale a alrededor de 280 millones de personas. La depresión es un tema serio, ya que usualmente interfiere con la vida cotidiana de las personas, llevándolos a múltiples sentimientos de tristeza, ira o frustración. En el peor de los casos, este trastorno puede llevar a las personas al suicidio, tomandose su propia vida. Actualmente, más de 700,000 personas se suicidan al año, convirtiéndolo en la cuarta causa de muerte en las personas entre 15 a 29 años (Depresión). Por esta razón, se decidió crear un modelo de aprendizaje automático que pueda determinar si una persona está intentando, va a intentar o ha intentado suicidarse con el fin de prevenir este acontecimiento.

Martin Ubaque
Alejandro Tovar
Gabriela Paez

b. Objetivos y criterios de éxito:

Crear un modelo de aprendizaje automático que sea capaz de determinar, a partir de textos planos, si una persona está intentando, va a intentar o ha intentado suicidarse con el fin de prevenir estos acontecimientos y/o brindar la ayuda necesaria a la gente que lo necesita. Para determinar el éxito de este proyecto, se debe garantizar, con pruebas, que la calidad del modelo es satisfactoria a la hora de clasificar dichos textos. En otras palabras, se debe garantizar que, para una gran cantidad de los casos, el modelo es capaz de, leyendo el texto, decidir correctamente si la persona que lo escribió ha intentado suicidarse.

Oportunidad/problema Negocio	Prevención de intentos de suicidio y posibilidad de brindar oportunidades de ayuda en cuanto a la salud mental del público general.		
Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje de máquina)	Crear un modelo capaz de analizar textos planos, convertirlos en datos y utilizar dichos datos para clasificar los textos del público general definiendo si pueden ser propensos a cometer suicidio en el futuro.		
Organización y rol dentro de ella que se beneficia con la oportunidad definida	Las organizaciones beneficiadas son fundaciones y organizaciones dedicadas a ofrecer servicios de salud mental para el público general como la Fundación Salud Mental del Valle o la Fundación Saldarriaga Concha. Los roles dentro de estas organizaciones que se benefician de dicha oportunidad son los encargados de explorar las redes en busca de gente que se beneficie de dichos servicios de salud.		
Técnicas y algoritmos a utilizar	Árboles de decisión	KNN	Bernoulli NB

2. Entendimiento y perfilamiento de los datos

a. Explicación de los datos

Nombre de Columna	Valores	Explicación
Text	Strings	Parrafos sacados de Reddit.

Class	suicide, non-suicide	
Words	Strings	Palabras claves procesadas tomadas de la columna “Text” con las que se determina el tipo de “Class”
Index	1 - 195,639	

b. Para el preprocesamiento de los datos se siguieron las siguientes etapas:

Para entender los datos se tomaron muestras para familiarizarse con la escritura e identificar posibles correcciones iniciales.

En primer lugar, se realizó la eliminación de los caracteres que no son ASCII, ya que esto generaba conflictos con el primer paso de preprocesamiento: contractions.

Luego de separar las contracciones, pasamos a realizar el tokenize de las palabras antes de continuar con el preprocesamiento. En el notebook se evidencia el proceso de estas funciones paso a paso, esto fue debido a que se necesitó debuggear las funciones para entender bien su funcionamiento sin demorar una ejecución fallida de todas las funciones (como se proponía hacer usando una única función de preprocesamiento). Se realizaron en este orden:

to_lowercase
remove_punctuation
replace_numbers
remove_stopwords

De esto se aprendió que el proceso más pesado es el de remover las stop words. Esto es debido a que se debe revisar el diccionario de stopwords completo, cada vez que se procesa una palabra. Esta función nos tomó 43 minutos en ejecutarse.

Para finalizar se realizó un stemming y lematización de las palabras para luego vectorizar las palabras resultantes.

Es importante mencionar que cada que se realizó una función que pudo haber borrado alguna palabra se eliminaron las filas que terminaron con valores NaN o

de longitud 0. Esto porque estos valores ponen en conflicto a las funciones de preprocesamiento.

