Universidad Nacional del Nordeste

Facultad de Ciencias Exactas y Naturales y Agrimensura

Cátedra: Base de datos II

Año: 2020

Trabajos Prácticos 2da. Parte

Alumno: Vargas Cristian Raúl

L.U. N°: 39513

D.N.I. N°: 32836858

Serie Ejercicios Prácticos 5
Bases de Datos Relacionales Extendidas

Modelo relacional anidado

1. Dado el siguiente esquema (en forma de tupla):

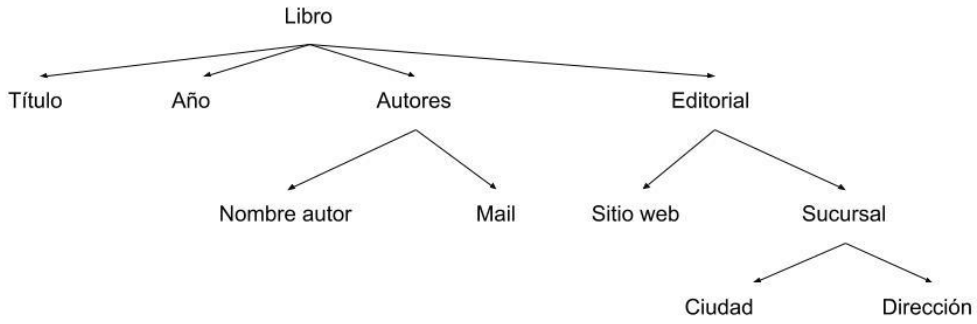
Libro

Titulo	Año	Autores		Editorial		
		Nombre autor	mail	Sitio web	Sucursal	
					Ciudad	Dirección

a) Realice la representación en árbol del esquema Libro

b) Obtenga la definición del esquema Libro

- **Libro** = (título, año, autores, editorial)
- **Autores** = (nombre autor, mail)



- **Editorial** = (sitio web, sucursal)
- **Sucursal** = (ciudad, dirección)

2. Sea la siguiente definición del esquema Universidad:

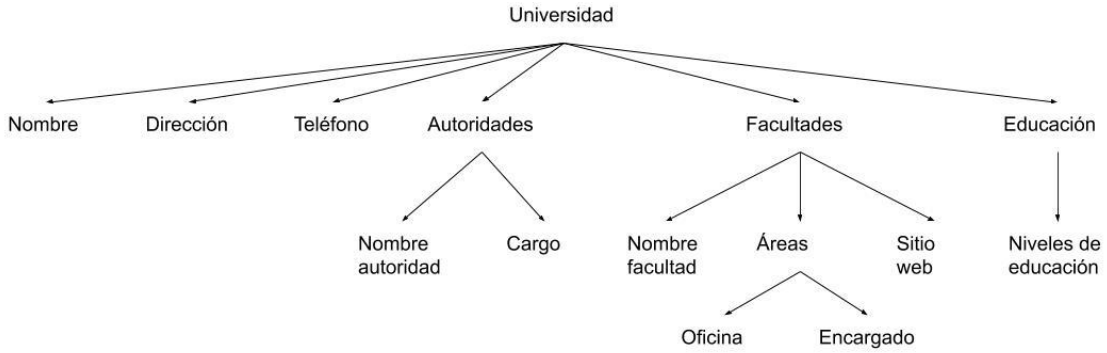
Universidad = (Nombre, Dirección, Teléfono, Autoridades, Facultades, Educación)
Autoridades = (Nombre autoridad, Cargo)
Facultades = (Nombre facultad, Áreas, Sitio web)
Educación = (Niveles de educación)
Áreas = (Oficina, Encargado)

a) Represente el esquema con el formato de una tupla

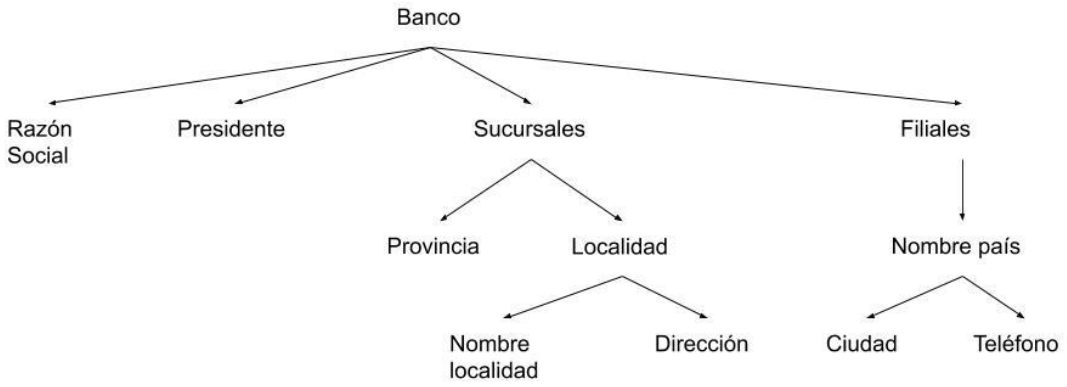
Universidad

Nombre	Dirección	Teléfono	Autoridades		Facultades				Educación
			Nombre autoridad	Cargo	Nombre facultad	Áreas		Sitio web	Niveles de educación
						Oficina	Encargado		

b) Realice la representación en árbol del esquema



3. Sea la siguiente representación en árbol del esquema Banco:



a) Obtenga la definición del esquema Banco

Banco = (Razón Social, Presidente, Sucursales, Filiales)
Sucursales = (Provincia, Localidades)
Filiales = (Nombre del país)
Localidades = (Nombre localidad, Dirección)
Nombre del país = (Ciudad, Teléfono)

b) Represente el esquema con el formato de una tupla

Razón social	Presidente	Sucursales			Filiales	
		Provincia	Localidades		Nombre país	
			Nombre localidad	Dirección	Ciudad	Teléfono

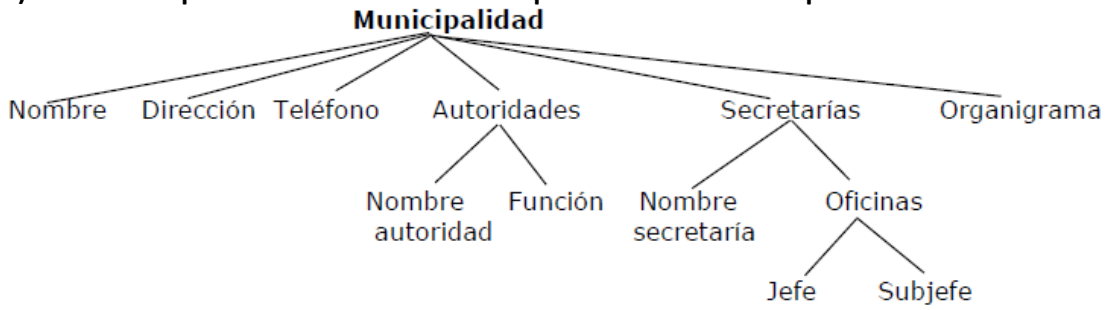
4) Sea la siguiente definición del esquema anidado Municipalidad:
Municipalidad = (Nombre, Dirección, Teléfono, Autoridades, Secretarías, Organigrama)
Autoridades = (Nombre autoridad, Función)
Secretarías = (Nombre secretaría, Oficinas)
Oficinas = (Jefe, Subjefe)

a) Represente el esquema anidado Municipalidad con el formato de una tupla

Municipalidad

Nombre	Dirección	Teléfono	Autoridades		Secretarías			Organigrama
			Nombre autoridad	Función	Nombre secretaría	Oficinas		
						Jefe	Subjefe	

b) Realice la representación en árbol del esquema anidado Municipalidad



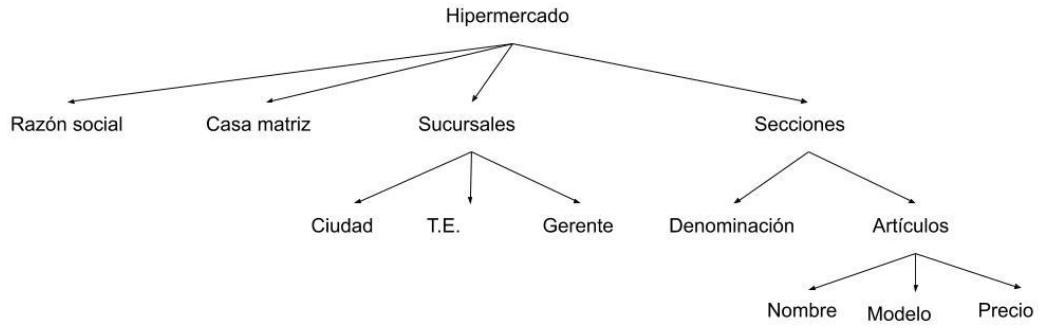
5) Dado el siguiente esquema (en formato de tupla):
Hipermercado

Razón Social	Casa matriz	Sucursales			Secciones			
		Ciudad	T.E.	Gerente	Denominación	Artículos		
						Nombre	Modelo	Precio

a) Realice la representación en árbol del esquema

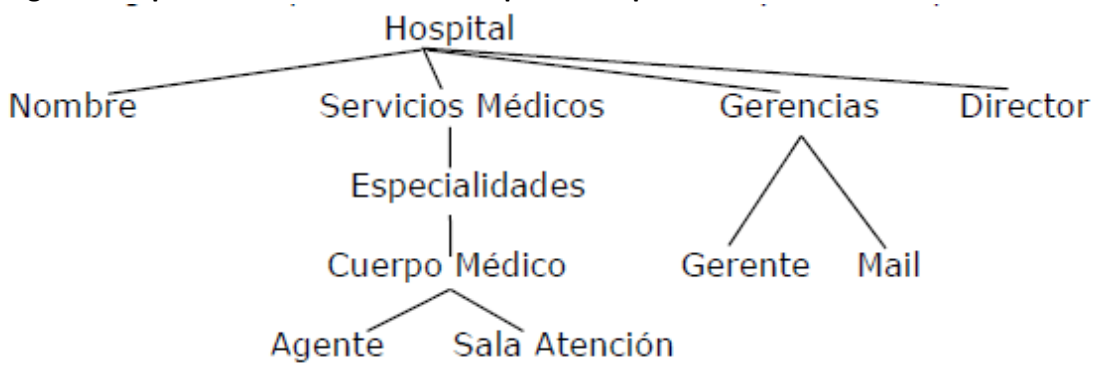
b) Obtenga la definición del esquema Hipermercado

Hipermercado = (Razón Social, Casa matriz, Sucursales, Secciones)



Sucursales = (Ciudad, T.E., Gerente)
Secciones = (Denominación, Artículos)
Artículos = (Nombre, Modelo, Precio)

6. Siguiente representación en árbol del esquema Hospital:



a) Obtenga la definición del esquema Hospital
Hospital = (Nombre, Director, Servicios Médicos, Gerencias)
Servicios Médicos = (Especialidad)
Gerencias = (Gerente, Mail)
Especialidad = (Cuerpo médico)
Cuerpo médico = (Nombre, Sala atención)

b) Represente el esquema con el formato de una tupla
Hospital

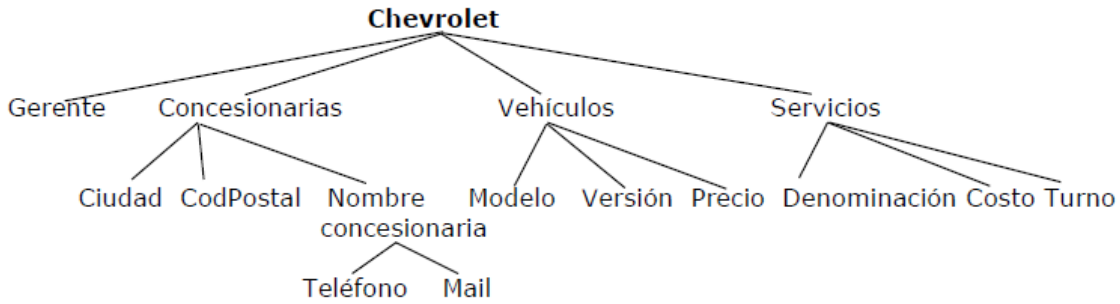
Nombre	Director	Servicios Médicos		Gerencias	
		Especialidad		Gerente	Mail
		Cuerpo medico			
		Nombre	Sala Atención		

7) Sea la siguiente definición del esquema anidado de la marca Chevrolet:
Chevrolet = (Gerente, Concesionarias, Vehículos, Servicios)
Concesionarias = (Ciudad, Cod. Postal, Nombre concesionaria)
Nombre concesionaria = (Teléfono, Email)
Vehículos = (Modelo, Versión, Precio)
Servicios = (Denominación, Costo, Turno)

a) Represente el esquema anidado Chevrolet con el formato de una tupla

Gerente	Concesionarias				Vehículos			Servicios		
	Ciudad	CodPostal	Nombre concesionaria		Modelo	Versión	Precio	Denominación	Costo	Turno
			Teléfono	Mail						

b) Realice la representación en árbol del esquema anidado Chevrolet



Base datos temporales
Esquema de base de datos de tiempo válido
8. Dado el siguiente esquema de base de datos:
Empleado

Apellido	Dni	Sueldo	Cargo
Vidal	11122233	80000	10
Pruyas	22117789	25500	15
Reyes	32190784	39700	19

a) Realizar los cambios necesarios para representar la base de datos con el esquema de tiempo válido, teniendo en cuenta los siguientes valores de tiempos de inicio y final válidos para cada tupla:

	Tiv (Vst)	Tfv (Vet)
Vidal	01-06-2019	Ahora
Pruyas	20-05-2020	Ahora
Reyes	01-02-2017	30-09-2020

Apellido	Dni	Sueldo	Cargo	Tiv	Tfv
Vidal	11122233	80000	10	01-06-2019	Ahora
Pruyas	22117789	25500	15	20-05-2020	Ahora
Reyes	32190784	39700	19	01-02-2017	30-09-2020

b) Actualizar el sueldo del empleado Vidal a 92000, que será efectivo a partir del día 01-09-2020, representar la modificación.

Apellido	Dni	Sueldo	Cargo	Tiv	Tfv
Vidal	11122233	80000	10	01-06-2019	**31-08-2020**
Vidal	11122233	**92000**	10	**01-09-2020**	**Ahora**

c) Modificar el cargo de Pruyas a 20, que será efectivo a partir del día 31-08-2020

Apellido	Dni	Sueldo	Cargo	Tiv	Tfv
Pruyas	22117789	25500	15	20-05-2020	**30-08-2020**
Pruyas	22117789	25500	**20**	**31-08-2020**	**Ahora**

d) Dar de baja al empleado de apellido Pruyas Dni 22117789, que deja de prestar servicios a partir del día 31-03-2021.

Apellido	Dni	Sueldo	Cargo	Tiv	Tfv
Pruyas	22117789	25500	20	31-08-2020	**31-03-2021**

e) Incorporar la siguiente tupla que corresponde al empleado de apellido Méndez, que comienza a prestar servicios desde el día 14-09-2020:

Apellido	Dni	Sueldo	Cargo
Méndez	32076123	58500	12

Apellido	Dni	Sueldo	Cargo	Tiv	Tfv
Méndez	32076123	58500	12	**14-09-2020**	**Ahora**

Esquema de base de datos bitemporal

9. Dado el siguiente esquema de base de datos bitemporal:

Empleado

Apellido	Dni	Sueldo	Tiv (Vst)	Tfv (Vet)	Tit (Tst)	Tft (Tet)
Pérez	11122233	48000	15-06-2020	Ahora	08-06-2020 13:05:33	Uc
Campos	22117789	37000	20-08-2019	Ahora	20-08-2019 11:18:54	07-01-2020 14:33:25
Campos	22117789	51000	20-08-2019	31-01-2020	07-01-2020 14:33:25	Uc
Campos	22117789	65000	01-02-2020	Ahora	07-01-2020 14:33:25	Uc
Torres	36876321	39800	01-05-2020	Ahora	27-04-2020 16:22:17	Uc

Representar los cambios necesarios tanto en los tiempos válidos, como en los tiempos de transacción, para las siguientes operaciones sobre la base de datos bitemporal:

a) Actualizar el sueldo del empleado Pérez a 63500, que será efectivo a partir del día 01-10-2020, representar la modificación, siendo el tiempo de actualización de la transacción ´14-09-2020 10:30:46' (marca de tiempo de la transacción).

