**Proyecto 1 - Salud Mental: Fase 1 y 2**

[**Prefacio - Trabajo en grupo**](#_26on834wexkj) **2**

[a. Fase 1: Lista de trabajos a realizar](#_6ow3f9ibce4t) 2

[b. Fase 2: Lista de trabajos a realizar](#_sw2vtit04jj1) 2

[i. Roles](#_gizefxx5mozw) 2

[ii. Reuniones:](#_uk09ohea8p73) 3

[iii. Tareas realizadas](#_r0d5iylp7c11) 3

[iv. 100 puntos](#_inzoxjvj12sj) 4

[**1. Comprensión del negocio y enfoque analítico**](#_j5mzzg32zwvw) **4**

[a. Marco teórico:](#_imjxoybzaksk) 4

[b. Objetivos y criterios de éxito:](#_hpsrhl5pdux) 4

[**2. Entendimiento y perfilamiento de los datos**](#_i26h5plkevm8) **5**

[a. Explicación de los datos](#_pqnij6b9480) 5

[b. Para el preprocesamiento de los datos se siguieron las siguientes etapas:](#_qf2vlh8sm73d) 6

[**3. Modelado y evaluación**](#_pynu4r5fsbhr) **6**

[a. Árbol de decisión](#_rwu49e1oc0a4) 6

[b. KNN](#_8299uv3mjv4p) 7

[c. Bernoulli NB](#_uuo7vxae73op) 8

[**4. Proceso de Automatización**](#_1f09jfwef8pd) **8**

[**5. Aplicación y justificación:**](#_60sqndl2685) **9**

[a. Aplicacion:](#_vd28f2he0du4) 9

[b. Justificación:](#_ib6zqrhvniqa) 9

[**6. Resultados:**](#_pj8qjyjkx0zg) **9**

[a. Video](#_k1tauq16l09b) 9

[b. Descripción resultados de los modelos](#_cdyitudpf8z) 10

[**Referencias**](#_2umhju2gxm88) **11**

# 

# *Prefacio - Trabajo en grupo*

## **Fase 1: Lista de trabajos a realizar**

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

**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.

## **Fase 2: Lista de trabajos a realizar**

### **Roles**

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

### **Reuniones:**

1. **Reunión de lanzamiento y planeación:**
   1. 25 de Octubre: Definición de nuevos roles y lluvia de ideas para el desarrollo de la aplicación.
2. **Reuniones de seguimiento:** 
   1. 28 de Octubre: Primera reunión semanal de seguimiento. (10m)
   2. 2 de Noviembre: Segunda reunión semanal de seguimiento. (30m)
   3. 5 de Noviembre: Tercera reunión semanal de seguimiento. (10m)
   4. 10 de Noviembre: Cuarta reunión semanal de seguimiento. (45m)

En las reuniones de seguimiento se hace seguimiento de las tareas establecidas, corrección de errores y replanteamiento de las cosas que no estén funcionando.

1. **Reunión de finalización:**
   1. 13 de Noviembre: Se consolidó el trabajo de la aplicación y el documento. Se revisó que todo estuviera corriendo correctamente, y se unieron los documentos de la fase 1 y 2 para tener uno maestro.

### **iii. Tareas realizadas**

**A. Proceso de automatización**

Encargado: Diego Alejandro Tovar

Horas de trabajo: 6

**B. Aplicacion**

Encargado: Martín Ubaque

Horas de trabajo: 8

**C. Aplicacion**

Encargado: Gabriela Paez

Horas de trabajo: 2

**D. Justificación**

Encargado: Gabriela Paez

Horas de trabajo: 2

**E. Evaluación y resultados**

Encargado: Martín Ubaque

Horas de trabajo: 3.

**F. Video**

Encargado: Martin Ubaque

Horas de trabajo: 2

**G. Organización del proyecto**

Encargado: Gabriela Paez

Horas de trabajo: 3.

### **iv. 100 puntos**

Finalmente, de acuerdo con la división de trabajo, si se tuvieran 100 puntos se dividirán de la siguiente manera:

Martin: 40

Alejandro: 30

Gabriela: 30

Al hacer la reflexión, se concluyo que Martín trabajó más en la parte 2 del proyecto 1, debido a que se enfoco y le metió más horas a la creación de la aplicación

# **Comprensión del negocio y enfoque analítico**

## **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.

## **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 |

# **Entendimiento y perfilamiento de los datos**

## **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 |  |

## **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.

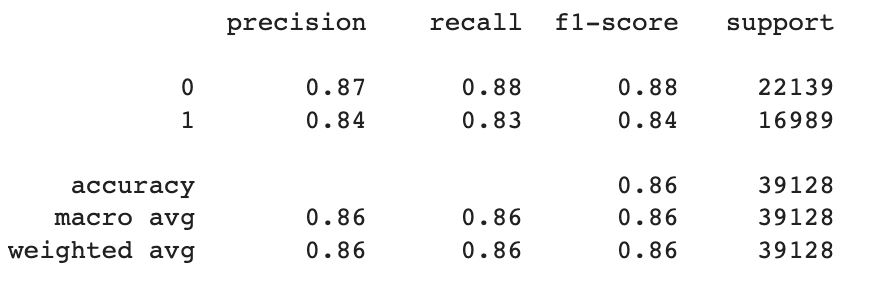
Es importante mencionar que cada que se realizó una función que pudo haber borrado alguna palabra se eliminaron las filas que terminaron con valores NaN o de longitud 0. Esto porque estos valores ponen en conflicto a las funciones de preprocesamiento.

