EXAMEN PRIMER MODULO: BASES DE DATOS

Diplomado: Minería de Datos y Técnicas Computacionales

Chávez Clavellina Ángel Uriel

11/03/2022

Pregunta 1

.

Sintetizando y Diagramando

1. **Dibuje un diagrama usando software, para explicar una metodología que involucre minería de datos.**

   La metodología CRISP-DM ha sido fuente de inspiración de otros estándares como SEMMA o ASUM-DM. Se conceptualiza en 6 fases.
   - Primera Fase: Entendimiento del Negocio
   - Segunda Fase: Entendimiento de los datos
   - Tercera Fase: Preparación de lo datos
   - Cuarta Fase: Modelado
   - Quinta Fase: Evaluación
   - Sexta Fase: Despliegue

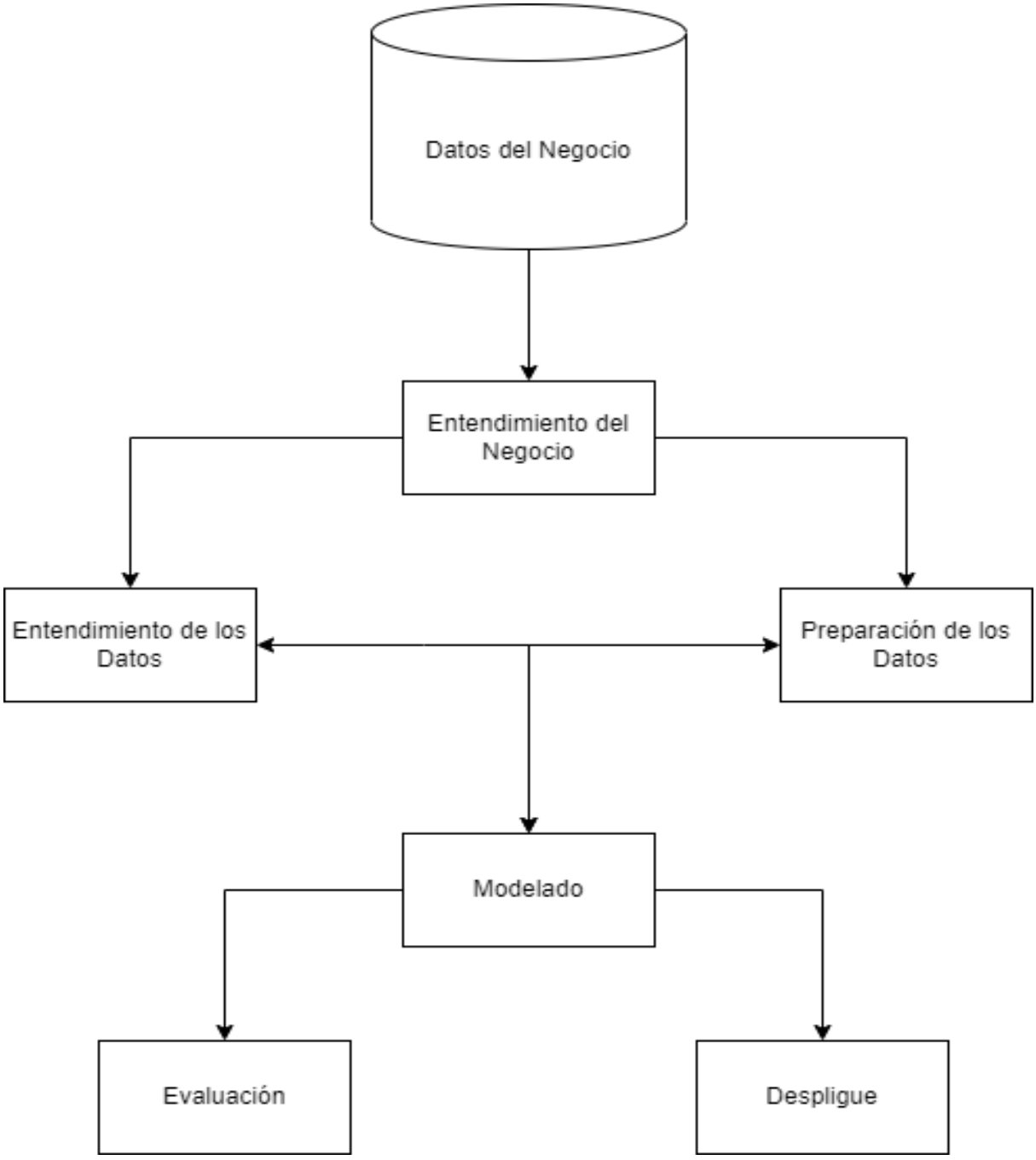
   Esta metodología surge cuando las empresas empiezan a implementar “Minería de datos”, pero ha tenido que ser modificada desde que la llamada “Ciencia de Datos”, una especie de evolución de la Minería, responde a muchas de las necesidades de negocio, que pueden tener las empresas.