Apellido	Dni	Sueldo	Tiv (Vst)	Tfv (Vet)	Tit (Tst)	Tft (Tet)
Pérez	11122233	48000	15-06-2020	Ahora	08-06-2020 13:05:33	**14-09-2020 10:30:46**
Pérez	11122233	48000	15-06-2020	30-09-2020	14-09-2020 10:30:46	Uc
Pérez	11122233	**63500**	**01-10-2020**	Ahora	**14-09-2020 10:30:46**	Uc

b) Eliminar el empleado Torres, que deja de prestar servicios el día 30-09-2020, siendo el tiempo de actualización de la transacción ´13-09-2020 12:25:31'.

Apellido	Dni	Sueldo	Tiv (Vst)	Tfv (Vet)	Tit (Tst)	Tft (Tet)
Torres	36876321	39800	01-05-2020	Ahora	27-04-2020 16:22:17	**13-09-2020 12:25:31**
Torres	36876321	39800	01-05-2020	**30-09-2020**	**13-09-2020 12:25:31**	Uc

c) Insertar el siguiente empleado, que comienza a prestar servicios desde el día de la fecha (considerar la fecha actual para Tiv), siendo la marca de tiempo de actualización de la transacción también la fecha actual y hora actual (expresada con el formato dd-mm-aaaa hh-mm-ss), esta fecha corresponde al día de desarrollo de este práctico el 14-09-2020 y una hora a elección.

Apellido	Dni	Sueldo
Ríos	24785930	71200

Apellido	Dni	Sueldo	Tiv (Vst)	Tfv (Vet)	Tit (Tst)	Tft (Tet)
Ríos	24785930	71200	**14-09-2020**	**Ahora**	**14-09-2020 21:22:37**	**Uc**

10. Dado el siguiente esquema de base de datos: Vehículos

Denominación	Versión	Precio	Año
Cruze	MT	915000	2019
Focus	Se Plus	950000	2019
Etios	XLS	1055000	2020
Civic	EXL	442000	2017
Suran	Feline	51600	2018

a) Realizar los cambios necesarios para representar la base de datos con el esquema de tiempo válido, teniendo en cuenta los siguientes valores de tiempos de inicio y final válidos para cada tupla:

	Tiv (Vst)	Tfc (vet)
Cruze	23-04-2019	Ahora
Focus	01-09-2019	Ahora
Etios	05-02-2020	30-11-2020
Civic	23-11-2017	Ahora
Suran	13-09-2018	Ahora

Denominación	Versión	Precio	Año	Tiv	Tfv
Cruze	MT	915000	2019	**23-04-2019**	**Ahora**
Focus	Se Plus	950000	2019	**01-09-2019**	**Ahora**
Etios	XLS	1055000	2020	**05-02-2020**	**30-11-2020**
Civic	EXL	442000	2017	**23-11-2017**	**Ahora**
Suran	Feline	51600	2018	**13-09-2018**	**Ahora**

b) Actualizar el precio del vehículo Cruze a 986500, que será efectivo a partir del día 01-10-2020, representar la modificación.

Denominación	Versión	Precio	Año	Tiv	Tfv
Cruze	MT	915000	2019	23-04-2019	**30-09-2020**
Cruze	MT	**986500**	2019	**01-10-2020**	**Ahora**

c) Insertar el siguiente vehículo, que comienza a estar vigente desde el día de la fecha (considerar la fecha actual para Tiv, tomamos la fecha para esta clase práctica):

Denominación	Versión	Precio	Año
Amarok	TrendLine	2177000	2020

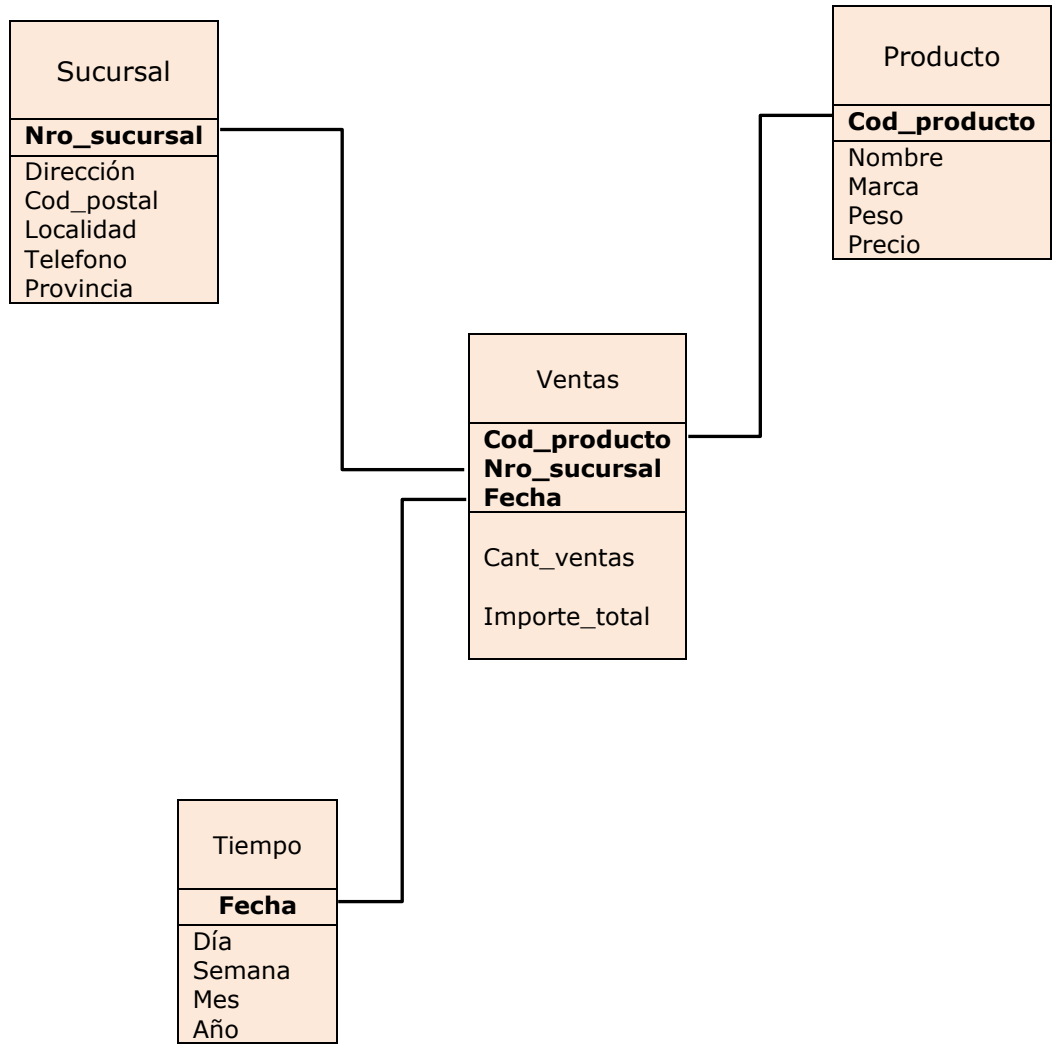
Denominación	Versión	Precio	Año	Tiv	Tfv
Cruze	MT	915000	2019	23-04-2019	Ahora
Cruze	MT	986500	2019	01-10-2020	Ahora
Focus	Se Plus	950000	2019	01-09-2019	Ahora
Etios	XLS	1055000	2020	05-02-2020	30-11-2020
Civic	EXL	442000	2017	23-11-2017	Ahora
Suran	Feline	51600	2018	13-09-2018	Ahora
Amarok	**TrendLine**	**2177000**	**2020**	**14-09-2020**	**Ahora**

Serie ejercicios prácticos 6
Almacenes de Datos

1) La cadena de supermercados Supermax, desea crear una base de datos de soporte a la decisión de las ventas realizadas de productos, en cuanto a cantidades realizadas y el importe total ya sea en forma diaria, semanal, mensual o anual.

Para cada producto se va a almacenar la siguiente información: código del producto, nombre, marca, peso y precio.
Además, se almacenarán los datos de cada sucursal: número de la sucursal, dirección, código postal, localidad, teléfono y provincia.
Considerar para la tabla tiempo, los atributos fecha, día, semana, mes y año.

a) Representar el esquema, según corresponda para el almacén de datos solicitado, especificando los atributos de la tabla de hechos y de las tablas de dimensiones necesarias para este caso.



2) La empresa de electrodomésticos "Megatone", desea crear una base de datos de soporte a la decisión de las ventas realizadas de sus productos, en cuanto a cantidades realizadas y el importe total ya sea en forma diaria, semanal, mensual, trimestral o anual.

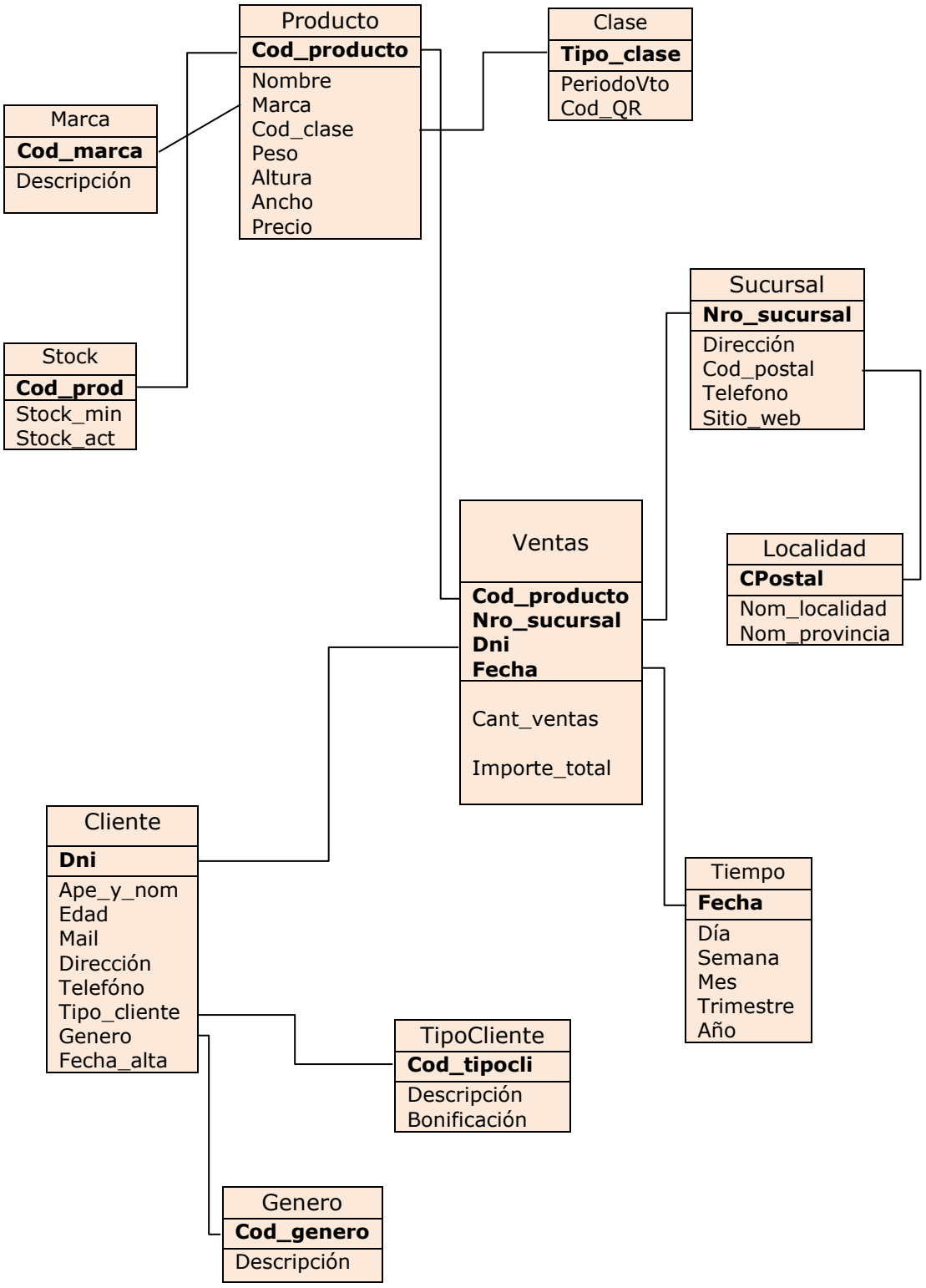
Para cada producto se va a almacenar la siguiente información: código del producto, nombre del producto, marca, clase, peso, altura, ancho y precio, teniendo en cuenta que para marca se almacenara en otra tabla el código de marca y su descripción. Con el código de producto, se accede a la tabla de stock de cada producto, donde se almacena el stock mínimo y stock actual. Con el atributo clase acceder a la tabla clases de producto, a fin de obtener el período de vencimiento y a su código QR.

Además, se almacenarán los datos de cada sucursal: número de la sucursal, dirección, código postal, teléfono y sitio web. Con el código postal obtener de otra tabla el nombre de la localidad y nombre de la provincia a la cual corresponde.

Para cada cliente, se registrarán el documento, nombre y apellido, edad, mail, dirección, teléfono, género y fecha de alta. En cuanto al código de género se almacenará en otra tabla el código y la descripción del género. Con el documento se accede a la tabla de los pedidos de cada cliente, a fin de obtener la orden del pedido, su descripción, cantidad y fecha del pedido.

Considerar para la tabla tiempo, los atributos fecha, día, semana, mes, año, trimestre y día festivo.

a) Representar el esquema, según corresponda para el almacén de datos solicitado, especificando los atributos de la tabla de hechos y de las tablas de dimensiones necesarias para este caso.


Producto
Cod_producto
Nombre
Marca
Cod_clase
Peso
Altura
Ancho
Precio
Clase
Tipo_clase
PeriodoVto
Cod_QR
Marca
Cod_marca
Descripción
Stock
Cod_prod
Stock_min
Stock_act
Sucursal
Nro_sucursal
Dirección
Cod_postal
Telefono
Sitio_web
Localidad
CPostal
Nom_localidad
Nom_provincia
Ventas
Cod_producto
Nro_sucursal
Dni
Fecha
Cant_ventas
Importe_total
Cliente
Dni
Ape_y_nom
Edad
Mail
Dirección
Telefóno
Tipo_cliente
Genero
Fecha_alta
Tiempo
Fecha
Día
Semana
Mes
Trimestre
Año
TipoCliente
Cod_tipocli
Descripción
Bonificación
Genero
Cod_genero
Descripción

3) Una empresa de telefonía desea crear una base de datos de soporte a la decisión de las ventas realizadas de sus servicios, en cuanto a cantidades realizadas y el importe total ya sea en forma diaria, semanal, mensual, trimestral o anual, teniendo en cuenta cada una de sus sucursales y nómina de clientes registrados.

