

Universidad de sonora



Maestría en ciencia de datos

Curso propedéutico: Base de datos

## **Proyecto final**

Minjares Neriz Victor Manuel

Junio 2023

# Índice

1. Introducción	3
2. Objetivo	3
3. Explicación de la base de datos	3
4. Manipulación de la base de datos	5
5. Conclusiones	8

## 1. Introducción

Las bases de datos son una **colección de información estructurada**, con el objetivo de facilitar el manejo de una gran cantidad de datos. Una base de datos consiste en una colección de datos estructurados que se almacenan en tablas relacionales. Estas tablas están compuestas por filas y columnas, donde cada fila representa una entidad y cada columna representa un atributo o característica de esa entidad.

Aprender a manejar las bases de datos es primordial para un científico de datos, saber usarlas mejora exponencialmente la eficiencia y la cantidad de información que se puede extraer de los datos. Para esto, se usó **MySQL** como sistema de gestión de bases de datos relacionales (RDBMS por sus siglas en inglés) para explorar una base de datos relacional. MySQL es una de las opciones más populares y ampliamente utilizadas en la industria para administrar bases de datos relacionales. Proporciona una plataforma robusta y confiable para almacenar y manipular datos.

En resumen, el uso de bases de datos y la comprensión de cómo trabajar con ellas son habilidades cruciales para los científicos de datos. Al aprovechar herramientas como MySQL, es posible optimizar el manejo y análisis de grandes volúmenes de datos, lo que permite extraer información valiosa y tomar decisiones basadas en datos sólidos.

## 2. Objetivo

El objetivo de este trabajo es aprender a leer una base de datos, manipularla y saber hacer consultas (queries) significativas para extraer información de los datos de la base.

## 3. Explicación de la base de datos

La base de datos que se utilizó es [Mental Health in the Tech Industry](#), la cual fue obtenida de kaggle. Los datos son encuestas del 2014, 2016, 2017, 2018 y 2019 realizadas por

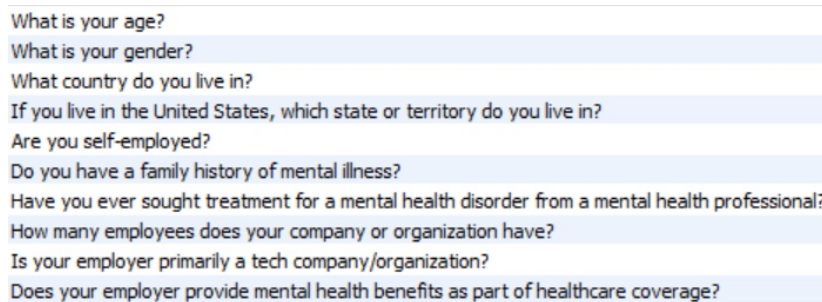
Open Source Mental Illness (OSMI). Las encuestas recolectan información sobre la salud mental y la frecuencia de trastornos de la salud mental en los lugares de trabajo en tecnología.

La base de datos contiene 3 tablas, **survey** (encuesta), **question** (pregunta) y **answer** (respuesta), cada una con 2, 2, 4 registros respectivamente. En la siguiente lista podemos ver los tipos de cada campo

- Survey: SurveyID (PRIMARY KEY, SMALLINT), Description (VARCHAR(29))
- Question: QuestionID (PRIMARY KEY, SMALLINT), QuestionText (VARCHAR(224))
- Answer: UserID (PRIMARY KEY, SMALLINT), SurveyID (FOREIGN KEY, SMALLINT), QuestionID (FOREIGN KEY, SMALLINT), AnswerText (TEXT)

La tabla **survey** contiene los años en los que se realizaron las encuestas los años 2014, 2016, 2017, 2018 y 2019. La tabla **question** tiene las preguntas que se realizaron. Por último, la tabla **answer** contiene las respuestas dadas y la relación entre las otras dos tablas, relacionando la respuesta, con la pregunta, con el usuario y el año de la encuesta.

