1.- Describe a detalle:

a) Que es la minería de datos?

La minería de datos es un área de las ciencias computacionales encargada de procesar y analizar colecciones de información, atraves de métodos computacionales y estadísticos, con el propósito de encontrar patrones, asociaciones, irregularidades o intereses para entender y explicar los datos. El objetivo de la minería de datos es extraer la información de una colección, agrupación o conjunto de datos para transformarlo en una estructura de datos entendible para su uso posterior.

A manera general, la minería de datos es utilizada para:

- Filtrar el ruido y obtener solo la información relevante de un conjunto de datos.
- Analizar y entender la información obtenida.
- En base a la información filtrada tomar decisiones útiles.
- Acelerar el ritmo de la toma de decisiones.

b) **Ejemplos de la minería de datos.**

- En anomalías, un caso especifico es la detección de fraudes en las transacciones bancarias. En base a cada usuario se empieza a recolectar información sobre los montos, las paginas, las pasarelas de pago utilizadas, intervalo de fechas con mas compras, etc. Y gracias a esto se pude formar un perfil de usuario sobre el uso que le da a su cuenta. Entonces cuando se produce un cargo por una cantidad o concepto atipico, un sitio desconocido (como puede ser de un país diferente al que reside el usuario) o pasarela inusual el sistema bloquea la transacción automáticamente y alerta al usuario.
- Cluster/Reglas para encontrar o crear un perfil persuasivo del usuario, las empresas de tecnología recaudan nuestra información y en base a esto nos muestran publicidad/contenido personalizado. Esto es debido a la minería de datos y al aprendizaje de maquina. A través del contenido que consumimos cada día, las paginas que visitamos, artículos que revisamos se puede ir generando un perfil persuasivo de cada usuario en base a la información recolectada e ir prediciendo que llamara la atención de cada usuario, ir encontrando los gustos, disgustos, patrones de comportamiento e interés sobre temas específicos, y formar estrategias para que compren productos en base a sus compras anteriores o recomendar videos en base al historial, etc.

c) **Procesos generales de la minería de datos**

1. Definición del problema que vamos a solucionar, y sabes si se puede resolver con minería de datos. Ej. Encontrar patrones entre los enfermos de COVID-19 asintomaticos.
2. Obtención de datos. **Acceso:** Definir quien nos brindara los datos, si son de acceso publico o pivado, si hay que pagar por ellos. **Muestreo:** Centrarnos en el usuario objetivo para obtener la información de estos. **Almacenamiento:** En donde los vamos a almacenar, en que formato. Ej. Se podría obtener la información a través de los hospitales, los formularios u

hoja de datos de cada paciente de COVID-19, y almacenarla en una base de datos estructurada para acceder a la información mas eficazmente.

3. Preprocesamiento de los datos. **Limpieza:** Limpiar los datos, ejemplo: Eliminar datos innecesarios de los pacientes, o dependiendo del dataset, eliminar a los pacientes no asintomáticos o pacientes no objetivos. **Normalizar:** Dejar los datos en mediadas o datos estandarizados para todos, ejemplo: Dejar alturas y pesos expresados en metros, kg para cada paciente. **Transformación:** Pasar el contenido textual a datos que puedan ser procesados de manera mas fácil, ejemplo: el genero pasarlo a "H", "M" o valores binarios.
4. Extracción de patrones: Procesos estadísticos. **Modelos de machine learning:** Que el sistema sea capaz de aprender a través de los datos. **Análisis de datos:** Visualizar y encontrar los patrones de la información. A través de un análisis estadístico, generar un método estadístico capaz de analizar y aprender de los datos para encontrar la representación y entender la información obtenida.
5. Despliegue del conocimiento. **Reportes:** entregar la información recaudada. Ej. Grafias sobre los patrones encontrados en las personas asintomáticas

2. Describe a detalle cada una de las áreas que intervienen en la minería de datos y como es que intervienen. Da un ejemplo de contribución de cada área.