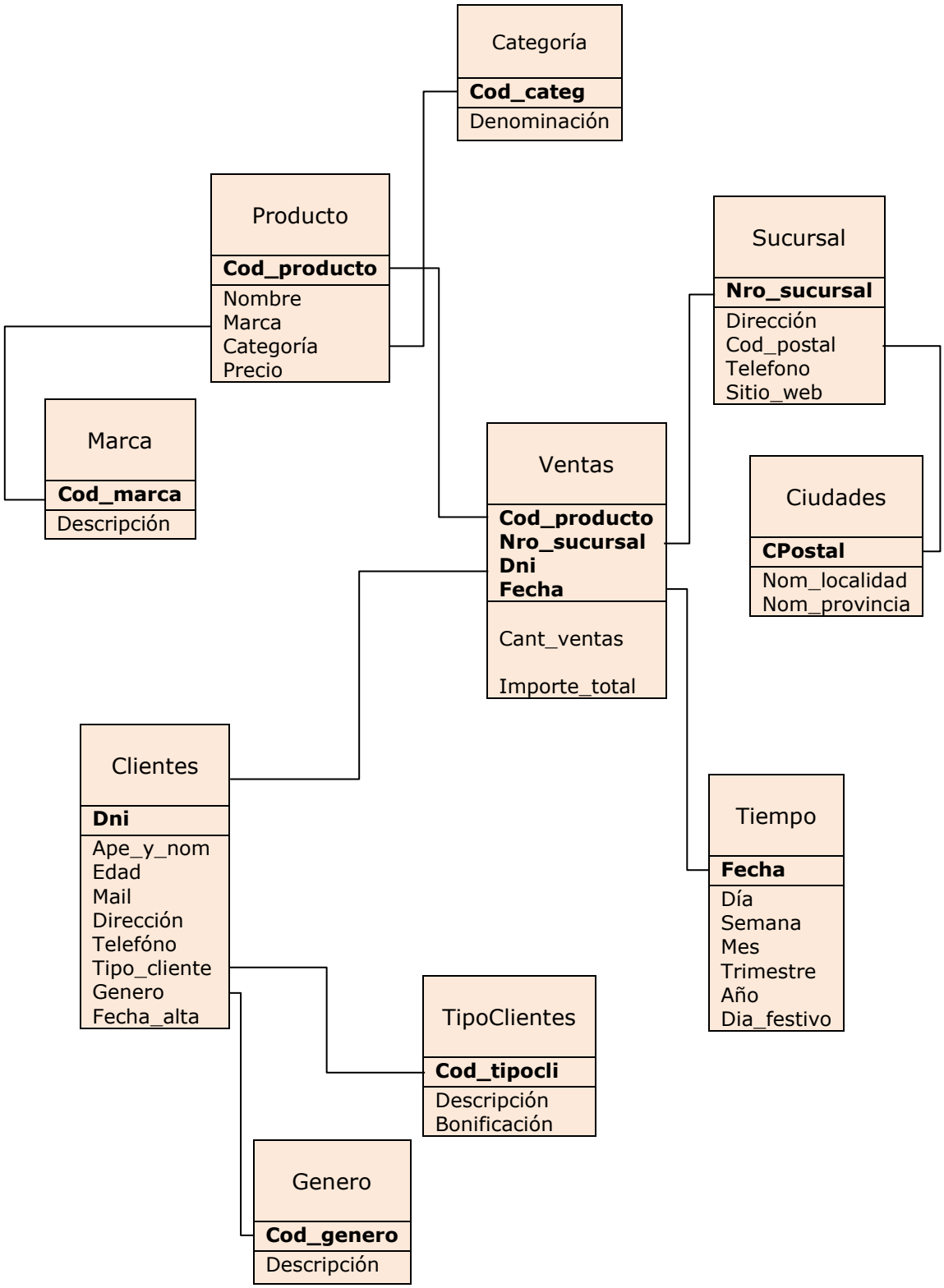
Para cada producto se va a almacenar la siguiente información: código de producto, nombre, marca, categoría y precio, teniendo en cuenta que para el atributo marca se guardará en otra tabla el código de marca y su descripción, y también para la categoría se registrará en otra tabla el código de categoría y su denominación.

Además, se almacenarán los datos de cada sucursal: número de la sucursal, dirección, código postal, teléfono y sitio web. Con el código postal obtener de otra tabla el nombre de la localidad y nombre de la provincia a la cual corresponde.

Para cada cliente, se registrarán el documento, nombre y apellido, edad, mail, dirección, teléfono, tipo cliente, género y fecha de alta. En cuanto al código de género se guardarán en otra tabla el código de género y la descripción, y respecto al tipo de cliente se almacenarán en otra tabla el código del tipo de cliente, su descripción y la bonificación.

Considerar para la tabla tiempo, los atributos: fecha, día, semana, mes, año, trimestre y día festivo.

a) Representar el esquema, según corresponda para el almacén de datos solicitado, especificando los atributos de la tabla de hechos y de las tablas de dimensiones necesarias para este caso.


Categoría
Cod_categ
Denominación
Producto
Cod_producto
Nombre
Marca
Categoría
Precio
Sucursal
Nro_sucursal
Dirección
Cod_postal
Telefono
Sitio_web
Marca
Cod_marca
Descripción
Ventas
Cod_producto
Nro_sucursal
Dni
Fecha
Cant_ventas
Importe_total
Ciudades
CPostal
Nom_localidad
Nom_provincia
Clientes
Dni
Ape_y_nom
Edad
Mail
Dirección
Telefóno
Tipo_cliente
Genero
Fecha_alta
Tiempo
Fecha
Día
Semana
Mes
Trimestre
Año
Dia_festivo
TipoClientes
Cod_tipocli
Descripción
Bonificación
Genero
Cod_genero
Descripción

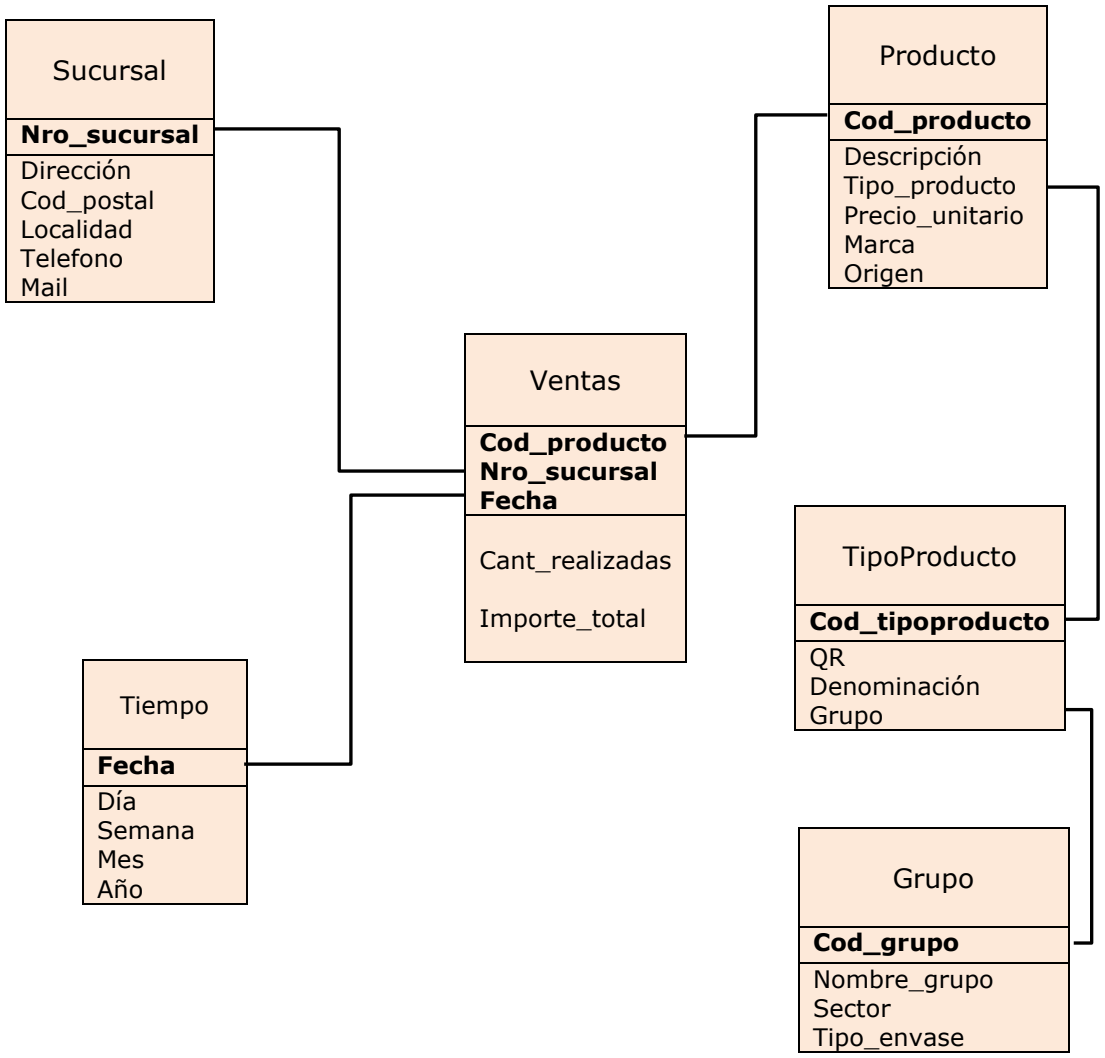
4) Un hipermercado desea crear una base de datos de soporte a la decisión de los productos comercializados para cada una de sus sucursales distribuidas en distintas provincias, en cuanto a cantidades realizadas y el importe total ya sea en forma diaria, semanal, mensual o anual.

Para cada producto se va a almacenar la siguiente información: código del producto, descripción, tipo de producto, precio unitario, marca y origen. En cuanto al tipo de producto se registrarán en otra tabla el código del tipo de producto, código QR, su denominación y código de grupo al cual corresponde. Respecto al código de grupo, se almacenarán en otra tabla el código de grupo, nombre del grupo, sector y tipo de envase.

Además, se almacenarán los datos de cada sucursal: número de la sucursal, dirección, código postal, localidad, teléfono y mail.

Considerar para la tabla tiempo, los atributos: fecha, día, semana, mes y año.

a) Realizar el esquema, según corresponda para el almacén de datos solicitado, especificando los atributos de las tablas de hechos y de dimensiones necesarias para este caso.

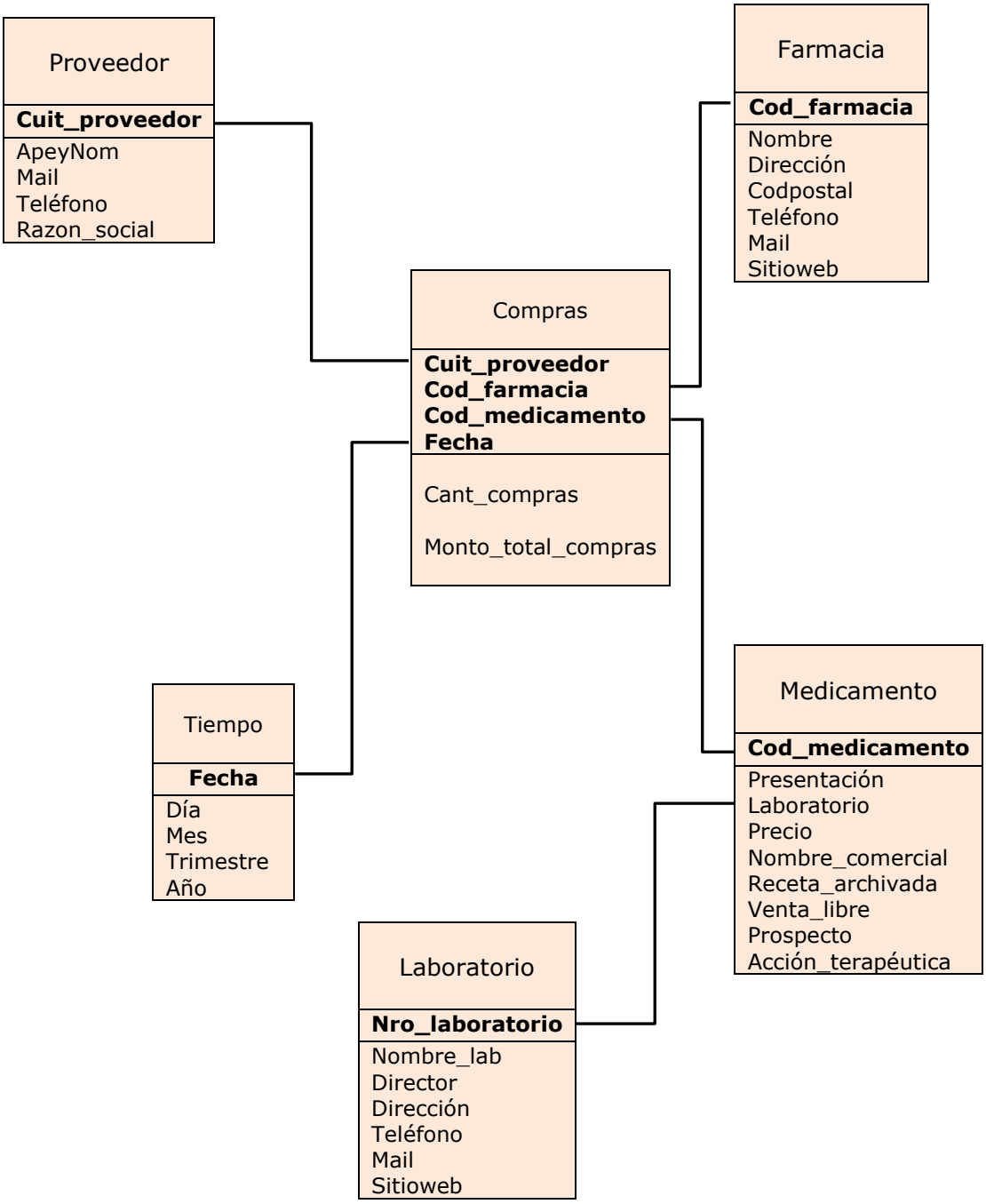

Sucursal
Nro_sucursal
Dirección
Cod_postal
Localidad
Telefono
Mail
Producto
Cod_producto
Descripción
Tipo_producto
Precio_unitario
Marca
Origen
Ventas
Cod_producto
Nro_sucursal
Fecha
Cant_realizadas
Importe_total
TipoProducto
Cod_tipoproducto
QR
Denominación
Grupo
Tiempo
Fecha
Día
Semana
Mes
Año
Grupo
Cod_grupo
Nombre_grupo
Sector
Tipo_envase

Diseño y armado de almacén de datos

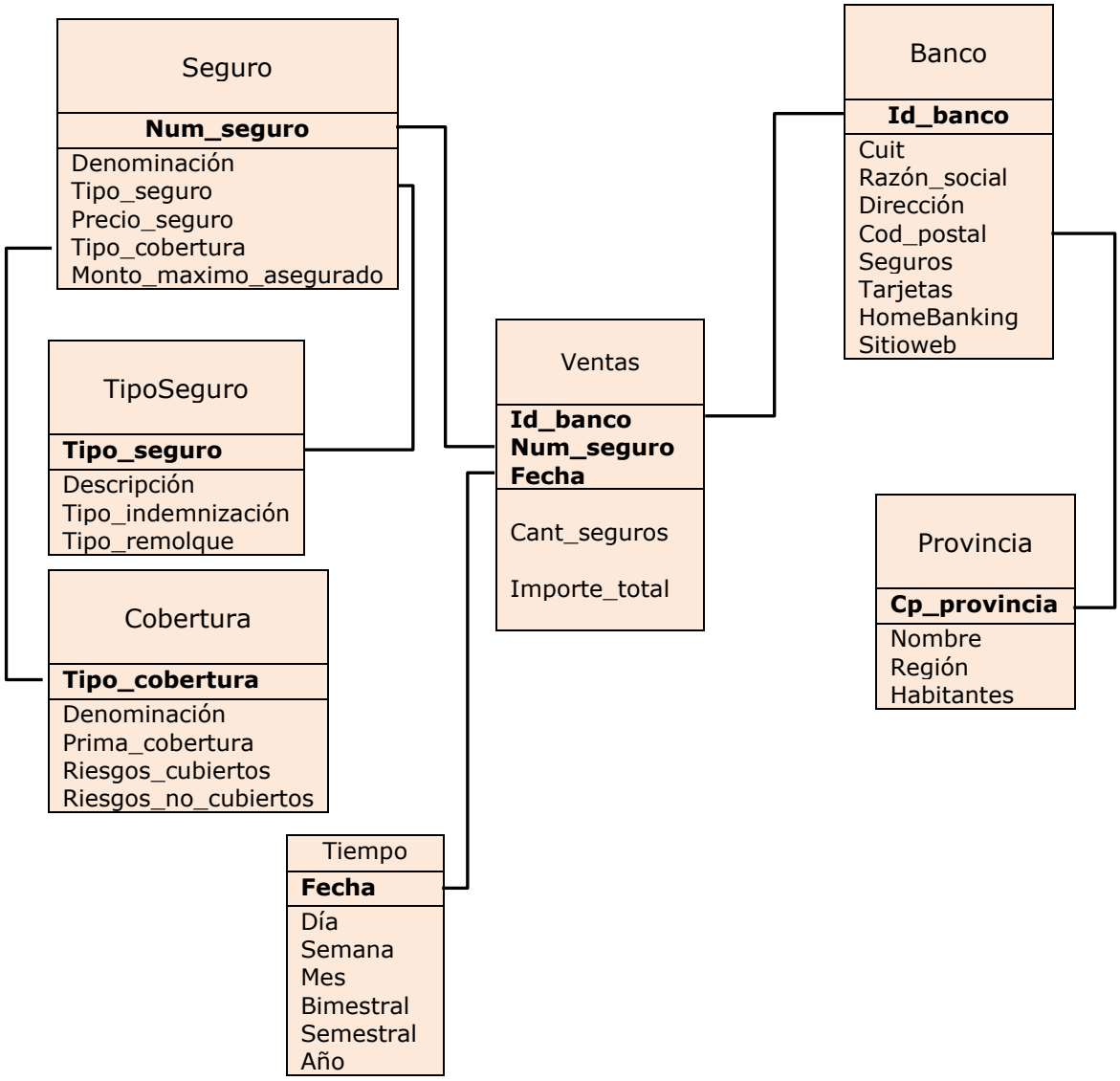
5) Teniendo en cuenta los pasos a seguir para el diseño de un almacén de datos, realice su armado:

- Paso 1: Elegir un proceso de la organización para modelar.
- Paso 2: Decidir el gránulo (nivel de detalle) de la información que se desea almacenar.
- Paso 3: Identificar las dimensiones necesarias para el proceso, determinar cuáles son los atributos relevantes de cada dimensión y las jerarquías naturales que se dan entre ellos.
- Paso 4: Decidir la información a almacenar sobre el proceso (información sobre la actividad que se almacenara en cada tupla de la tabla de Hechos).