Algunas preguntas realizadas se ven en la siguiente figura



What is your age?  
What is your gender?  
What country do you live in?  
If you live in the United States, which state or territory do you live in?  
Are you self-employed?  
Do you have a family history of mental illness?  
Have you ever sought treatment for a mental health disorder from a mental health professional?  
How many employees does your company or organization have?  
Is your employer primarily a tech company/organization?  
Does your employer provide mental health benefits as part of healthcare coverage?

Figura 1: Algunas preguntas de la tabla **question**.

## 4. Manipulación de la base de datos

Todas las modificaciones y los queries comentados se pueden encontrar en este [repositorio de GitHub](#).

Lo primero que realice cuando entre a la base de datos fue ver las tablas y contar el numero de registros de ellas. La tabla **answer** cuenta con 236,898 registros, la tabla **question** con 105 y la **survey** con 5. Con respecto a los campos, la descripción de la base de datos en kaggle estaba mal, ya que ninguna de las tablas tenia un campo como PRIMARY KEY, además de que las tablas no estaban relacionadas al hacer ingeniería inversa para obtener el diagrama identidad-relación.

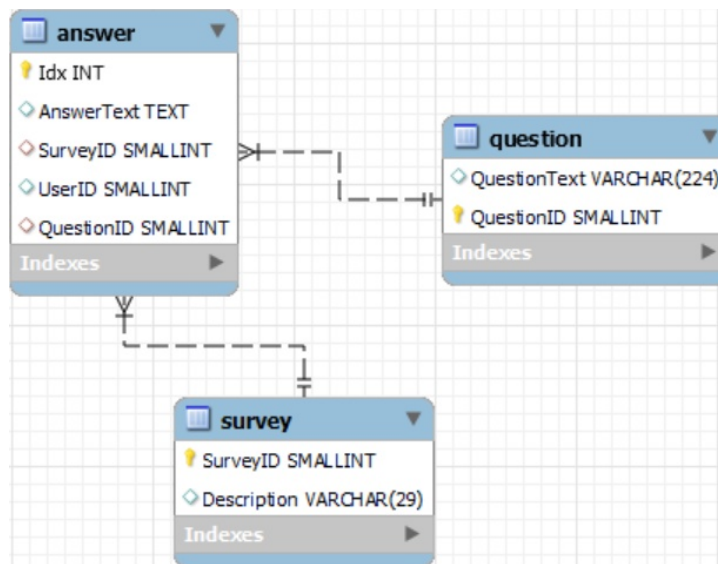


Figura 2: Diagrama entidad-relación de nuestra base de datos.

Para corregirlo primero tuve que agregar un campo nuevo en la tabla **answer**, porque el campo **UserID**, el cual esta descrito como PRIMARY KEY en el kaggle, tiene registros repetidos por lo tanto no puede ser PRIMARY KEY. A este nuevo campo la llame **Idx**, sería el índice de cada registro, y la designe como PRIMARY KEY para la tabla **answer**, se auto incrementa con cada registro. Después seleccione **SurveyID** y **QuestionID** como

PRIMARY KEY para las tablas **survey** y **question** respectivamente. Por último, asigne como FOREIGN KEY's a SurveyID y QuestionID en la tabla **answer** , con esto termine todas las modificación a la base de datos y obtuve el diagrama de entidad-relación que se ve en la figura 2.

Como datos interesantes que pude extraer de la base de datos están los siguientes

Año de la encuesta	Personas encuestadas
2014	1260
2016	1433
2017	756
2018	417
2019	352

Cuadro 1: Número de personas encuestadas en cada año.

En la tabla podemos ver que cada año bajo el número de personas encuestadas.

Edad del más joven	Edad del más viejo
11	99

Cuadro 2: Persona más joven y más vieja encuestada.

Las respuestas de las personas con respecto a su edad algunas eran falsas, ya que habían edades no positivas. Además, como vemos en el cuadro 2 las edades son muy extremas, por lo tanto muchas personas no querían dar su edad y ponían cualquier número.