1. **Modelado matemático.** Estadistias. La estadística consiste en métodos, procedimientos y formulas que permiten recolectar información para luego analizarla y extraer de ella conclusiones relevantes. Ejemplo: las estadísticas que obtiene el INEGI para analizar la calidad de vida de los mexicanos, las actividades y empleos que realizan las personas, sueldos, etc.
2. **Ciencias de administración e información.** Bases de datos. Interviene brindando las estructuras que nos permite almacenar y manipular la información de los datos. Ej. Las bases de datos estructuradas que almacenan la información de una empresa.
3. **Inteligencia artificial:** La inteligencia artificial consiste en desarrollar sistemas capaces de percibir, razonar y aprender para la resolución de un problema especifico. Dentro de la minería de datos se utiliza el machine learning, que nos permite desarrollar métodos estadísticos que permiten a nuestro sistema aprender a través de los datos. O el Deep learningque consiste en capas de redes neuronales que aprenden directamente de la información. Ej. En los autos inteligentes en base a la información obtenida de los sensores, el sistema puede predecir acciones y así evitar choques.
4. **Visualizacion de la información.** Area que nos permite analizar las representaciones graficas que expresan y sintetizan los datos. Ej. Histogramas sobre variables especificas.

3. Describe a detalle cada uno de los tipos de distribuciones que podemos encontrar en estadística.

1. **Uniforme.** Esta distribución se caracteriza por tener "la misma" cantidad de datos para cada valor observado. (No necesariamente tenemos la misma cantidad de datos, pero esta muy cercana).
2. **Unimodal simétrica.** Esta distribución se caracteriza por tener un pico al centro de los datos, se puede visualizar como una campana de gauss "simetrica" para ambos lados.

3. **Sesgada a la derecha:** Esta distribución se caracteriza por tener un pico del lado izquierdo de la información observada de la variable (como una campana de gauss), mientras que para el lado derecho esta información observada va decreciendo.
4. **Sesgada a la izquierda:** Esta distribución se caracteriza por tener un pico del lado derecho de la información observada de la variable (como una campana de gauss), mientras que para el lado izquierdo esta información observada va decreciendo.
5. **Bimodal.** Esta distribución se caracteriza por tener un 2 picos en los datos, podría decirse que es una combinación de dos campanas de gauss que se pueden observar.
6. **Multimodal.** Esta distribución esta caracterizada por tener múltiples picos, es decir, podemos observar en su grafica múltiples campanas de gauss unidas.

4. Describe a detalle para cada tipo de información (estructurada, texto e imágenes): a) Las características del tipo de información b) Las tareas que se pueden realizar con ese tipo de información c) Los problemas que se pueden presentar con ese tipo de información.

a) **Estructurada.** Esta información esta llevada por una estructura donde los datos observados pueden ser descritos por alguna instancia de la estructura antes dada, es decir, podemos ver estos datos como una tabla, donde tenemos variables y sus datos observados. Algunas tareas pueden ser asociación de variables, como lo podría ser “Cuantos hombres tienen consola” o “Cuantas mujeres tienen mascotas”, como se esta viendo en clase. Los problemas que puede presentar este tipo de información son que los datos observados estén incompletos, o que al momento de llenarlos no se siguió un mismo formato de manera que habría que transformar todo a un mismo formato.
b) **Texto.** Los datos con los que se trabaja son textos escritos. Se pueden realizar análisis semántico, traducción entre idiomas. Los problemas que puede presentar este tipo de información es la gran cantidad de palabras que puedan surgir de estos textos, o las palabras que se escriben igual, pero dependiendo del contexto tienen significados diferentes (lo cual puede meter ruido en varias tareas).
c) **Imágenes.** Los datos con los que se trabaja son grupos de imágenes donde se busca extraer o encontrar alguna representación de estas. Las tareas realizadas con esta información suelen ser de clasificacion, reconocimiento de objetos. Los problemas que puede presentar este tipo de información es que la imagen no tenga buena calidad, o este movida y se vea difuminada, o para el reconocimiento de objetos, que estos se presenten cortados por algún otro objeto y no se vea gran parte de este.