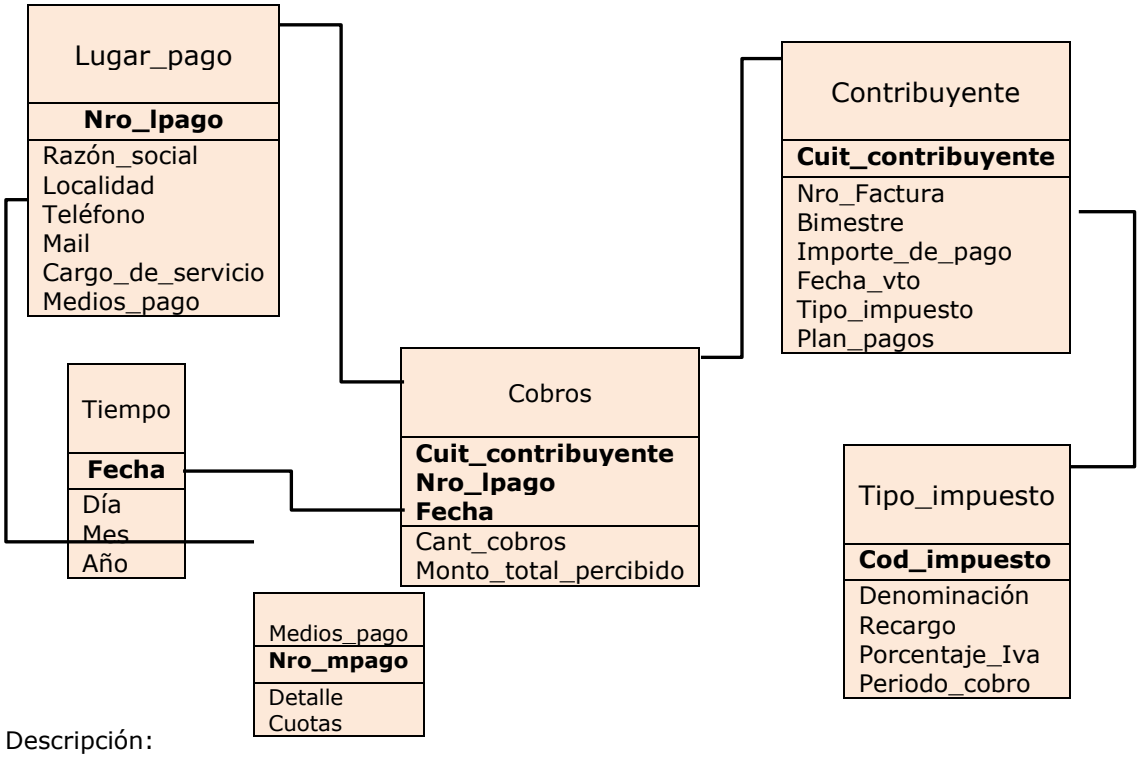
Considerando:
- Organización: una red farmacéutica.
- Proceso: compras de medicamentos de cada farmacia a las droguerías proveedoras.
- Gránulo: se requiere almacenar, las compras diarias de medicamentos de cada farmacia, en cuanto a las cantidades e importe total de las compras en el día.
- Dimensiones: considerar al menos las de tiempo (diario, mensual, trimestral, anual), farmacias (sus datos identificatorios, ubicación, localidad, etc.), proveedores y medicamentos (código, descripción, presentación, precio, entre otros).


Proveedor
Cuit_proveedor
ApeyNom
Mail
Teléfono
Razon_social
Farmacia
Cod_farmacia
Nombre
Dirección
Codpostal
Teléfono
Mail
Sitioweb
Compras
Cuit_proveedor
Cod_farmacia
Cod_medicamento
Fecha
Cant_compras
Monto_total_compras
Tiempo
Fecha
Día
Mes
Trimestre
Año
Medicamento
Cod_medicamento
Presentación
Laboratorio
Precio
Nombre_comercial
Receta_archivada
Venta_libre
Prospecto
Acción_terapéutica
Laboratorio
Nro_laboratorio
Nombre_lab
Director
Dirección
Teléfono
Mail
Sitioweb

6) Diseñe y arme el almacén de datos, considerando a una entidad bancaria que requiere registrar las ventas semanales de los diferentes tipos de seguros que comercializa (para la vivienda, personales, vehículos, etc.), acerca de cantidad de cada rubro y el importe total de cada uno de ellos.
Las dimensiones posibles pueden ser: tiempo, banco, seguros, provincia, tipos de seguros y tipos o clases de coberturas de los seguros.
Identifique y represente el esquema elegido (estrella o copo de nieve), de acuerdo a las dimensiones y tabla de hechos determinados.



7) Diseñe y arme un almacén de datos, optando a su criterio la unidad de organización de trabajo y los demás componentes necesarios para el caso de estudio, prever al menos 4 dimensiones para el esquema elegido.

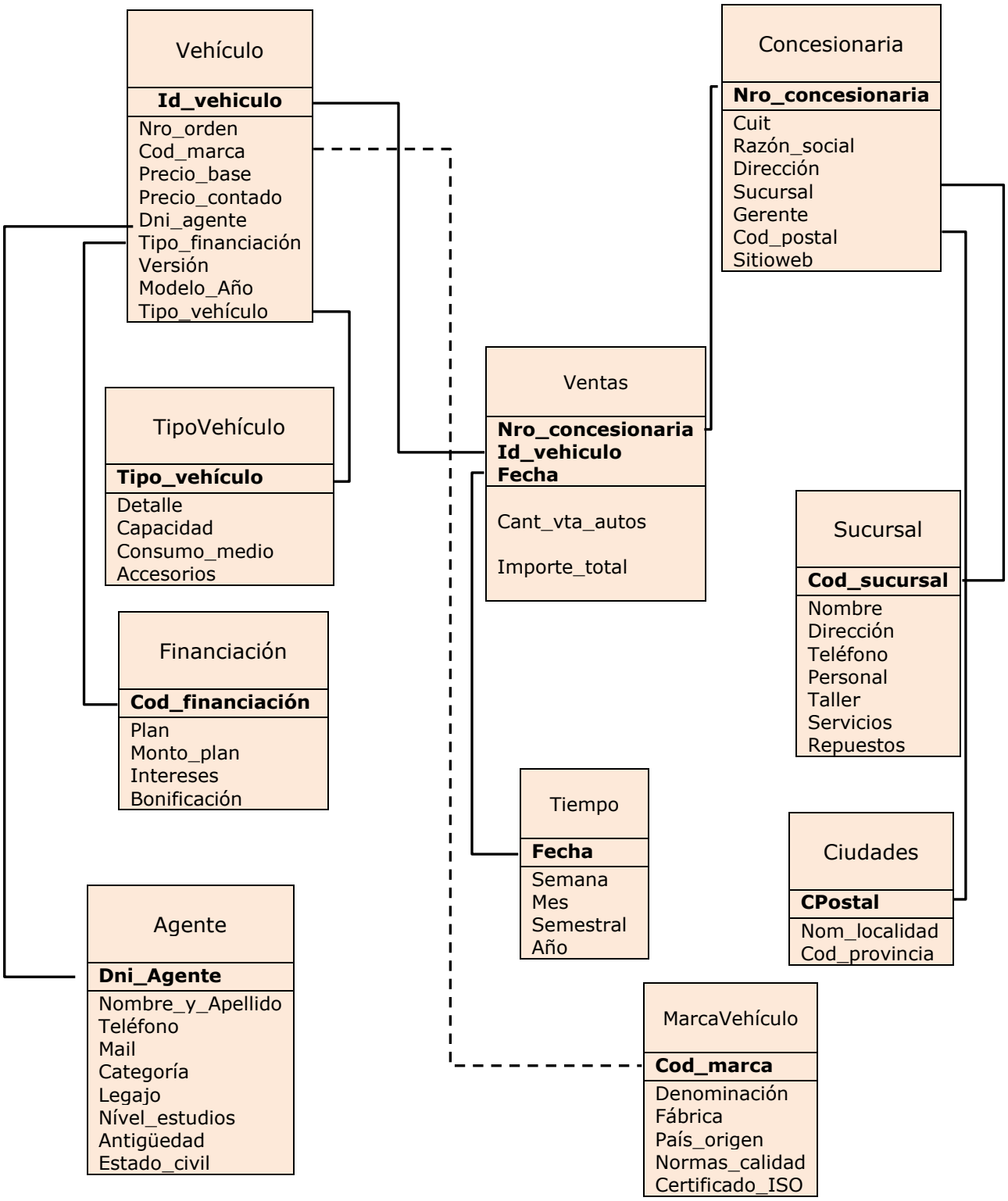


Descripción:
Organización: Entidad recaudadora de impuestos Afip.
Proceso: Cantidad de impuestos cobrados y el monto total de la cobranza.
Gránulo: se requiere obtener la cantidad de cobros de impuestos y el monto total, en forma diaria, mensual y anual.
Dimensiones: Tiempo, Lugar de Pago, Contribuyente, Tipo Impuesto y Medios pago.

8) Teniendo en cuenta los pasos a seguir para el diseño de un almacén de datos, realice su armado:

- Paso 1: Elegir un proceso de la organización para modelar.
- Paso 2: Decidir el gránulo (nivel de detalle) de la información que se desea almacenar.
- Paso 3: Identificar las dimensiones necesarias para el proceso, determinar cuáles son los atributos relevantes de cada dimensión y las jerarquías naturales que se dan entre ellos.
- Paso 4: Decidir la información a almacenar sobre el proceso (información sobre la actividad que se almacenara en cada tupla de la tabla de Hechos).

Considerando:
- Organización: una concesionaria de autos.
- Proceso: ventas de vehículos de cada sucursal en el país.
- Gránulo: se requiere almacenar, las ventas mensuales de vehículos de cada sucursal de la concesionaria, en cuanto a las cantidades e importe total de dichas ventas.
- Dimensiones: considerar al menos las de tiempo (semanal, mensual, semestral, anual), sucursales (sus datos identificatorios, dirección, localidad, medios de pagos, etc.), agentes comercializadores (Nombre y apellido, mail, teléfono) y vehículos (Marca, modelo, año, precio, tipo de vehículo (familiar, camioneta, sedán), versiones, entre otros).


Vehículo
Id_vehiculo
Nro_orden
Cod_marca
Precio_base
Precio_contado
Dni_agente
Tipo_financiación
Versión
Modelo_Año
Tipo_vehículo
Concesionaria
Nro_concesionaria
Cuit
Razón_social
Dirección
Sucursal
Gerente
Cod_postal
Sitioweb
Ventas
Nro_concesionaria
Id_vehiculo
Fecha
Cant_vta_autos
Importe_total
TipoVehículo
Tipo_vehículo
Detalle
Capacidad
Consumo_medio
Accesorios
Sucursal
Cod_sucursal
Nombre
Dirección
Teléfono
Personal
Taller
Servicios
Repuestos
Financiación
Cod_financiación
Plan
Monto_plan
Intereses
Bonificación
Tiempo
Fecha
Semana
Mes
Semestral
Año
Ciudades
CPostal
Nom_localidad
Cod_provincia
Agente
Dni_Agente
Nombre_y_Apellido
Teléfono
Mail
Categoría
Legajo
Nível_estudios
Antigüedad
Estado_civil
MarcaVehículo
Cod_marca
Denominación
Fábrica
País_origen
Normas_calidad
Certificado_ISO

Serie Ejercicios Prácticos 7
OLAP

1) Considerando la siguiente consulta para el análisis de datos:

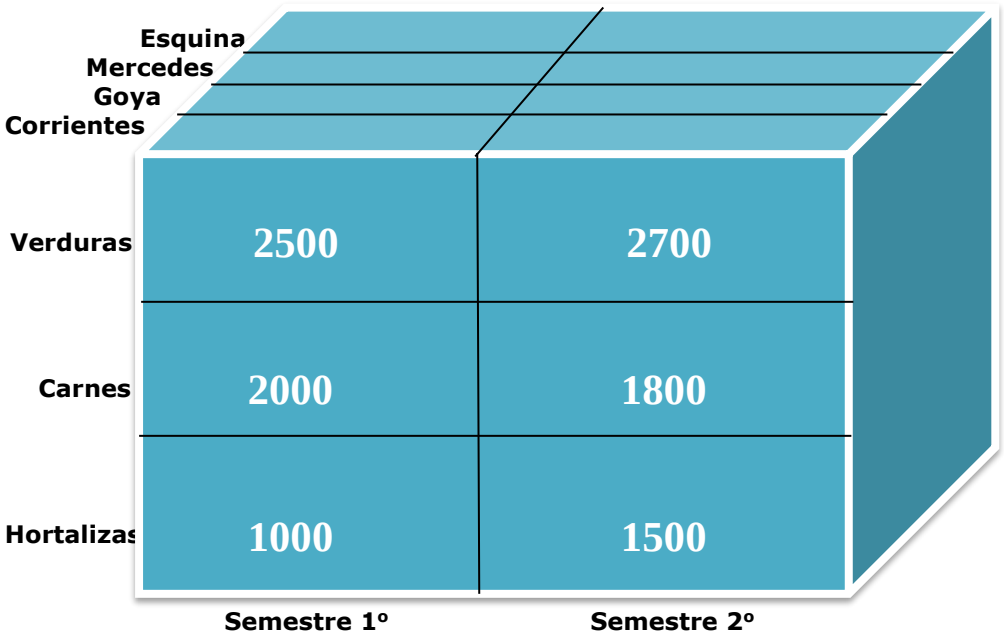
Categoría	Semestre	Ciudad	Lotes vendidos
Verduras	Primero	Corrientes	1250
Verduras	Primero	Corrientes	1250
Verduras	Segundo	Corrientes	1350
Verduras	Segundo	Corrientes	1350
Verduras	Primero	Goya	900
Verduras	Primero	Goya	900
Verduras	Segundo	Goya	900
Verduras	Segundo	Goya	900
Verduras	Primero	Mercedes	1100
Verduras	Primero	Mercedes	1100
Verduras	Segundo	Mercedes	900
Verduras	Segundo	Mercedes	900
Verduras	Primero	Esquina	1100
Verduras	Primero	Esquina	1100
Verduras	Segundo	Esquina	800
Verduras	Segundo	Esquina	800
Carnes	Primero	Corrientes	1000
Carnes	Primero	Corrientes	1000
Carnes	Segundo	Corrientes	900
Carnes	Segundo	Corrientes	900
Carnes	Primero	Goya	1000
Carnes	Primero	Goya	1000
Carnes	Segundo	Goya	900
Carnes	Segundo	Goya	900
Carnes	Primero	Mercedes	1000
Carnes	Primero	Mercedes	1000
Carnes	Segundo	Mercedes	900
Carnes	Segundo	Mercedes	900
Carnes	Primero	Esquina	950
Carnes	Primero	Esquina	950
Carnes	Segundo	Esquina	750
Carnes	Segundo	Esquina	750
Hortalizas	Primero	Corrientes	500
Hortalizas	Primero	Corrientes	500
Hortalizas	Segundo	Corrientes	750
Hortalizas	Segundo	Corrientes	750
Hortalizas	Primero	Goya	1000
Hortalizas	Primero	Goya	1000
Hortalizas	Segundo	Goya	900
Hortalizas	Segundo	Goya	900
Hortalizas	Primero	Mercedes	1100
Hortalizas	Primero	Mercedes	1100
Hortalizas	Segundo	Mercedes	900
Hortalizas	Segundo	Mercedes	900
Hortalizas	Primero	Esquina	700
Hortalizas	Primero	Esquina	700
Hortalizas	Segundo	Esquina	750
Hortalizas	Segundo	Esquina	750

a) Disgregar en 2 nuevos grupos cada Categoría/Semestre/Ciudad de la consulta original, considerando que las ciudades se conforman por área urbana y suburbana, y los lotes vendidos se distribuyen en igual proporción.