2. **¿Qué tipo de diagrama realizó y porque lo selección**

   El tipo de mapa mental seleccionado es un mapa conceptual, con el describí de forma clara y secuencial las fases que componen la metodología.

3. **Aspectos y validaciones que el mapa debe cumplir**
   - ✓ No líneas/flecas que conecten nodos sin sentido.
   - ✓ El flujo del mapa debe de ser claro y secuencial
   - ✓ Uso de objetos visuales
   - ✓ Un nodo raíz

Mapa Conceptual



Pregunta 2.

Sintetizando y Diagramando

1. **Investigue, seleccione y analice, un artículo relacionado con la definición y aplicación de minería de datos.**

   El artículo “Minería de datos”, correspondiente a la lista de sugerencia, lo expone SAS en su página oficial. El texto hace una breve recopilación histórica sobre la forma en como este proceso conocido como “ El arte de hurgar entre grandes volúmenes de datos con el fin de descubrir conocimiento” ha evolucionado a través del tiempo. Desde 1990, año en el que oficialmente se emplea el término “Minería de Datos” y en donde las prácticas manuales y tediosas tomaban mucho tiempo, hasta lo que hoy en día y gracias a los avances en el poder y la velocidad de procesamiento, nos han permitido hacer, análisis fáciles, rápidos y automatizados.

   La importancia que tiene la Minería de datos en la industria es detallada en la última década, hoy en día el 90% del mundo digital esta conformado por datos No Estructurados, listos para ser procesados, entendiendo el fin relevante, para la mejor toma de decisiones, funciona a través de la intersección entre el modelado descriptivo, predictivo y prescriptivo.

   Es de gran prioridad en muchas industrias y disciplinas, tal como Seguros, Educación, Manufactura, Bancos, Retail, etc.

2. **Escriba 5 puntos, que considere los más importantes de la lectura, justifique su respuesta.**

- “Cuanto más complejos son los conjuntos de datos recopilados, mayor es el potencial que hay para descubrir insights relevantes”.

  Los insights que aportan valor, sin duda son un gran tesoro dentro de grandes volúmenes de datos, pueden hacer que cualquier organización cambie totalmente su paradigma sobre algo en específico.
  Recuerdo haber leído alguna vez una noticia relacionada a un insight que particularmente Wal-Mart encontró en sus datos de ventas, se dieron cuenta que había una alta correlación entre comprar artículos para bebé y comprar cervezas, ¿Cómo?, ¿Por qué?, se estudio el comportamiento, concluyendo que mientras las mujeres post-embazaro están más al tanto del bebé y de sus necesidades, los hombres son quienes se encargan de proveer y comprar lo necesario. “Si debo de ir a comprar pañales para mi hijo, por qué no de paso compro también un par de cervezas”.

  Este hallazgo dentro los propios datos de Wal-Mart hicieron que se crearan campañas publicitarias para promocionar mejor ambos artículos.

- “En la última década, los avances en el poder y la velocidad de procesamiento nos han permitido llegar más allá de las prácticas manuales, tediosas y que toman mucho tiempo al análisis de datos rápido, fácil y automatizado”.

  Como bien menciona el artículo, el avance tecnológico que hoy en día tenemos, no es de hace mas de 20-30 años, es demasiado reciente. No vimos como entró la era digital, pero si como cada que pasa el tiempo, se apodera de más áreas en la industria, incluso está empezando a sustituir actividades que hace todavía 10 años eran cruciales en la toma de decisiones.
  La minería de datos, los eficaces procesamientos de máquina y el uso particular que se les pueda dar, nos están llevando a automatizar cualquier cosa que se nos ponga enfrente, desde una simple acción de texto hasta alimentar algoritmos con millones de datos que tienen el poder de predecir y/o contestarte cualquier cosa. ¿Seremos capaces de poder crear un algoritmo que sea capaz de incluso predecir nuestro muerte?, ¿Seremos capaces de sustituir a los profesionales de la salud por algoritmos muy bien entrenados con el poder suficiente de darnos respuesta a la misma atención médica?, suena un tanto utópico, pero no hace mas de 30 años hubiéramos imaginado conectarnos con otra persona al otro lado del mundo en tiempo real.

