

MINERÍA DE DATOS

AVANCE I PROYECTO INTEGRADOR

EQUIPO 2

GRUPO 002 FRECUENCIA: MIÉRCOLES 7PM-10PM

1802920 JUÁREZ RODRÍGUEZ FLOR KARINA
1815973 MENDOZA ÁLVAREZ PILAR ABIGAIL
1802473 ORDAZ RUIZ MARGARITA

M.C. MAYRA CRISTINA BERRONES REYES

**SEMESTRE 7
LICENCIADO EN ACTUARÍA**

SAN NICOLÁS DE LOS GARZA A 28 DE OCTUBRE DE 2020

BASE DE DATOS

La base de datos a utilizar se obtuvo de la página Kaggle y se encuentra con el nombre de “Suicide Rates Overview 1985 to 2016”

URL: <https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016>

DESCRIPCIÓN DE LOS DATOS

El conjunto de datos que muestra la base de datos fue extraído de otros cuatro conjuntos de datos vinculados por tiempo y lugar, y se creó con la finalidad de encontrar señales correlacionadas con el aumento de las tasas de suicidio entre diferentes cohortes a nivel mundial, en todo el espectro socioeconómico.

Los datos se muestran en formato de tabla, la cual cuenta con 27,821 filas y 12 columnas con las siguientes descripciones: *país*, en esta columna el tipo de dato es texto, debido a que nos da el nombre del país donde ocurrieron los suicidios, *año*, en esta columna el tipo de dato es numérico, debido a que nos da el año en que ocurrieron los suicidios, *sexo*, esta columna sólo contiene dos tipos de valores “male” y female”, *grupo de edad*, esta columna contiene un rango de edad de 10 años de las personas que se suicidaron, *recuento de suicidios*, esta columna contiene el número de suicidios ocurridos, *población*, la columna tiene datos numéricos, ya que muestra el número de habitantes en ese año, *tasa de suicidio*, esta columna muestra un porcentaje el cual está dado por el número de suicidios ocurridos entre la población dividida entre mil, *composite key de país-año*, en esta columna se muestra el país seguido por el año, *índice de desarrollo humano (HDI) por año*, esta columna muestra un porcentaje, *PIB por año*, *PIB per cápita*, en estas 2 columnas las cantidades están dadas en dólares y generación (basado en el promedio de agrupación de edad), aquí el tipo de dato es texto, dado que nos da el nombre de la generación de las personas que se suicidaron.

JUSTIFICACIÓN DEL USO DE DATOS

La elección de esta base de datos se debió a que nos llamó la atención que tratara sobre el suicidio dado que es un tema muy importante y del cuál se debería de hablar más para saber de él y tratar de prevenirlo, además, dado que estamos atravesando una pandemia y muchas personas nos encontramos en confinamiento los problemas de salud mental se vieron incrementados y entre ellos se encuentra el suicidio. Luego de que abrimos y analizamos un poco más la base de datos vimos que aparte de ser un gran tema el que trata, tenía mucha información con la cuál trabajar y esto nos beneficiaría a la hora de elegir nuestro objetivo.

Esperamos que trabajar con este conjunto de datos nos ayude a entender más sobre el tema del suicidio y los diferentes factores que influyen en él para así poder crear un poco más de conciencia y ver cómo esto se puede implementar en los programas sociales que existen para la prevención de este.

PLANTEAMIENTO DEL PROBLEMA

Luego de un análisis un poco más profundo en la base datos y viendo lo que ésta nos brinda, consideramos que nuestra problemática a tratar en el proyecto es precisamente las tasas de suicidios. A simple vista pudimos observar que las tasas tienen una variación muy diferente según sea la ciudad y el año, esto hizo que nos cuestionáramos las razones del por qué sucede; tal vez ocurrió algo en cierta ciudad en cierto año que hizo que la tasa se fuera a la baja o en su defecto al alza y esperamos que el encontrar la respuesta a estas variaciones nos ayuden a lograr nuestro objetivo final.

OBJETIVO FINAL

Nuestro objetivo principal se basa en que, se observan variaciones en las tasas de suicidios, y queremos encontrar si existe relación alguna entre las variables que se presentan en la base de datos que se utilizará a lo largo del proyecto (presentadas como columnas), y determinar cuáles son aquellas características que podrían ser factores en común en los que cuales pueden enfocarse alternativas y/o programas sociales con la finalidad de decrementar en la medida de lo posible estas tasas.

Además de nuestro objetivo principal, también queremos enfocarnos en un objetivo secundario el cual sería investigar si los tiempos de crisis son un factor detonante a la alteración de las tasas de suicidio, incluyendo también la región/país en la que se encuentren.

PLANEACIÓN DE LAS HERRAMIENTAS A UTILIZAR

Visualización de datos

Nos interesó aplicar esta técnica, porque nos va a ayudar a encontrar nuestro objetivo, como poder detectar en que años ocurrieron más suicidios, y así poder investigar más a fondo, que sucesos ocurrieron en estos años, para entender que fue lo orilló a las personas a suicidarse, además de poder ver los rangos de edades, los países en donde se concentran más los suicidios, y en donde es donde se concentran menos.

Reglas de Asociación

Decidimos utilizar esta técnica, debido a que nuestro objetivo es encontrar una relación entre las columnas y el número de suicidios, entonces queremos agrupar datos con las mismas características, como lo son la generación, los países, el nivel socioeconómico del país, con el fin de encontrar relaciones con el número de suicidios que hubo.