Función Disgregación (Drill)

Categoría	Semestre	Ciudad	Área	Lotes vendidos
Verduras	Primero	Corrientes	Urbana	1250
Verduras	Primero	Corrientes	Suburbana	1250
Verduras	Segundo	Corrientes	Urbana	1350
Verduras	Segundo	Corrientes	Suburbana	1350
Verduras	Primero	Goya	Urbana	900
Verduras	Primero	Goya	Suburbana	900
Verduras	Segundo	Goya	Urbana	900
Verduras	Segundo	Goya	Suburbana	900
Verduras	Primero	Mercedes	Urbana	1100
Verduras	Primero	Mercedes	Suburbana	1100
Verduras	Segundo	Mercedes	Urbana	900
Verduras	Segundo	Mercedes	Suburbana	900
Verduras	Primero	Esquina	Urbana	1100
Verduras	Primero	Esquina	Suburbana	1100
Verduras	Segundo	Esquina	Urbana	800
Verduras	Segundo	Esquina	Suburbana	800
Carnes	Primero	Corrientes	Urbana	1000
Carnes	Primero	Corrientes	Suburbana	1000
Carnes	Segundo	Corrientes	Urbana	900
Carnes	Segundo	Corrientes	Suburbana	900
Carnes	Primero	Goya	Urbana	1000
Carnes	Primero	Goya	Suburbana	1000
Carnes	Segundo	Goya	Urbana	900
Carnes	Segundo	Goya	Suburbana	900
Carnes	Primero	Mercedes	Urbana	1000
Carnes	Primero	Mercedes	Suburbana	1000
Carnes	Segundo	Mercedes	Urbana	900
Carnes	Segundo	Mercedes	Suburbana	900
Carnes	Primero	Esquina	Urbana	950
Carnes	Primero	Esquina	Suburbana	950
Carnes	Segundo	Esquina	Urbana	750
Carnes	Segundo	Esquina	Suburbana	750
Hortalizas	Primero	Corrientes	Urbana	500
Hortalizas	Primero	Corrientes	Suburbana	500
Hortalizas	Segundo	Corrientes	Urbana	750
Hortalizas	Segundo	Corrientes	Suburbana	750
Hortalizas	Primero	Goya	Urbana	1000
Hortalizas	Primero	Goya	Suburbana	1000
Hortalizas	Segundo	Goya	Urbana	900
Hortalizas	Segundo	Goya	Suburbana	900
Hortalizas	Primero	Mercedes	Urbana	1100
Hortalizas	Primero	Mercedes	Suburbana	1100
Hortalizas	Segundo	Mercedes	Urbana	900
Hortalizas	Segundo	Mercedes	Suburbana	900
Hortalizas	Primero	Esquina	Urbana	700
Hortalizas	Primero	Esquina	Suburbana	700
Hortalizas	Segundo	Esquina	Urbana	750
Hortalizas	Segundo	Esquina	Suburbana	750

b) Presente matricialmente los datos seleccionados.



c) Realice las operaciones necesarias para expresar la consulta obtenida en el punto a) en forma anual.

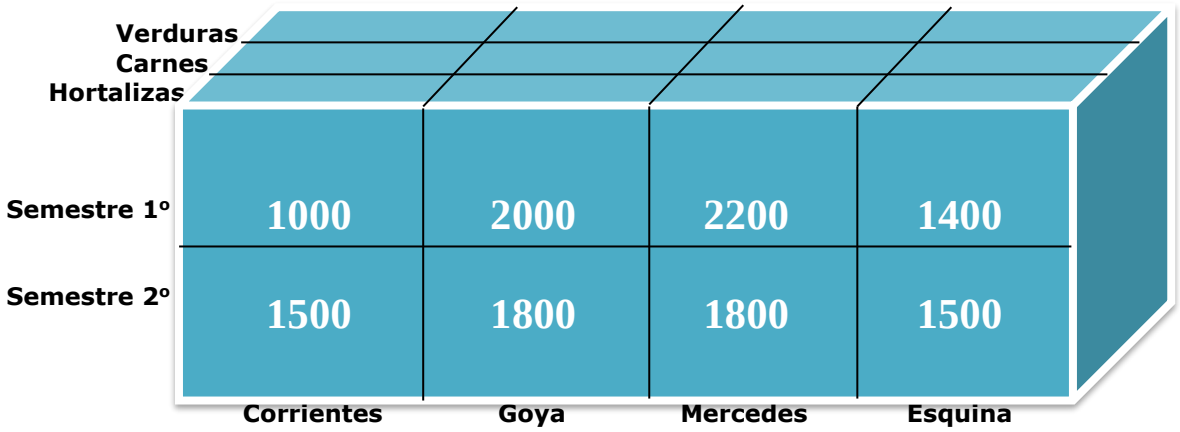
Categoría	Anual	Ciudad	Área	Lotes vendidos
Verduras	Anual	Corrientes	Urbana	2600
Verduras	Anual	Corrientes	Suburbana	2600
Verduras	Anual	Goya	Urbana	1800
Verduras	Anual	Goya	Suburbana	1800
Verduras	Anual	Mercedes	Urbana	2000
Verduras	Anual	Mercedes	Suburbana	2000
Verduras	Anual	Esquina	Urbana	1900
Verduras	Anual	Esquina	Suburbana	1900
Carnes	Anual	Corrientes	Urbana	1900
Carnes	Anual	Corrientes	Suburbana	1900
Carnes	Anual	Goya	Urbana	1900
Carnes	Anual	Goya	Suburbana	1900
Carnes	Anual	Mercedes	Urbana	1900
Carnes	Anual	Mercedes	Suburbana	1900
Carnes	Anual	Esquina	Urbana	1700
Carnes	Anual	Esquina	Suburbana	1700
Hortalizas	Anual	Corrientes	Urbana	1250
Hortalizas	Anual	Corrientes	Suburbana	1250
Hortalizas	Anual	Goya	Urbana	1900
Hortalizas	Anual	Goya	Suburbana	1900
Hortalizas	Anual	Mercedes	Urbana	2000
Hortalizas	Anual	Mercedes	Suburbana	2000
Hortalizas	Anual	Esquina	Urbana	1450
Hortalizas	Anual	Esquina	Suburbana	1450

d) ¿Qué operador es necesario utilizar para obtener a partir de a) una consulta por Categoría/Lotes vendidos? Resuélvalo.
El operador de manipulación de consultas que debemos de aplicar es **Roll**, en particular **Roll-across**, para condensar las unidades por Categoría/Lotes vendidos, resultando:

Categoría	Lotes vendidos
Verduras	16600
Carnes	14800
Hortalizas	13200

Se debe de acumular para cada Categoría evaluada, las unidades de lotes vendidos, sin considerar en este punto a que ciudad, área o período pertenece.

e) Represente el punto b) luego de utilizar el operador tipo **Pívot** Luego de aplicar el operador de tipo rotación **Pívot** al cubo del punto **b),** obtenemos:



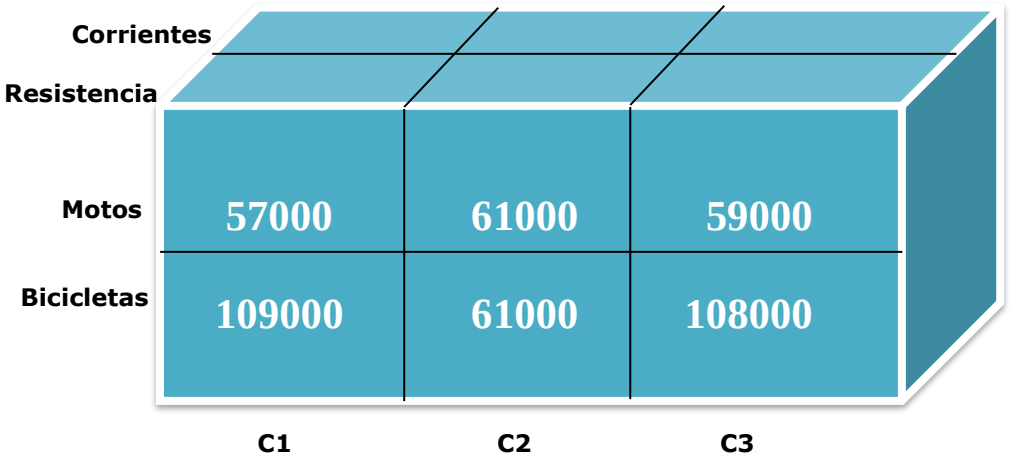
2) Considerando la siguiente consulta para el análisis de datos:

Vehículo	Cuatrimestre	Unidades Vendidas
Motos	C1	114000
Motos	C2	122000
Motos	C3	118000
Bicicletas	C1	218000
Bicicletas	C2	122000
Bicicletas	C3	216000

a) Disgregar en 2 nuevos grupos cada Vehículo/Cuatrimestre de la consulta original, considerando que las ciudades de Corrientes y Resistencia conforman el universo estudiado, y las unidades vendidas se distribuyen en igual proporción en dichas ciudades.

Vehículo	Cuatrimestre	Ciudad	Unidades vendidas
Motos	C1	Corrientes	57000
Motos	C1	Resistencia	57000
Motos	C2	Corrientes	61000
Motos	C2	Resistencia	61000
Motos	C3	Corrientes	59000
Motos	C3	Resistencia	59000
Bicicletas	C1	Corrientes	109000
Bicicletas	C1	Resistencia	109000
Bicicletas	C2	Corrientes	61000
Bicicletas	C2	Resistencia	61000
Bicicletas	C3	Corrientes	108000
Bicicletas	C3	Resistencia	108000

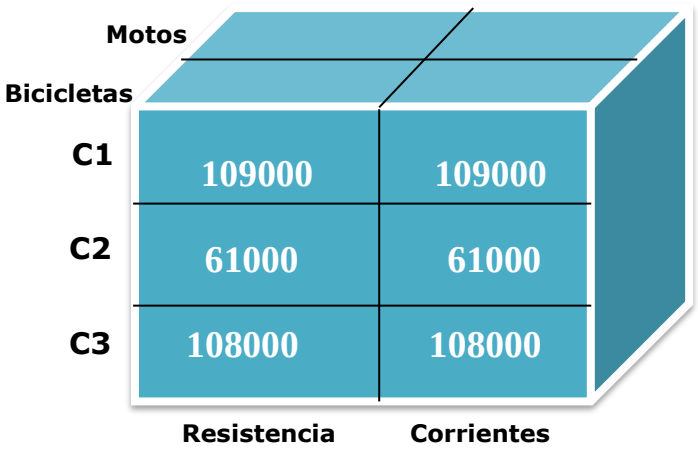
b) Presente matricialmente los datos seleccionados.



c) ¿Qué operador es necesario utilizar para obtener a partir de a) una consulta por Vehículo/Unidades vendidas? Resuélvalo. Operador **Roll** (agrupar):

Vehículo	Unidades vendidas
Motos	354000
Bicicletas	556000

d) Represente el punto b) luego de utilizar el operador **Pivot**



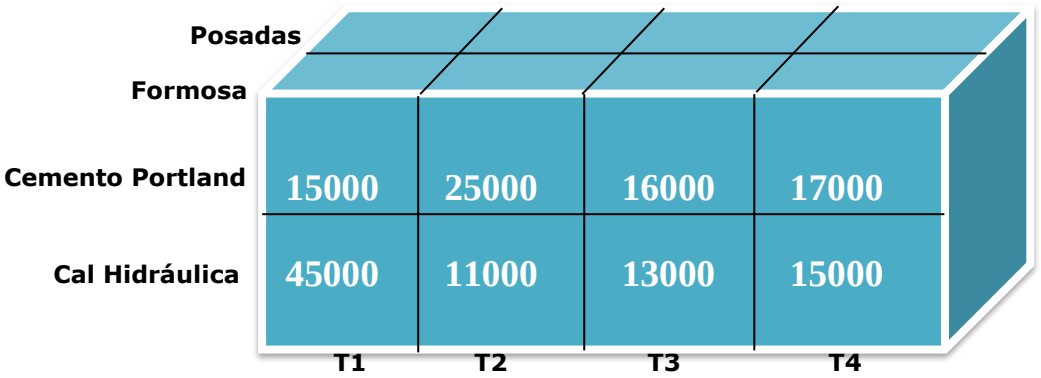
3) Considerando la siguiente consulta para el análisis de datos:

Material	Trimestre	Bolsas vendidas
Cemento Portland	T1	30000
Cemento Portland	T2	50000
Cemento Portland	T3	32000
Cemento Portland	T4	34000
Cal Hidráulica	T1	90000
Cal Hidráulica	T2	22000
Cal Hidráulica	T3	26000
Cal Hidráulica	T4	30000

a) Disgregar en 2 nuevos grupos cada Material/Trimestre de la consulta original, considerando que las ciudades de Formosa y Posadas conforman el universo estudiado, y las bolsas vendidas se distribuyen en igual proporción en dichas ciudades.

Material	Trimestre	Ciudad	Bolsas vendidas
Cemento Portland	T1	Formosa	15000
Cemento Portland	T1	Posadas	15000
Cemento Portland	T2	Formosa	25000
Cemento Portland	T2	Posadas	25000
Cemento Portland	T3	Formosa	16000
Cemento Portland	T3	Posadas	16000
Cemento Portland	T4	Formosa	17000
Cemento Portland	T4	Posadas	17000
Cal Hidráulica	T1	Formosa	45000
Cal Hidráulica	T1	Posadas	45000
Cal Hidráulica	T2	Formosa	11000
Cal Hidráulica	T2	Posadas	11000
Cal Hidráulica	T3	Formosa	13000
Cal Hidráulica	T4	Posadas	13000
Cal Hidráulica	T4	Formosa	15000
Cal Hidráulica	T4	Posadas	15000

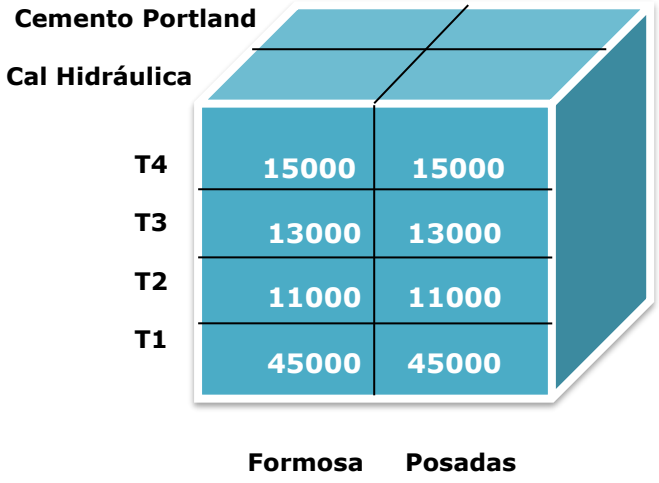
b) Presente matricialmente los datos seleccionados.



c) Realice las operaciones necesarias para expresar la consulta obtenida en el punto a) en semestres.