- “Lo que era antiguo es nuevo otra vez, ya que la minería de datos continúa evolucionando para igualar el ritmo del potencial sin límites del big data y poder de cómputo asequible”

  Se piensa que la tan muy famosa “Era predictiva” es algo reciente, una disciplina que nació con la digitalización, esto es totalmente un error, porque todos esos algoritmos no son mas que modelos matemáticos y estadísticos que llevan años existiendo, lo que en efecto es evidente es el enorme potencial que tienen, ante una ola de información que sin duda lleva poco tiempo, jamás en la historia de la humanidad se habían generado tantos datos como sucede hoy día.
  Una red neuronal hace 20 años no tenía el potencial que tiene hoy, no por su estructura, sino por la limitación tecnológica que se tenía.

- “Los algoritmos automatizados ayudan a los bancos a entender a su base de clientes y también los miles de millones de transacciones en el corazón del sistema financiero”.

  Los sistemas financieros no son más que máquinas que siguen funcionando porque hay alguien que día con día está analizando los datos que entran, si el día de mañana se apaga el mundo digital, sin duda se entraría en una recesión económica devastadora.

- “ Hacemos un mejor trabajo de analizar lo que realmente necesitamos analizar y de predecir lo que realmente deseamos predecir”

Así como hay olas y olas de información navegando en grandes volúmenes de datos, también es cierto que mucha de esta información es insignificante. Desde el inicio se tiene que saber específicamente que se pretende hacer y a donde se quiere llegar, si no se logra determinar conscientemente una manta de propósitos, intentar usar analítica avanzada solo será una pérdida de tiempo.

Pregunta 3

Analizando

1. **Seleccione uno de los temas propuestos**

   Seleccioné el tema 'De control de inventarios", un grupo de socios decidieron expandir su negocio, venden artículos para deportistas, atletas y/o personas que regularmente practican algún deporte y/o actividad física.

   Empezaron con una tienda local hace 6 años en la CDMX, actualmente han podido abrir en una sola acción, una sucursal en cada una de las ciudades más importantes económicamente hablando, tales como, Guadalajara, Monterrey y Tijuana. Lograron atraer a un inversionista, quien se dio cuenta del gran impacto que el negocio está teniendo en sus operaciones de compra y venta dentro de la Ciudad de México.

   Se sabe que desde hace algunos años esta creciendo la demanda de personas que empiezan a hacer algún deporte y/o actividad física. Particularmente este negocio ha sabido posicionarse bien en el mercado dada su gran estrategia de publicidad que han venido implementado desde que inició la pandemia, los socios son expertos en marketing y entendieron a cómo jugar con las redes sociales para obtener rendimiento. Tik-Tok es una fuente masiva inmediata de atención, que si se ocupa con inteligencia se puede hacer crecer un negocio en semanas.

   Esta acción de colocar sucursales ha hecho que aquello que inició como un autoempleo, 6 años después este consolidada ya como una empresa con operaciones nacionales e internacionales.

   Las tiendas llevan ya aproximadamente 6 meses en producción y los socios se han dado cuenta del gran problema interno que la empresa empieza a tener, sus sistemas de control de inventarios están ya obsoletos, causando retrasos en los envíos, una mala administración del dinero que entra por cada una de las ventas de cada una de las sucursales y dado que el inventario general de la empresa se ha expandido, constantemente pierden el control y el monitoreo de los artículos que se están comercializando. No se tomó en cuenta la gestión de la información, se pensó que adoptar el mismo proceso rudimentario que se ocupaba en la primera tienda sería lo ideal, solía registrarse absolutamente todo lo que entraba en carpetas físicas y como no había necesidad de hacer envíos internacionales, no logró crearse una estrategia adecuada para ello.