3. Modelado y evaluación

a. Árbol de decisión

Los modelos de árboles de decisión son modelos de aprendizaje supervisado que funcionan para clasificar datos basándose en etiquetas dadas. Este algoritmo agrupa observaciones con valores similares y utiliza dichas agrupaciones para definir tanto la agrupación que mejor se ajuste a la variable dada como las agrupaciones que más se acercan con el fin de predecir los valores en las variables que más se ajustan a un resultado esperado en una variable objetivo. Para este caso, después de la modificación de los textos, se utilizó un arbol de decisión para definir si un texto fue escrito por una persona que intentó cometer suicidio con el fin de predecir (clasificar) los textos de personas que puedan ser propensas a hacerlo. En este caso, vectorizamos los datos utilizando el conteo de palabras presentes, pues es el que más se ajusta a nuestros objetivos. Utilizando el árbol de esta manera, nos dieron las siguientes métricas:

Esto nos demuestra que el modelo fue bastante efectivo a la hora de clasificar dichos textos.

	precision	recall	f1-score	support
0	0.87	0.88	0.88	22139
1	0.84	0.83	0.84	16989
accuracy			0.86	39128
macro avg	0.86	0.86	0.86	39128
weighted avg	0.86	0.86	0.86	39128

b. KNN

KNN es un algoritmo el cual es muy útil en un espacio multidimensional para emparejar a un punto con sus k-vecinos más cercanos. La idea es que los valores faltantes pueden ser aproximados basados en otras variables, por los puntos más cercanos a este. Este algoritmo es particularmente útil porque los datos pueden ser de diferentes tipos, ya sean continuos, discretos, ordinales y categóricos. En este caso, la variable "class", la cual es en la que se va a enfocar este estudio, es categórica.

Para este caso particular, el texto son párrafos sacados de Reddit, que es la variable que se tiene, basándose en esto se sacan las palabras claves de lo que se encontró en el texto. Entonces basándose en esto, personas que tengan palabras claves parecidas, deberían tener un tipo de clase parecido, en este caso suicide o non-suicide.

c. Bernoulli NB

Es un algoritmo derivado de la familia de los Bayesianos Ingenuos. Son algoritmos basados en el teorema de Bayes, es decir, en las distribuciones probabilísticas de los datos. Se les llama ingenuos por las asunciones fuertes sobre la independencia de las variables necesaria para simplificar el teorema de Bayes. En el caso de Bernoulli, es un algoritmo muy útil para clasificar variables binarias. Es importante mencionar que el modelo se alimentó por batches por las limitaciones físicas de la máquina. Por eso, se puede ver en el notebook que se hizo uso del método partial_fit para esto. Se realizó la predicción igualmente por batches. En este caso, obtuvimos las siguientes métricas:

Martin Ubaque
Alejandro Tovar
Gabriela Paez

	precision	recall	f1-score	support
0	0.77	0.91	0.84	22139
1	0.85	0.65	0.74	16989
accuracy			0.80	39128
macro avg	0.81	0.78	0.79	39128
weighted avg	0.81	0.80	0.79	39128

4. Resultados:

a. Video

Ver link en la wiki del proyecto.

b. Descripción resultados de los modelos

Como se puede ver, los 3 algoritmos usaron las mismas 3 métricas las cuales fueron: precisión, recall, f1-score y support. Teniendo esto en cuenta, la comparación entre las métricas fue muy sencilla. En las 4 métricas, las más altas fueron las del Árbol de Decisión, razón por lo que se concluye que este fue el mejor modelo entre los 3 tratados. La métrica que tiene en cuenta las otras tres, y la más precisa, se llama F1-score, que en este caso tuvo un puntaje de .84, un valor relativamente alto. Lo que esto significa es que es el mejor algoritmo para predecir si un mensaje o un texto escrito fue escrito por una persona que es más probable que sea suicida, o si no parece ser una persona suicida.

Martin Ubaque
Alejandro Tovar
Gabriela Paez

Referencias

OMS (2022). Depresión. Recuperado de https://www.who.int/es/news-room/fact-sheets/detail/depression#:~:text=La%20depresi%C3%B3n%20es%20una%20enfermedad,personas%20tienen%20depresi%C3%B3n%20(1).

Clínica Universidad de Navarra (2022). Depresión. Recuperado de https://www.cun.es/enfermedades-tratamientos/enfermedades/depresion