Material	Semestre	Ciudad	Bolsas vendidas
Cemento Portland	S1	Formosa	40000
Cemento Portland	S1	Posadas	40000
Cemento Portland	S2	Formosa	33000
Cemento Portland	S2	Posadas	33000
Cal Hidráulica	S1	Formosa	56000
Cal Hidráulica	S1	Posadas	56000
Cal Hidráulica	S2	Formosa	28000
Cal Hidráulica	S2	Posadas	28000

d) Represente el punto b) luego de utilizar el operador tipo **Pivot**.



4) Considerando la siguiente consulta para el análisis de datos:

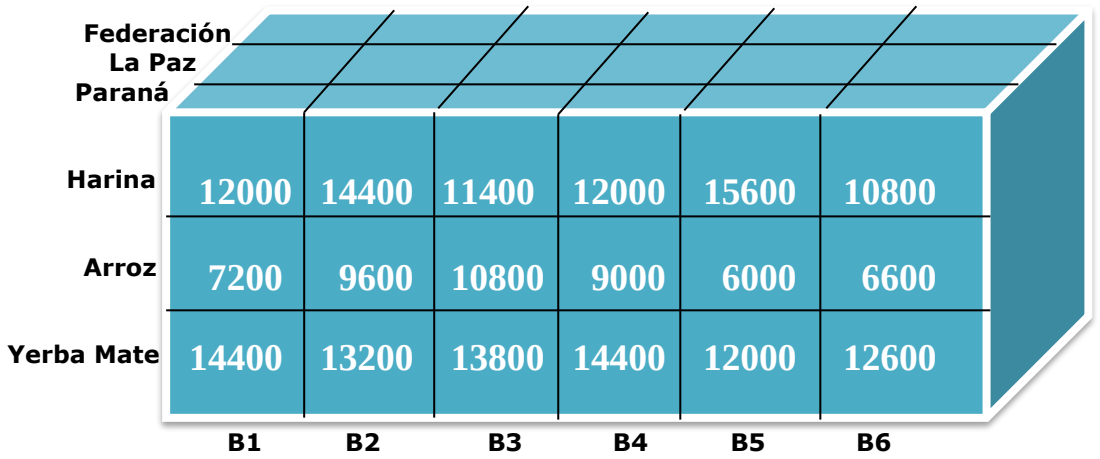
Producto	Bimestre	Kilogramos Vendidos
Harina	B1	20000
Harina	B2	24000
Harina	B3	19000
Harina	B4	20000
Harina	B5	26000
Harina	B6	18000
Arroz	B1	12000
Arroz	B2	16000
Arroz	B3	18000
Arroz	B4	15000
Arroz	B5	10000
Arroz	B6	11000
Yerba mate	B1	24000
Yerba mate	B2	22000
Yerba mate	B3	23000
Yerba mate	B4	24000
Yerba mate	B5	20000
Yerba mate	B6	21000

a) Disgregar en 3 nuevos grupos cada Producto/Bimestre de la consulta original, considerando que las ciudades de Paraná, La Paz y Federación conforman el universo estudiado, y los kilogramos vendidos se distribuyen en las proporciones de 3/5, 1/5 y 1/5 respectivamente en cada ciudad.

Producto	Bimestre	Ciudad	Kilogramos vendidos
Harina	B1	Paraná	12000
Harina	B1	La Paz	4000
Harina	B1	Federación	4000
Harina	B2	Paraná	14400
Harina	B2	La Paz	4800
Harina	B2	Federación	4800
Harina	B3	Paraná	11400
Harina	B3	La Paz	3800
Harina	B3	Federación	3800
Harina	B4	Paraná	12000
Harina	B4	La Paz	4000
Harina	B4	Federación	4000
Harina	B5	Paraná	15600
Harina	B5	La Paz	5200
Harina	B5	Federación	5200
Harina	B6	Paraná	10800
Harina	B6	La Paz	3600
Harina	B6	Federación	3600
Arroz	B1	Paraná	7200
Arroz	B1	La Paz	2400
Arroz	B1	Federación	2400
Arroz	B2	Paraná	9600
Arroz	B2	La Paz	3200
Arroz	B2	Federación	3200
Arroz	B3	Paraná	10800
Arroz	B3	La Paz	3600
Arroz	B3	Federación	3600
Arroz	B4	Paraná	9000
Arroz	B4	La Paz	3000
Arroz	B4	Federación	3000

Arroz	B5	Paraná	6000
Arroz	B5	La Paz	2000
Arroz	B5	Federación	2000
Arroz	B6	Paraná	6600
Arroz	B6	La Paz	2200
Arroz	B6	Federación	2200
Yerba mate	B1	Paraná	14400
Yerba mate	B1	La Paz	4800
Yerba mate	B1	Federación	4800
Yerba mate	B2	Paraná	13200
Yerba mate	B2	La Paz	4400
Yerba mate	B2	Federación	4400
Yerba mate	B3	Paraná	13800
Yerba mate	B3	La Paz	4600
Yerba mate	B3	Federación	4600
Yerba mate	B4	Paraná	14400
Yerba mate	B4	La Paz	4800
Yerba mate	B4	Federación	4800
Yerba mate	B5	Paraná	12000
Yerba mate	B5	La Paz	4000
Yerba mate	B5	Federación	4000
Yerba mate	B6	Paraná	12600
Yerba mate	B6	La Paz	4200
Yerba mate	B6	Federación	4200

b) Presente matricialmente los datos seleccionados.



c) Realice las operaciones necesarias para expresar la consulta obtenida en el punto a) en cuatrimestres.

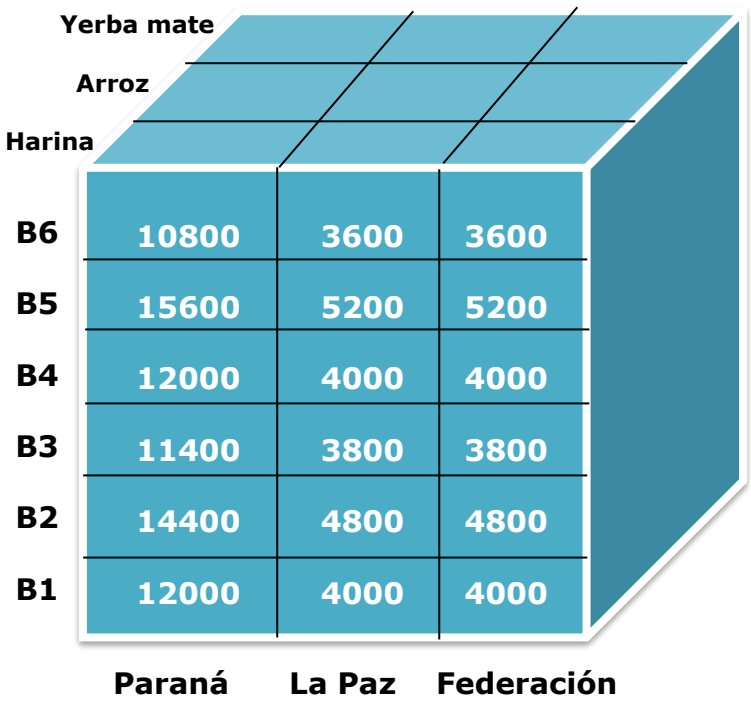
Producto	Cuatrimestre	Ciudad	Kilogramos vendidos
Harina	C1	Paraná	26400
Harina	C1	La Paz	8800
Harina	C1	Federación	8800
Harina	C2	Paraná	23400
Harina	C2	La Paz	7800
Harina	C2	Federación	7800
Harina	C3	Paraná	26400
Harina	C3	La Paz	8800
Harina	C3	Federación	8800
Arroz	C1	Paraná	16800
Arroz	C1	La Paz	5600
Arroz	C1	Federación	5600
Arroz	C2	Paraná	19800
Arroz	C2	La Paz	6600
Arroz	C2	Federación	6600

Arroz	C3	Paraná	12600
Arroz	C3	La Paz	4200
Arroz	C3	Federación	4200
Yerba mate	C1	Paraná	27600
Yerba mate	C1	La Paz	9200
Yerba mate	C1	Federación	9200
Yerba mate	C2	Paraná	28200
Yerba mate	C2	La Paz	9400
Yerba mate	C2	Federación	9400
Yerba mate	C3	Paraná	24600
Yerba mate	C3	La Paz	8200
Yerba mate	C3	Federación	8200

d) ¿Qué operador es necesario utilizar para obtener a partir de a) una consulta por Producto/Kilogramos vendidos? Resuélvalo.
Aplicamos el operador **Roll**:

Producto	Kilogramos vendidos
Harina	127000
Arroz	82000
Yerba mate	134000

e) Represente el punto b) luego de utilizar el operador tipo **Pívot**.



Resoluciones Serie Ejercicios Prácticos 8
Minería de Datos

1) Algoritmo ID3

Para resolver estos casos hay que tener en cuenta el algoritmo denominado ID3.
ID3 es un algoritmo matemático que permite generar un árbol de decisión, se construye de arriba hacia abajo y emplea el concepto de ganancia de la información para seleccionar el atributo más útil en cada caso.

Donde tenemos que:

- Cada nodo que no sea hoja representa un atributo
- Las aristas o ramas están etiquetados con cada uno de los posibles valores que puede tomar el atributo
- Los nodos hojas contiene los valores de la clasificación

De la tabla de datos proporcionada, para saber si se puede jugar al fútbol, la entropía que tiene el sistema respecto a la clasificación de si se juega o no, es aplicando la siguiente fórmula:

$$\textbf{Entropia(S)} = \textbf{E(S)} = -\sum_{i=1}^{C} p_i * \log_2 (p_i)$$

La **entropía** permite calcular el grado de incertidumbre de una muestra, si la muestra es totalmente homogénea la entropía es igual a 0 y si la muestra es igualmente distribuida tiene entropía igual a 1.

Donde S es el conjunto de muestras (el sistema analizado), C es el número de diferentes clasificaciones que usamos (Si y No), y cada p_i es la proporción/probabilidad de ejemplos/casos que hay de la clasificación i en la muestra.

Recordemos que podemos calcular $\log_2 (p_i) = \log_{10}(p_i) / \log_{10}(2)$ y que debido a que no está definido el $\log_x(0)$, tomaremos siempre que $0 \log_2(0) = 0$.

En el caso particular de una clasificación binaria (que podríamos denotar como casos positivos – Si / negativos - No), la fórmula anterior queda como:

$$\textbf{E(S)= (- P log}_2\textbf{(P)) + (- N log}_2\textbf{(N))}$$

donde P y N son, respectivamente, la proporción de casos positivos y negativos, positivo es si es posible jugar y negativo cuando no es posible jugar al fútbol.

Considerando el lote de datos en análisis:

Jugar fútbol	
Si	No
9	5
Probabilidad=9/14	Probabilidad=5/14

Se evalúan 14 días en total de la muestra considerada.

Entropía(Jugarfútbol) = Entropía(9,5) = $(-9/14 \log_2 (9/14)) + (-5/14 \log_2 (5/14))$ = **0,94**

Si obtenemos $\log_2$ de cada probabilidad:
$\log_2 (9/14) = \log_2 (0,64) = \log_{10}(0,64) / \log_{10}(2) = -0,64$
$\log_2 (5/14) = \log_2 (0,36) = \log_{10}(0,36) / \log_{10}(2) = -1,47$

E(Jugarfútbol) = (-9/14*(-0,64)) + (-5/14*(-1,47)) = (-0,64*(-0,64)) + (-0,36*(-1,47)) = 0,9388 = 0,94

Ganancia de la información:

- Se basa en el decremento de la entropía cuando el conjunto de datos se divide en los valores de un atributo
- ¿Qué atributo crea las ramas más homogéneas?
  - ✓ Se calcula la entropía total (Entropia(Jugarfútbol))
  - ✓ Se divide el conjunto de datos en función de los diferentes atributos
  - ✓ Se calcula la entropía de cada rama y se suman proporcionalmente las ramas para calcular la entropía del total
  - ✓ Se resta este resultado de la entropía total (E(Jugarfútbol))

- ✓ El resultado es la ganancia de información, descenso de entropía
- ✓ El atributo con mayor ganancia se selecciona como nodo de decisión
- ✓ Una rama con entropía 0 se convierte en hoja (nodo-respuesta, todos sus casos ya están clasificados), ya que representa una muestra completamente homogénea, en la que todos los casos tienen la misma clasificación
- ✓ Si no es así, la rama debe seguir subdividiéndose, para clasificar mejor sus nodos
- ✓ El algoritmo ID3 se ejecuta recursivamente en nodos que no son hojas, hasta que se llegue a nodos hoja

¿Qué atributo crea las ramas más homogéneas y, por tanto, proporciona una ganancia de información mayor?
Para ello, hay que ir determinando la entropía asociada a cada atributo X, sería:

$$\textbf{Entropía}(S,X) = E(S,X) = \sum_{c \in X}^{C} P(c)\, E(c)$$

c son los posibles valores que puede tomar el atributo X
E(c) es la entropía respecto a los casos positivos "Si" y casos negativos "No" que puede tomar **c**, $\mathbf{E(P_c,N_c)}$.

Y la ganancia de información que aportaría dividir respecto a los valores de ese atributo, resulta de:

$$\textbf{Ganancia}(S,X) = G(S,X) = E(S) - E(S,X)$$

Calculemos la ganancia que obtendríamos si hiciéramos una división usando el primer atributo ***Tiempo***:

		Jugar fútbol		
		Si	No	
Tiempo	soleado	2	3	5
	nublado	4	0	4
	lluvia	3	2	5
				14

Entropía(Jugarfútbol, Tiempo) =

E(Jugarfútbol, Tiempo) = P(soleado) * E(2,3) + P(nublado) * E(4,0) + P(lluvia) * E(3,2) = (5/14) * 0,971) + (4/14) * 0 + (5/14) * 0,971 = 0,3751 * 0,971 + 0,3751 * 0,971 = 0,693

Para este atributo, obtendremos las entropías de los casos positivos y negativos que pueden tomar cada uno de sus valores:

E(soleado) = E(2,3) = E(2/5,3/5) = E(0,4 , 0,60)= (-0,4 $\log_2$ (0,4)) + (-0,60 $\log_2$ (0,60)) = = (-0,4 * -1,3219) + (-0,6 * -0,7369) = 0,52876 + 0,44214 = 0,9709
por lo que
E(soleado) = E(2,3) = 0,971

Para calcular los logaritmos tenemos:
$\log_2$ (2/5) = $\log_2$ (0,4) = $\log_{10}$(0,40) / $\log_{10}$(2) = -0,39794 / 0,301029 = -1,3219
$\log_2$ (3/5) = $\log_2$ (0,6) = $\log_{10}$(0,60) / $\log_{10}$(2) = -0,22184 / 0,301029 = -0,7369

E(nublado) = E(4,0) = E(4/4,0/4) = E(1,0) = (-1 $\log_2$ (1)) + (-0 $\log_2$ (0)) = 0 + 0 = 0

$\log_2$ (1) = $\log_{10}$(1) / $\log_{10}$(2) = 0 / 0,301029 = 0
$\log_2$ (0) = $\log_{10}$(0) = 0 al no existir logaritmo en base$_{10}$ para el valor 0, consideramos igual a 0

E(lluvia) = E(3,4) = E(3/5,2/5) = E(0,6 , 0,40)= (-0,6 $\log_2$ (0,6)) + (-0,40 $\log_2$ (0,40)) = = (-0,6 * -0,7369) + (-0,4 * -1,3219) = 0,44214 + 0,52876 = 0,9709
por lo que
E(lluvia) = E(3,2) = 0,971

Ahora debemos de obtener la ganancia de información para el atributo tiempo, para ello consideramos la diferencia entre la entropía total y la entropía tiempo:

G(Jugarfútbol, Tiempo) = E(Jugarfútbol) – E(Jugarfútbol, Tiempo) = 0,94 – 0,693 = 0,247

Ganancia(Jugarfútbol, Tiempo)= 0,25

Ahora debemos de replicar este procedimiento para los demás atributos de los casos:

Ganancia del atributo ***Temperatura***:

		Jugar fútbol		
		Si	No	
Temperatura	alta	2	2	4
	templada	4	2	6
	frío	3	1	4
				14