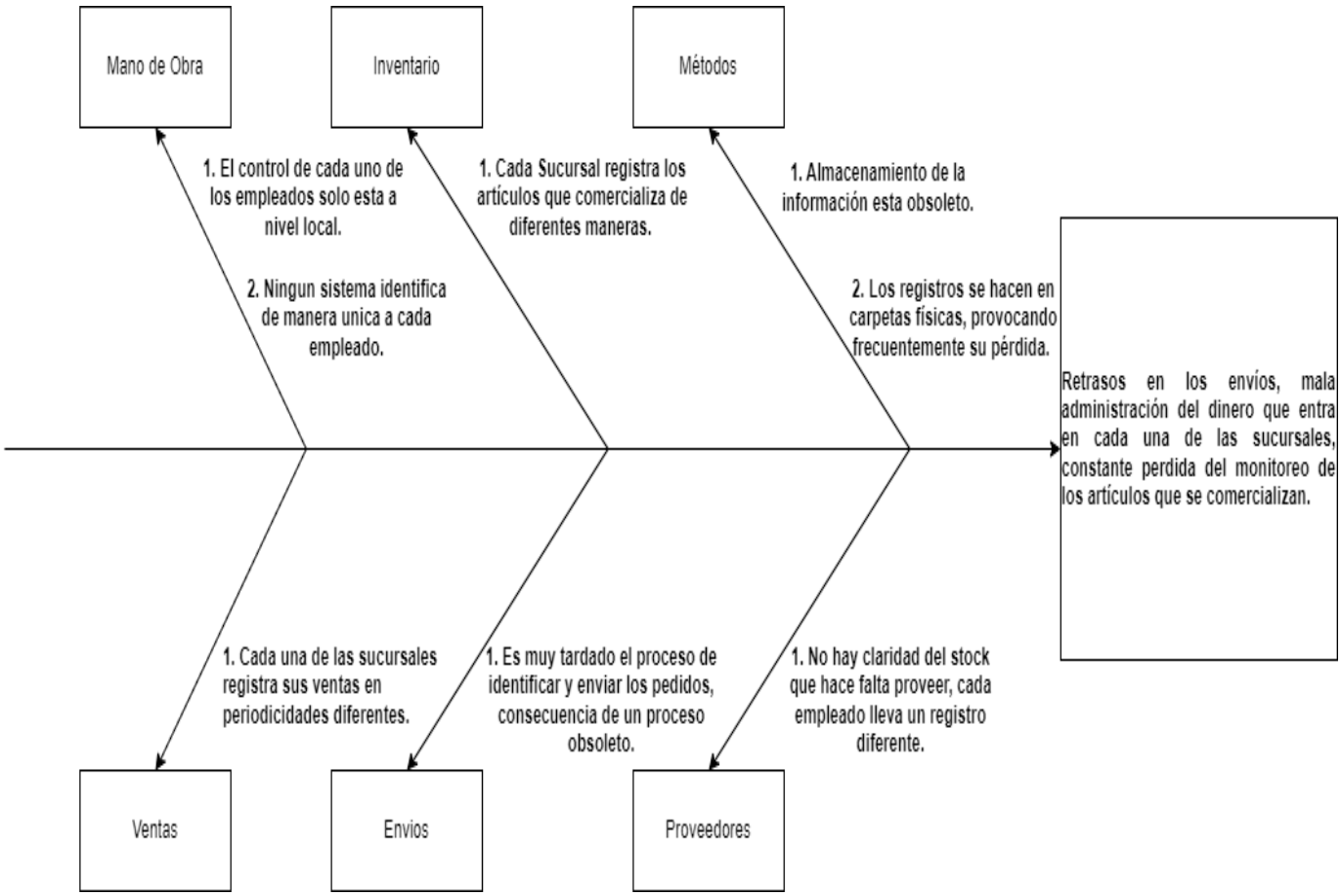
Los socios están sumamente preocupados porque están perdiendo el control de la información que los gerentes les comparten cada semana, cada tienda hace lo que quiere, tienen sus propios estándares de negocio, un mismo artículo puede estar registrado de todas las maneras posibles, dificultando su administración. Se pretende mejorar lo antes posible estas problemáticas, porque se está empezando a perder mucho dinero.

2. Definición de alcance:

Se estandarizará y se homologará toda la información que la empresa genera día con día, con el objetivo de mejorar todas las practicas posibles, se procederá a la creación de una base de datos que permitirá llevar un registro de las ventas de los artículos así como del mismo stock. Se revisará el expediente de todos los empleados y de ser necesario se reasignará un identificador único para monitorear sus actividades.