País	Número de personas entrevistadas
Estados unidos	1853
México	12

Cuadro 3: Cantidad de personas entrevistadas por país.

La mayor cantidad de personas entrevistadas vienen de estados unidos, y 12 personas entrevistadas son mexicanas.

Respuesta	Número de respuestas
1	2412
0	1806

Cuadro 4: Respuesta a pregunta, ¿has buscado tratamiento para un trastorno de la salud mental con un profesional?

En la descripción de la base de datos, no dice que significa que respondieran 1 o 0 pero si suponemos que 1 significa *sí* y 0 *no*, entonces 2412 si han buscado tratamiento, 1806 no han buscado.

¿con compañeros?		
Año de encuesta	Número de personas	Porcentaje
2016	275	19.1 %
2017	206	27.2 %
2018	112	26.8 %
2019	89	25.2 %
¿con supervisores?		
2016	428	29.8 %
2017	256	33.8 %
2018	131	31.4 %
2019	117	33.23 %

Cuadro 5: Cantidad de personas que se sienten cómodas discutiendo de salud mental con compañeros y supervisores en cada encuesta.

Esta pregunta no la hicieron en la encuesta del 2014. Del cuadro 1 sabemos las personas totales y sacamos el porcentaje de las personas que contestaron que sí. Como vemos en la tabla 5 cada año va aumentando la comunicación sobre la salud mental en los trabajos en tecnología, lo cual es muy positivo.

En el siguiente cuadro también calculamos los porcentajes con el cuadro 1 y tampoco se hizo la pregunta en el 2014. En este cuadros podemos ver que los porcentajes de las respuestas no varían mucho año con año, manteniéndose casi constantes, con la diferencia que en la encuesta del 2016 existía la respuesta “tal vez” y en los años posteriores se cambió por las respuestas “no sé” y “posiblemente”.

Año de encuesta	Respuesta	Número de respuestas	Porcentaje
2016	Sí	575	40.1 %
	No	531	37 %
	tal vez	327	22.8 %
2017	Sí	324	42.8 %
	No	222	29.3 %
	No sé	66	8.7 %
	Posiblemente	144	19 %
2018	Sí	191	45.8 %
	No	112	26.8 %
	No sé	32	7.6 %
	Posiblemente	82	19.6 %
2019	Sí	147	41.7 %
	No	104	29.5 %
	No sé	26	7.3 %
	Posiblemente	75	21.3 %

Cuadro 6: Respuesta a pregunta, ¿Actualmente padeces de un trastorno de la salud mental?

## 5. Conclusiones

Al realizar este trabajo me di cuenta de la eficiencia de usar un sistema de gestión de bases de datos relacionales como MySQL para extraer información. Además, las consultas, por lo menos las que use, son intuitivas y no se requiere de aprenderse sintaxis complicadas.

Hablando un poco de la información que pude extraer de la base de datos que elegí, podemos observar como las personas, año con año, son más abiertos a hablar de su salud mental (cuadro 5), por lo menos entre las personas que trabajan en tecnología. Sin embargo, es importante destacar que los porcentajes no son muy altos y no han experimentado un crecimiento significativo en ese aspecto. Por otro lado, en el cuadro 6 vemos como más del 40 % de las personas encuestadas tenían, o tienen, un trastorno de salud mental. Esta cifra nos brinda una perspectiva sobre la relevancia y la necesidad de abordar de manera adecuada los problemas de salud mental en nuestra sociedad.



En conclusión, el uso de MySQL como sistema de gestión de bases de datos y la capacidad de realizar consultas intuitivas nos han permitido obtener información valiosa. En particular, la información obtenidos de nuestros datos resaltan la importancia de fomentar la apertura y conciencia en torno a la salud mental, aunque aún queda un camino largo por recorrer para lograr un mayor impacto y comprensión en este ámbito.