E(Jugarfútbol, Temperatura) = P(alta) * E(2,2) + P(templada) * E(4,2) + P(frío) * E(3,1) =
(4/14) * E(2/4,2/4) + (6/14) * E(4/6,2/6) + (4/14) * E(3/4,1/4) =
0,28571 * 1 + 0,42857 * 0,918301 + 0,28571 * 0,81127 = 0,91104

G(Jugarfútbol, Temperatura) = 0,94 – 0,91104 = 0,0289 = 0,03

Ganancia del atributo ***Humedad***:

		Jugar fútbol		
		Si	No	
Humedad	alta	3	4	7
	normal	6	1	7
				14

E(Jugarfútbol, Humedad) = P(alta) * E(3,4) + P(normal) * E(6,1) =
(7/14) * E(3/7,4/7) + (7/14) * E(6/7,1/7) =
0,5 * 0,9851 + 0,5 * 0,5915 = 0,7883

G(Jugarfútbol, Humedad) = 0,94 – 0,7883 = 0,15

Ganancia del atributo ***Viento***:

		Jugar fútbol		
		Si	No	
Viento	débil	6	2	8
	fuerte	3	3	6
				14

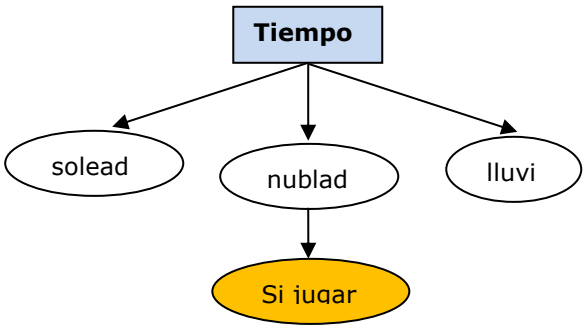
E(Jugarfútbol, Viento) = P(débil) * E(6,2) + P(fuerte) * E(3,3) =
(8/14) * E(6/8,2/8) + (6/14) * E(3/6,3/6) =
0,571 * 0,8112 + 0,428 * 1 = 0,8911

G(Jugarfútbol, Viento) = 0,94 – 0,8911 = 0,0489 = 0,05

Las ganancias de los atributos calculados son:

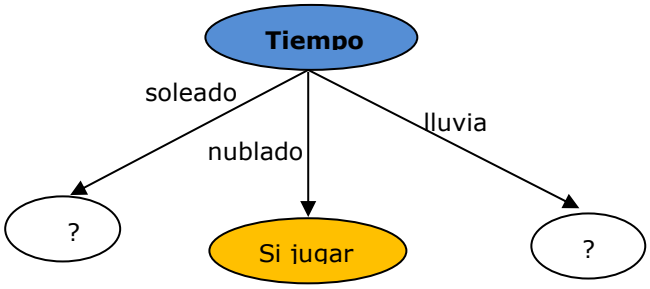
Atributos	Ganancia de información
Tiempo	0,25
Temperatura	0,03
Humedad	0,15
Viento	0,05

Vamos a elegir el atributo que proporciona mayor ganancia, en este caso corresponde a **Tiempo**, por lo que podemos construir el primer paso del árbol de decisión, identificando el primer nodo de decisión:



Un nodo que tenga entropía nula se convierte en un nodo respuesta, ya que representa una muestra homogénea en el que la clasificación final es la misma para todos los ejemplos/casos que contiene.

También podemos representar de la siguiente manera:



A continuación, nos situamos en cada uno de los subconjuntos de casos que define cada valor del atributo seleccionado y repetimos el proceso, construyendo poco a poco el árbol completo de decisión.

Ahora seleccionados de la tabla las filas que corresponde al valor "**soleado**" para el atributo Tiempo, repetimos los cálculos, pero con esta muestra menor.

Tiempo	Temperatura	Humedad	Viento	Jugar fútbol
soleado	alta	Alta	débil	no
soleado	alta	Alta	fuerte	no
soleado	templada	Alta	débil	no
soleado	frio	Normal	débil	si
soleado	templada	Normal	fuerte	si

Obtenemos la entropía de este lote de datos:

E(soleado) = **E(3,2)** = -(3/5) * log(3/5;2) + -(2/5) * log(2/5;2) = **0,97**

Ganancia del atributo ***Temperatura***:

		Soleado		
		Si	No	
Temperatura	alta	0	2	2
	templada	1	1	2
	frío	1	0	1
				5

E(soleado, Temperatura) = P(alta) * E(0,2) + P(templada) * E(1,1) + P(frío) * E(1,0) =
(2/5) * E(0/2,2/2) + (2/5) * E(1/2,1/2) + (1/5) * E(1/1,0/1) =
0,4 * 0 + 0,4 * 1 + 0,2 * 0 = 0,4
G(soleado, Temperatura) = 0,97 – 0,4 = 0,57

Ganancia del atributo ***Humedad***:

		Soleado		
		Si	No	
Humedad	alta	0	3	3
	normal	2	0	2
				5

E(soleado, Humedad) = P(alta) * E(0,3) + P(normal) * E(2,0) =
(3/5) * E(0/3,3/3) + (2/5) * E(2/2,0/2) = 0,6*0 + 0,4*0 = 0
G(soleado, Humedad) = 0,97 – 0 = 0,97

Ganancia del atributo ***Viento***:

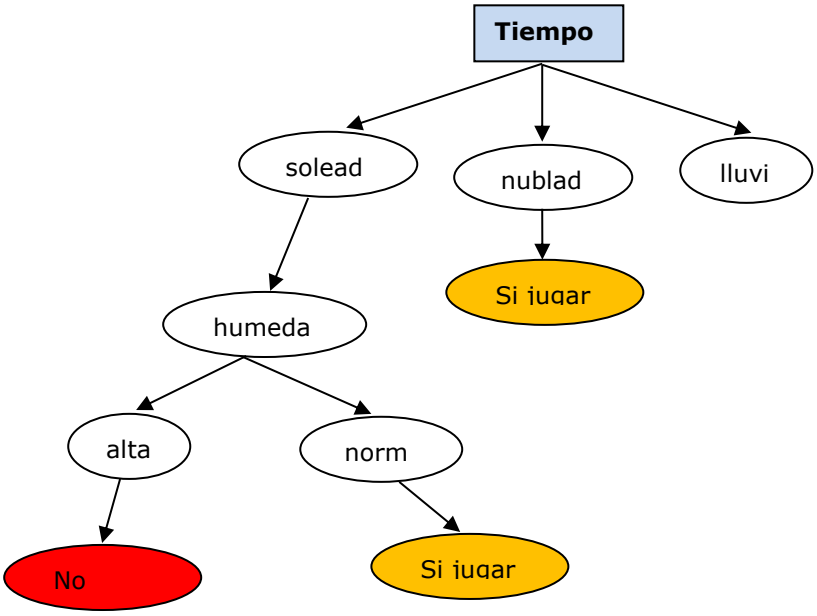
		Soleado		
		Si	No	
Viento	débil	1	2	3
	fuerte	1	1	2
				5

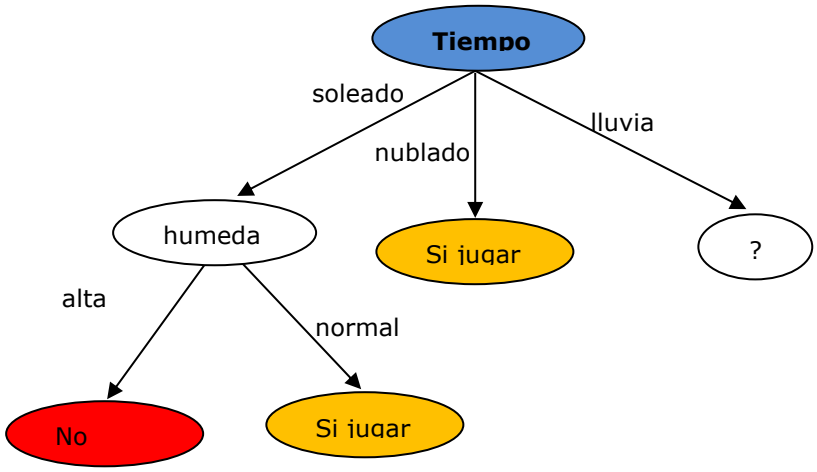
E(soleado, Viento) = P(débil) * E(1,2) + P(fuerte) * E(1,1) =
(3/5) * E(1/3,2/3) + (2/5) * E(1/2,1/2) =
0,6 * 0,9183 + 0,4 * 1 = 0,95098
G(soleado, Viento) = 0,97 – 0,95098 = 0,019 = 0,02

Las ganancias de los atributos calculados son:

Atributos	Ganancia de información
Temperatura	0,57
Humedad	0,97
Viento	0,02

Representamos en nuestro árbol de decisión:





Por último, nos queda evaluar de la tabla las filas que corresponde al valor "**lluvia**" para el atributo Tiempo.

Tiempo	Temperatura	Humedad	Viento	Jugar fútbol
lluvia	templada	Alta	débil	si
lluvia	frio	Normal	débil	si
lluvia	frio	Normal	fuerte	no
lluvia	templada	Normal	débil	si
lluvia	templada	Alta	fuerte	no

Obtenemos la entropía de este lote de datos:

E(lluvia) = **E(3,2)** = -(3/5) * log(3/5;2) + -(2/5) * log(2/5;2) = **0,97**

Ganancia del atributo ***Temperatura***:

		lluvia		
		Si	No	
Humedad	templada	2	1	3
	frío	1	1	2
				5

E(lluvia, Temperatura) = P(templada) * E(2,1) + P(frío) * E(1,1) =

(3/5) * E(2/3,1/3) + (2/5) * E(1/2,1/2) =
0,6 * 0,9183 + 0,4 * 1 = 0,9598
G(lluvia, Temperatura) = 0,97 – 0,9598 = 0,019 = 0,02

Ganancia del atributo ***Humedad***:

		lluvia		
		Si	No	
Humedad	alta	1	1	2
	normal	2	1	3
				5

E(lluvia, Humedad) = P(alta) * E(1,1) + P(normal) * E(2,1) =
(2/5) * E(1/2,1/2) + (3/5) * E(2/3,1/3) =
0,4*1 + 0,6*0,9183 = 0,9598
G(lluvia, Humedad) = 0,97 – 0,9598 = 0,019 = 0,02

Ganancia del atributo ***Viento***:

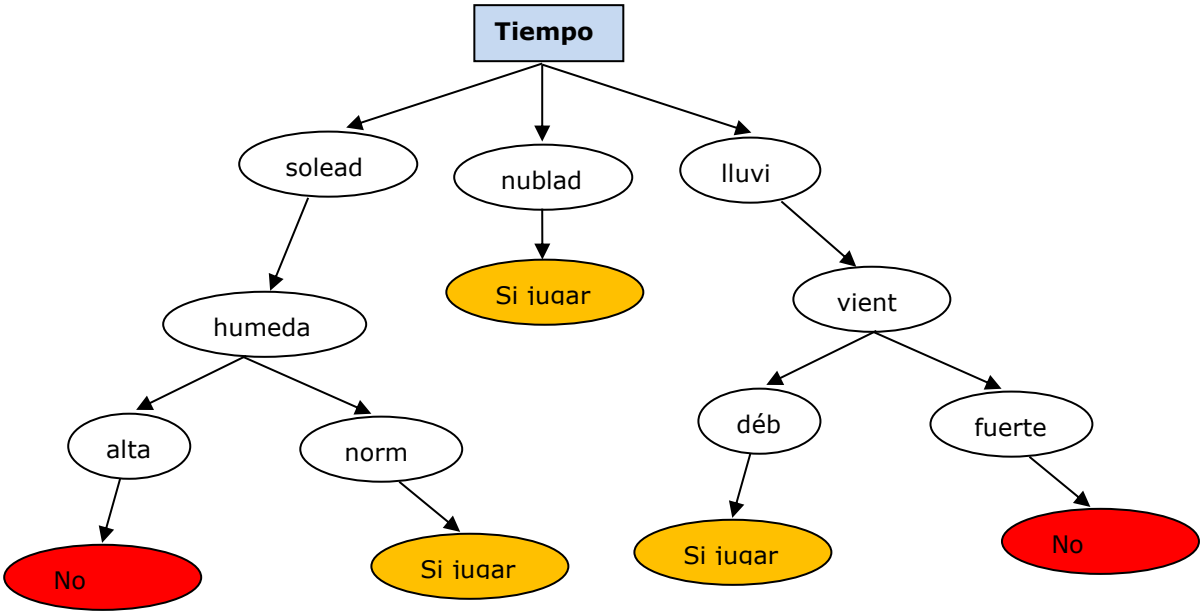
		lluvia		
		Si	No	
Viento	débil	3	0	3
	fuerte	0	2	2
				5

E(lluvia, Viento) = P(débil) * E(3,0) + P(fuerte) * E(0,2) =
(3/5) * E(3/3,0/3) + (2/5) * E(0/2,2/2) = 0,6 * 0 + 0,4 * 0 = 0
G(lluvia, Viento) = 0,97 – 0 = 0,97

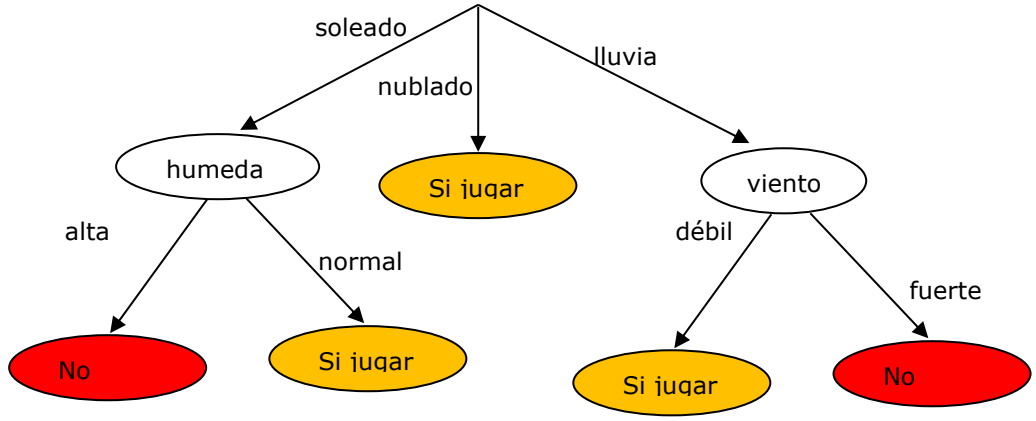
Las ganancias de los atributos calculados son:

Atributos	Ganancia de información
Temperatura	0,02
Humedad	0,02
Viento	0,97

El árbol de decisión final será:



Tiempo



La verificación de que, si es posible o no jugar al fútbol, la realizamos desde arriba y hacia abajo del árbol resultante.

2) Aplicar el algoritmo "**apriori**" para encontrar las reglas que predicen la ocurrencia de un ítem. Considerado al soporte mínimo = 2 (support – supp - minsup) y para la confianza = 0.75 (confidence – conf - minconf).

En la tabla se indican una colección de ítems que un cliente de un supermercado compra en una misma transacción, el problema viene dado por identificar el conjunto de ítems que son adquiridos en conjunto, siendo esta una de sus aplicaciones, la del análisis de los carros de compras.

Transacción	Ítems
1	Aceite, leche, azúcar
2	Aceite, yerba, cerveza, azúcar
3	Leche, yerba, cerveza, sal
4	Aceite, leche, yerba, cerveza, sal
5	Aceite, leche, yerba, sal

Transacción	Aceite	Leche	Yerba	Cerveza	Azúcar	Sal
1	1	1	0	0	1	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	1
5	1	1	1	0	0	1

Se deben de encontrar las reglas de asociación con valores de soporte y confianza, que cumplan con los umbrales mínimos establecidos por el usuario en cuanto al soporte mínimo y confianza mínima (minsup y minconf), en este caso minsup = 2 y minconf = 0.75.

Para lo cual tenemos:
$I = \{i_1, i_2, i_3, \ldots\ldots\ldots\ldots, i_n\}$ es el conjunto de ítems
D es conjunto de transacciones T_j
Cada transacción T_j es un subconjunto de ítems de entre los posibles ítems definidos en I.