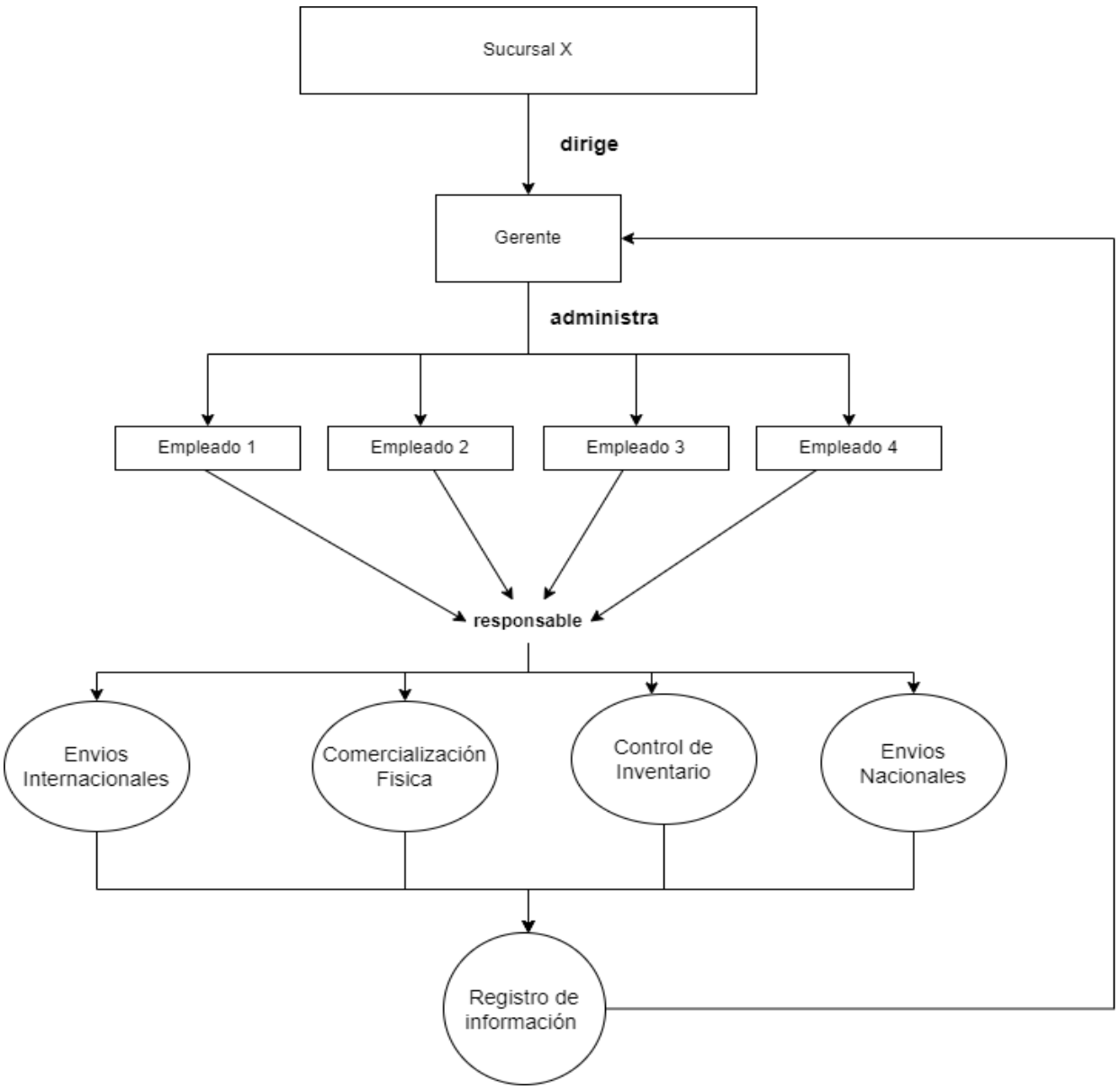
Diseñando

1. Realice un diagrama para analizar el problema, use un software para dibujarlo. Que diagrama seleccionó y porqué.



El diagrama mostrado corresponde a un diagrama de Pescado, en el se identifican los problemas, por los que la empresa está pasando y las causas que los están generando. Estos problemas deben de ser resueltos en vista de que muchas de las pérdidas económicas se están viendo ligadas a los procesos obsoletos que se siguen empleando.
Seleccioné este diagrama porque me parece que sin tener conocimientos acerca del propio diagrama, visualmente se logra entender lo que trata de comunicar, un problema con sus posibles causas.