# **Modelado y evaluación**

## **Árbol de decisión**

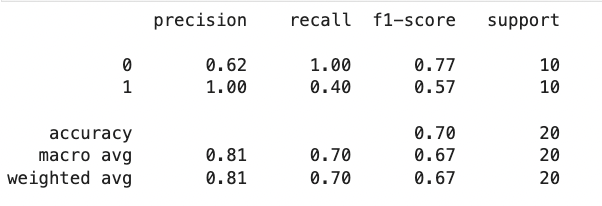
Los modelos de árboles de decisión son modelos de aprendizaje supervisado que funcionan para clasificar datos basándose en etiquetas dadas. Este algoritmo agrupa observaciones con valores similares y utiliza dichas agrupaciones para definir tanto la agrupación que mejor se ajuste a la variable dada como 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.



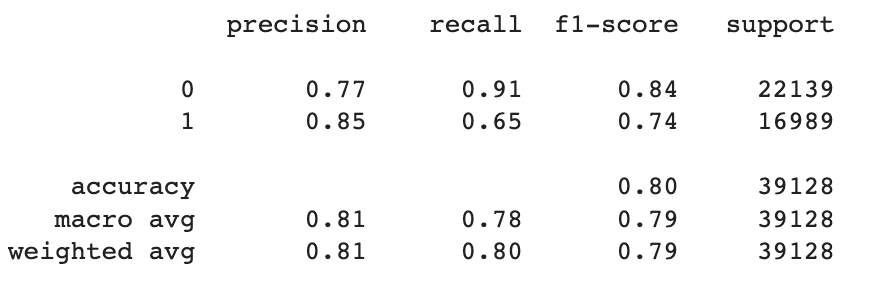
## **KNN**

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.

## **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:



# **Proceso de Automatización**

El código usado para este punto se encuentra en el repositorio. La parte la automatización del procesamiento de texto se puede encontrar en archivo titulado PipelineProyecto1\_V2, el cual contiene 1 clase con 15 “métodos”. Estos metodos hacen diferentes cosas, como remover todos los caracteres del texto que no son parte de ASCII, ya que teniendo en cuenta las palabras existentes, se puede determinar con una alta seguridad, que esto es un error. Adicionalmente, también pone todas las palabras de tokenizadas sin puntos, comas o cualquier tipo de puntuación. Por otro lado, convierte los números a palabras - ejemplo 1 se vuelve “uno”- todas las palabras se ponen en minúsculas, y se borran los stopwords, para así crear un mejor grupo de palabras claves.

# **Aplicación y justificación:**

## **Aplicacion:**

Para la creación de la aplicación, se utilizó el API FastAPI. Este nos permite fácilmente desplegar en un servidor aplicaciones simples que no tengan muchas funcionalidades, como esta. También se utilizaron Pipelines. Esto nos ayudó a automatizar el proceso del análisis de texto y guardar el modelo creado a partir de aquí en dicho pipeline. En otras palabras, al ingresar los datos en el pipeline, este automatizó el proceso de preprocesamiento explicado en el punto 2.b al igual que aplicarlos al modelo que mejor resultados nos dio, explicado en el punto 3.a. Con este pipeline, la aplicación recibe un input de texto de un usuario y, con este, devuelve una predicción sobre la situación de la persona que escribió dicho texto (determina si el paciente es propenso o no a cometer suicidio).

## 

## **Justificación:**

Como se mencionó previamente, alrededor del 5% de los adultos en el mundo sufren de depresión, lo cual resulta que alrededor de 700,000 personas se suicidan al año, sin embargo, actualmente no existen estadísticas de cuántas personas se intentan suicidar al año, pero si se sabe que es una cifra mucho más alta. Debido a esto, y al enfoque analitico inicial propuesto en la primera fase - 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.- decidimos la mejor aplicación era una que siguiera este enfoque analitico, el cual analizara texto planos para clasificarlos y poder determinar si la persona que lo escribió es propenso a suicidarse o no, pero que nos centraremos en usar este modelo en la aplicación para brindarle una ayuda al personal de salud enfocados en el área de salud mental. Por esto, se consideró necesario hacer aplicación pueda ayudar en esta área, para asistir en esta área, y ayudar a contribuir a los profesionales médicos con el objetivo de disminuir los suicidios. Esta aplicación está hecha para el sector de salud mental, específicamente los institutos del sistema nervioso y los centros psiquiátricos - ejemplo la Monserrat en Bogotá, Colombia - siendo usado como usuario final por los psiquiatras, psicólogos y médicos que trabajan en estos centros. La idea es que la aplicación funcione de apoyo, por la cual el personal médico pueda buscar información en las redes sociales del paciente, y pasar pasajes de texto que el paciente ha escrito por la aplicación, para ver si diría que el paciente es suicida o no. Esto, como apoyo a las notas que el personal de salud ha tomado, le puede ayudar a tomar una mejor determinación y dar un mejor diagnóstico.

**Organizaciones:** Los institutos del sistema nervioso y los centros psiquiátricos

**Usuarios finales:** El personal en estos centros, específicamente:médicos, psicólogos y psiquiatras.

**Rol de aplicación:** Apoyo para la determinación si un paciente es más probable a ser suicida o no.

# **Resultados:**

## **Video**

* + 1. Video Fase 1:

<https://youtu.be/fYUH9WwvSSY>

* + 1. Video Fase 2:

<https://www.youtube.com/watch?v=A5kzlGVyP6Y>

## **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.

# **Resultados de la aplicación**

# 

# 

# **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>