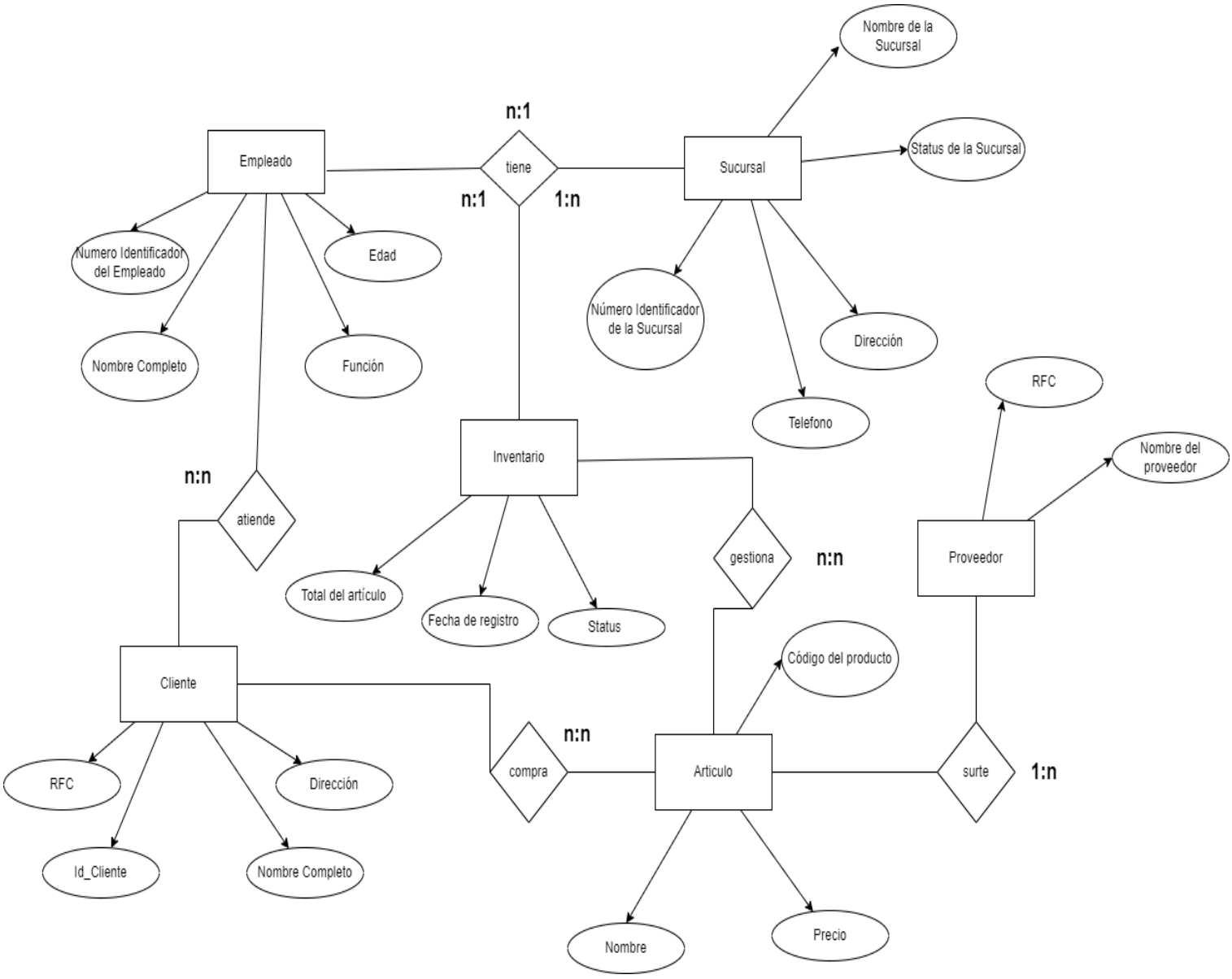
2. **Realice un mapa conceptual para mostrar las áreas involucradas en el proceso así como la importancia de cada área. ¿Por qué considera que el mapa realizado es correcto? Dibújelo con una herramienta de software.**



En el mapa conceptual se indican los elementos inmersos en la organización que cada una de las sucursales tiene actualmente, se aprecia la jerarquía y la relación que hay entre las actividades individuales, que es justo el gran problema que tiene este proceso, dado que todos tienen las mismas responsabilidades, todos lo hacen a su manera, no hay un orden, una organización y ni mucho menos registros homogéneos. El mapa es correcto porque indica explícitamente la organización actual.

Construyendo: Modelado Conceptual y Físico

3. Realice el modelo ER usando un software para diagramar (Entidades, atributos, relaciones y tipo de relación)

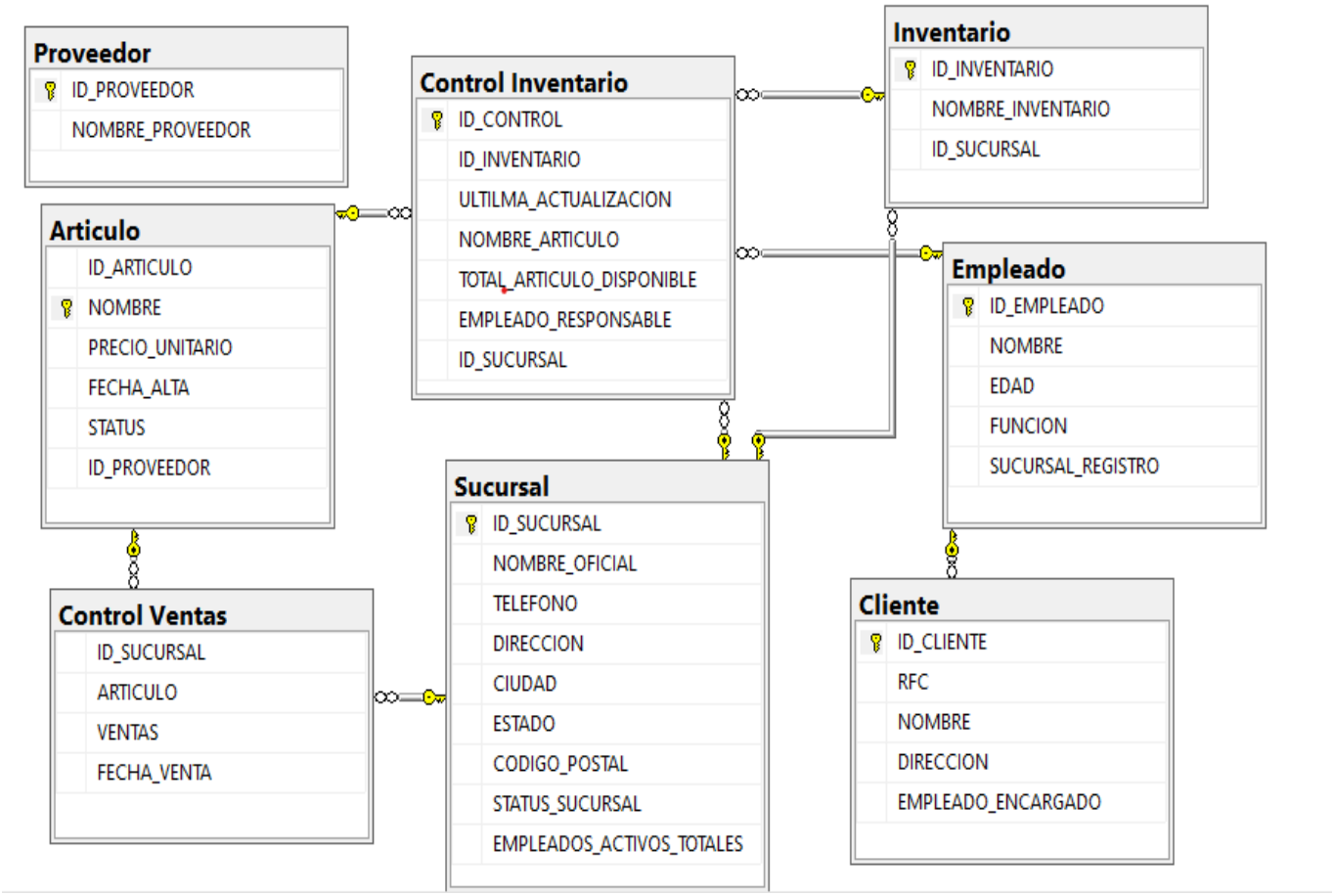


Se llevó acabo el proceso de normalización en la primera, segunda y tercera forma normal, comprobando los siguientes pasos:

- IFN:
  - Elimine grupos repetidos de tablas individuales.
  - Creación de tabla independiente para cada conjunto de datos relacionados.
  - Cada conjunto está identificando con un ID.
- 2FN:
  - Creación de tablas independientes para valores que se apliquen a varios registros.
  - Relacionar estas tablas con una clave externa.

- 3FN:
  - Elimine los campos que no dependan de la clave.

4. Convierta el modelo ER a un modelo R, usando una metodología y dibújelo usando un software para diagramar.



Manteniendo y obteniendo información del sistema

Se ha creado una Base de Datos, importando los datos de archivos csv.

Las consultas siguientes involucran al menos dos tablas, en cada uno de plantea el enunciado.

1. Los Artículos que más se venden e identificar de que sucursal provienen.

```
SELECT ID_SUCURSAL, ARTICULO, SUM(VENTAS) AS VENTAS_TOTALES_X_ARTICULO
FROM [Control Ventas]
GROUP BY ID_SUCURSAL, ARTICULO
ORDER BY SUM(VENTAS) DESC
```

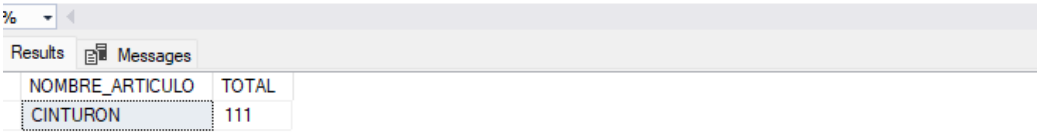
Results | Messages

ID_SUCURSAL	ARTICULO	VENTAS_TOTALES_X_ARTICULO
3	CINTA	78288
2	MU◆EQUERA	75932
1	SOGA	68614
3	GORRA	38704
3	FAJA	33320
1	BRAZALETE	30429
2	WHEY PROTEIN	24960
4	CINTURON	23320
3	MOCHILA	19890
2	CANDADOS	17664
3	BARRA ENERGETICA	8588

2. En la fecha más reciente, que articulo tiene mayor disponibilidad en stock.

```
SELECT NOMBRE_ARTICULO, SUM(TOTAL_ARTICULO_DISPONIBLE) AS TOTAL
FROM [Control Inventario]
WHERE  ULTILMA_ACTUALIZACION = (SELECT MAX(ULTILMA_ACTUALIZACION) FROM [Control Inventario])

GROUP BY NOMBRE_ARTICULO
ORDER BY SUM(TOTAL_ARTICULO_DISPONIBLE) DESC
```



Results | Messages

NOMBRE_ARTICULO	TOTAL
CINTURON	111

3. El nombre del Empleado que más registra stock y de que sucursal proviene.

```
CREATE VIEW AUX2 AS
SELECT E.NOMBRE, E.SUCURSAL_REGISTRO, CI.NOMBRE_ARTICULO, CI.TOTAL_ARTICULO_DISPONIBLE, S.NOMBRE_OFICIAL
FROM Empleado AS E
INNER JOIN
     [Control Inventario] AS CI
     ON E.ID_EMPLEADO = CI.EMPLEADO_RESPONSABLE
INNER JOIN
     Sucursal AS S
     on E.SUCURSAL_REGISTRO = S.ID_SUCURSAL

SELECT TOP 1 NOMBRE, SUCURSAL_REGISTRO, NOMBRE_OFICIAL, SUM(TOTAL_ARTICULO_DISPONIBLE) AS TOTAL_ARTICULO
FROM AUX2
GROUP BY NOMBRE, SUCURSAL_REGISTRO, NOMBRE_OFICIAL
ORDER BY SUM(TOTAL_ARTICULO_DISPONIBLE) DESC
```

Results | Messages

NOMBRE	SUCURSAL_REGISTRO	NOMBRE_OFICIAL	TOTAL_ARTICULO
Rosenda Infante Suarez	4	Centzontototl	2710

4. El empleado con la edad mínima.

```
SELECT TOP 1 NOMBRE, MIN(EDAD) AS EDAD_MINIMA
FROM EMPLEADO
GROUP BY NOMBRE
```

Results | Messages

NOMBRE	EDAD_MINIMA
Aaran Arce Checa	34

5. El empleado con la edad máxima.

```
SELECT TOP 1 NOMBRE, MAX(EDAD) AS EDAD_MAXIMA
FROM EMPLEADO
GROUP BY NOMBRE
ORDER BY EDAD_MAXIMA DESC
```

Results | Messages

NOMBRE	EDAD_MAXIMA
Andres Felipe Naranjo Chacon	50

Bibliografía

https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/

https://www.sas.com/es_mx/insights/analytics/data-mining.html