I = {Aceite, Leche, Yerba, Cerveza, Azúcar, Sal}
D = {1,2,3,4,5} = {{Aceite, Leche, Azúcar}, {Aceite, Azúcar, Yerba, Cerveza},
{Leche, Yerba, Cerveza, Sal}, {Aceite, Leche, Yerba, Cerveza, Sal}
{Aceite, Leche, Yerba, Sal}}

A toda colección de cero o más ítems se denomina *itemset*. Si un *itemset* contiene *k* ítems, se le llama *k-itemset.* En el ejemplo de la cesta de la compra anterior la transacción {Aceite, Azúcar, Yerba, Cerveza} es un ejemplo de un 4-itemset. Si tenemos un conjunto de datos con cero elementos también es válido y sería el *itemset* vacío.

El tamaño de una transacción lo definimos como el número de *ítems* presentes en la transacción. Una transacción t_j decimos que contiene un *itemset X* si *X* es un subconjunto de t_j. Por ejemplo, la segunda transacción de la Tabla contiene el *itemset* {Aceite, Yerba} pero no contiene el *itemset* {Aceite, Sal}.
Una propiedad que se asocia a un *itemset* es su frecuencia de ocurrencia que se define como el número de transacciones que contienen dicho *itemset*. Podemos ver que el *itemset* {Aceite, Yerba} aparece en 3 transacciones por lo que su frecuencia de ocurrencia es 3.

Una regla de asociación es una expresión que tiene la forma X → Y donde X e Y son itemsets disjuntos. Para medir la fuerza de una regla de asociación se utilizan los conceptos de soporte y confianza.

El **soporte** de una regla es la proporción de transacciones que contienen el itemset formado por los ítems que aparecen en el antecedente y consecuente de la regla.

La **confianza** de una regla es la proporción de transacciones que conteniendo los ítems del antecedente de la regla también contienen los ítems del consecuente de la regla, es decir, la proporción de transacciones para las que se cumple la regla.

Las definiciones formales que son las siguientes:

$$\textbf{soporte } \{X \rightarrow Y\} = \frac{X \cup Y}{N} \qquad \textbf{confianza } \{X \rightarrow Y\} = \frac{X \cup Y}{X}$$

Veamos un ejemplo de estos conceptos basado en los elementos del carrito de compra de nuestro caso.

Consideremos la regla {Aceite, Leche} → {Yerba}, como sabemos que la frecuencia de ocurrencia de {Aceite, Leche, Yerba} es 2 y el número total de transacciones es 5 el soporte de la regla es 2/5, es decir, igual a 0,4 que representa un 40%.

La confianza de la regla se obtiene dividendo la frecuencia de ocurrencia de {Aceite, Leche, Yerba} entre la frecuencia de ocurrencia de {Aceite, Leche}. Como hay 3 transacciones que contienen {Aceite, Leche} la confianza para esta regla es 2/3. Es decir, igual a 0,66, la regla se cumple en el 66% de los casos.

El proceso de generación de reglas de asociación consiste en encontrar, para un conjunto de transacciones, todas las reglas que tengan un soporte y confianza iguales o superiores a unos valores mínimos a los que se denomina minsup y minconf, fijados por el usuario.

Este proceso puede descomponerse en 2 fases para facilitar su implementación.

A.-**Generación de los itemsets frecuentes**

Cuyo objetivo es descubrir todos los itemsets que satisfagan el umbral minsup (soporte mínimo). Estos itemsets descubiertos serán los denominados itemsets frecuentes.

B.-**Generación de Reglas**

Cuyo objetivo es extraer, a partir de los itemsets frecuentes encontrados en el paso previo, todas las reglas que cumplan con el grado de confianza minconf (confianza mínima).

Se inicia el algoritmo identificando todos los *items* individuales (*itemsets* de un único *ítem*) y calculando su soporte.
Para la tabla presentada, el soporte de cada uno de los ítems es de:

Ítem	Soporte	Soporte mínimo
{Aceite}	4	4/5=0,80=80%
{Leche}	4	4/5=0,80=80%
{Yerba}	4	4/5=0,80=80%
{Cerveza}	3	3/5=0,60=60%
{Azúcar}	2	2/5=0,40=40%
{Sal}	3	3/5=0,60=60%

Indica la cantidad de veces que aparece cada ítem en las transacciones evaluadas.
Todos los *itemsets* de tamaño *k = 1* tienen un soporte igual o superior al mínimo establecido, por lo que todos superan la fase de filtrado (poda).

Otra forma de representar es:
S_1 = {{Aceite}:4, {Leche}:4, {Yerba}:4, {Cerveza}:3, {Azúcar}:2, {Sal}:3}

A continuación, veremos la generación de itemsets frecuentes usando el algoritmo a priori.

Fase A – Conjuntos Frecuentes

Combinaciones de 2 ítems, itemset (*k=2*):
S_2 = {
{Aceite, Leche}, {Aceite, Yerba}, {Aceite, Cerveza}, {Aceite, Azúcar}, {Aceite, Sal},
{Leche, Yerba}, {Leche, Cerveza}, {Leche, Azúcar}, {Leche, Sal},
{Yerba, Cerveza}, {Yerba, Azúcar}, {Yerba, Sal}
{Cerveza, Azúcar}, {Cerveza, Sal}
}

Vemos el soporte de cada regla:
S_2 = {
{Aceite, Leche}:3, {Aceite, Yerba}:3, {Aceite, Cerveza}:2, {Aceite, Azúcar}:2, {Aceite, Sal}:2,
{Leche, Yerba}:3, {Leche, Cerveza}:2, {Leche, Azúcar}:1, {Leche, Sal}:3,
{Yerba, Cerveza}:3, {Yerba, Azúcar}:1, {Yerba, Sal}:3,
{Cerveza, Azúcar}:1, {Cerveza, Sal}:2
}

Quitamos aquellas reglas que no llegan a cumplir con el minsup solicitado, igual a 2, quedando aquellas que son iguales o mayores a ese valor:
S_2 = {
{Aceite, Leche}:3, {Aceite, Yerba}:3, {Aceite, Cerveza}:2, {Aceite, Azúcar}:2, {Aceite, Sal}:2,
{Leche, Yerba}:3, {Leche, Cerveza}:2, {Leche, Sal}:3,
{Yerba, Cerveza}:3, {Yerba, Sal}:3,
{Cerveza, Sal}:2
}

Combinaciones de 3 ítems (recorriendo cada transacción) y contabilizando en una regla las ocurrencias que se repiten, itemset (*k=3*)::
S_3 = {
{Aceite, Leche, Azúcar}:1,
{Aceite, Yerba, Cerveza}:2, {Aceite, Yerba, Azúcar}:1, {Aceite, Cerveza, Azúcar}:1,
{Leche, Yerba, Cerveza}:2, {Leche, Yerba, Sal}:3, {Leche, Cerveza, Sal}:2,
{Aceite, Leche, Yerba}:2, {Aceite, Leche, Cerveza}:1, {Aceite, Leche, Sal}:2,
{Aceite, Yerba, Sal}:2,
{Aceite, Cerveza, Sal}:1, {Yerba, Cerveza, Sal}:2,
}

Y aquellas reglas que cumplen con el minsup de 2, son:
S_3 = {
{Aceite, Yerba, Cerveza}:2,
{Leche, Yerba, Cerveza}:2, {Leche, Yerba, Sal}:3, {Leche, Cerveza, Sal}:2,
{Aceite, Leche, Yerba}:2, {Aceite, Leche, Sal}:2,
{Aceite, Yerba, Sal}:2,
{Yerba, Cerveza, Sal}:2,
}

Combinaciones de 4 ítems (recorriendo cada transacción) y contabilizando en una regla las ocurrencias que se repiten, itemset (*k=4*)::
S_4 = {
{Aceite, Azúcar, Yerba, Cerveza}:1,
{Leche, Yerba, Cerveza, Sal}:2,
{Aceite, Leche, Yerba, Cerveza}:1, {Aceite, Leche, Yerba, Sal}:2,
{Aceite, Leche, Cerveza, Sal}:1, {Aceite, Yerba, Cerveza, Sal}:1,
}

Quedando las de soporte:
S_4 = {
{Leche, Yerba, Cerveza, Sal}:2,
{Aceite, Leche, Yerba, Sal}:2,
}

Combinaciones de 5 ítems solo se da una regla, itemset (*k=5*), que no configura con el soporte mínimo:
S_5 = {{Aceite, Leche, Yerba, Cerveza, Sal}:1}

De las reglas que cumplen con el soporte mínimo establecido, tenemos a los conjuntos conformados por 2, 3 y 4 ítems:

$S_{final} = S_2 \cup S_3 \cup S_4$ =
{

{Aceite, Leche}:3, {Aceite, Yerba}:3, {Aceite, Cerveza}:2, {Aceite, Azúcar}:2, {Aceite, Sal}:2,
{Leche, Yerba}:3, {Leche, Cerveza}:2, {Leche, Sal}:3,
{Yerba, Cerveza}:3, {Yerba, Sal}:3,
{Cerveza, Sal}:2
{Aceite, Yerba, Cerveza}:2,
{Leche, Yerba, Cerveza}:2, {Leche, Yerba, Sal}:3, {Leche, Cerveza, Sal}:2,
{Aceite, Leche, Yerba}:2, {Aceite, Leche, Sal}:2,
{Aceite, Yerba, Sal}:2,
{Yerba, Cerveza, Sal}:2,
{Leche, Yerba, Cerveza, Sal}:2,
{Aceite, Leche, Yerba, Sal}:2,
}

Fase B – Generación de Reglas, evaluación de la confianza

A partir de S_{final} de los soportes mínimos, ahora calcularemos la confianza para luego obtener aquellas reglas de asociación candidatas que cumplan al menos con el soporte y confianza en sus valores mínimos. Recordemos que: **soporte {X → Y} = $\frac{X \cup Y}{N}$** **confianza {X → Y} = $\frac{X \cup Y}{X}$**

Ítemsets frecuentes	Soporte mínimo	Confianza
{Aceite, Leche}	3	conf({Aceite, Leche}) = supp({Aceite, Leche}) / supp({Aceite}) =(3/5) / (4/5) = 3/4 = 0.75
{Aceite, Yerba}	3	conf({Aceite, Yerba}) = supp({Aceite, Yerba}) / supp({Aceite}) = (3/5) / (4/5) = 3/4 = 0.75
{Aceite, Cerveza}	2	conf({Aceite, Cerveza}) = supp({Aceite, Cerveza}) / supp({Aceite}) = (2/5) / (3/5) = 2/3 = 0.66
{Aceite, Azúcar}	2	conf({Aceite, Azúcar}) = supp({Aceite, Azúcar}) / supp({Aceite}) = (2/5) / (4/5) = 2/4 = 0.50
{Aceite, Sal}	2	conf({Aceite, Sal}) = supp({Aceite, Sal}) / supp({Aceite}) = (2/5) / (4/5) = 2/3 = 0.50
{Leche, Yerba}	3	conf({Leche, Yerba}) = supp({Leche, Yerba}) / supp({Leche}) = (3/5) / (4/5) = 3/4 = 0.75
{Leche, Cerveza}	2	conf({Leche, Cerveza}) = supp({Leche, Cerveza}) / supp({Leche}) = (2/5) / (4/5) = 2/3 = 0.50
{Leche, Sal}	3	conf({Leche, Sal}) = supp({Leche, Sal) / supp({Leche}) = (3/5) / (4/5) = 3/4 = 0.75
{Yerba, Cerveza}	3	conf({Yerba, Cerveza}) = supp({Yerba, Cerveza}) / supp({Yerba}) = (3/5) / (4/5) = 2/3 = 0.75
{Yerba, Sal}	3	conf({Yerba, Sal}) = supp({Yerba, Sal}) / supp({Yerba}) = (2/5) / (4/5) = 2/4 = 0.50
{Cerveza, Sal}	2	conf({Cerveza, Sal}) = supp({Cerveza, Sal}) / supp({Cerveza}) = (2/5) / (3/5) = 2/3 = 0.66
{Aceite, Yerba, Cerveza}	2	conf({Aceite, Yerba, Cerveza}) = supp({Aceite, Yerba, Cerveza}) / supp({Aceite, Yerba}) = (2/5) / (3/5) = 0.66
{Leche, Yerba, Cerveza}	2	conf({Leche, Yerba, Cerveza}) = supp({Leche, Yerba, Cerveza}) / supp({Leche, Yerba}) = (2/5) / (3/5) = 2/3 = 0.66
{Leche, Yerba, Sal}	3	conf({Leche, Yerba, Sal}) = supp({Leche, Yerba, Sal}) / supp({Leche, Yerba}) = (3/5) / (3/5) = 1
{Leche, Cerveza, Sal}	2	conf({Leche, Cerveza, Sal}) = supp({Leche, Cerveza, Sal}) / supp({Leche, Cerveza}) = (2/5) / (2/5) = 1
{Aceite, Leche, Yerba}	2	conf({Aceite, Leche, Yerba}) = supp({Aceite, Leche, Yerba}) / supp({Aceite, Leche}) = (2/5) / (3/5) = 2/3 = 0.66
{Aceite, Leche, Sal}	2	conf({Aceite, Leche, Sal}) = supp({Aceite, Leche, Sal}) / supp({Aceite, Leche}) = (2/5) / (3/5) = 0.66
{Aceite, Yerba, Sal}	2	conf({Aceite, Yerba, Sal}) = supp({Aceite, Yerba, Sal}) / supp({Aceite, Yerba}) = (2/5) / (2/5) = 1
{Yerba, Cerveza, Sal}	2	conf({Yerba, Cerveza, Sal}) = supp({Yerba, Cerveza, Sal}) / supp({Yerba, Cerveza}) = (2/5) / (3/5) = 2/3 = 0.66
{Leche, Yerba, Cerveza, Sal}	2	conf({Leche, Yerba, Cerveza, Sal}) = supp({Leche, Yerba, Cerveza, Sal}) / supp({Leche, Yerba, Cerveza}) = (2/5) / (2/5) = 1
{Aceite, Leche, Yerba, Sal}	2	conf({Aceite, Leche, Yerba, Sal}) = supp({Aceite, Leche, Yerba, Sal}) / supp({Aceite, Leche, Yerba}) = (2/5) / (2/5) = 1

Finalmente, las reglas de asociación que cumplen con los pisos mínimos de soporte = 2 y confianza = 0.75 son:

Reglas	Soporte	Confianza
{Aceite, Leche}	3	0.75
{Aceite, Yerba}	3	0.75
{Leche, Yerba}	3	0.75
{Leche, Sal}	3	0.75
{Yerba, Cerveza}	3	0.75
{Leche, Yerba, Sal}	3	1
{Leche, Cerveza, Sal}	2	1
{Aceite, Yerba, Sal}	2	1
{Leche, Yerba, Cerveza, Sal}	2	1
{Aceite, Leche, Yerba, Sal}	2	1

3) Clasificar cada persona como alta o baja, según el algoritmo On-Line **k-MEANS**.

Los datos considerados son:

Número (Persona)	Altura

1	190
2	160
3	158
4	177
5	189
6	171
7	160
8	154
9	200
10	172

Resolución:

a) Paso 1: obtener la altura promedio de las personas del lote de datos proporcionado:
∑ altura de las personas / cantidad de personas:
(190 + 160 + 158 + 177 + 189 + 171 + 160 + 154 + 200 + 172) /10 = **1731/10** = **173,1**

b) Paso 2: clasificar a las personas según tengan una altura superior a la altura promedio como personas altas, y aquellas que estén por debajo de está, serán consideradas personas bajas:

Número (Persona)	Altura	Clasificación
1	190	Alta
2	160	Baja
3	158	Baja
4	177	Alta
5	189	Alta
6	171	Baja
7	160	Baja
8	154	Baja
9	200	Alta
10	172	Baja

4) Sea la siguiente tabla de Navigation Trails:

ID	Trail
1	$A_1 \longrightarrow A_2 \longrightarrow A_3 \longrightarrow A_4$
2	$A_1 \longrightarrow A_5 \longrightarrow A_3 \longrightarrow A_4 \longrightarrow A_1$
3	$A_5 \longrightarrow A_2 \longrightarrow A_4 \longrightarrow A_6$
4	$A_5 \longrightarrow A_2 \longrightarrow A_3$
5	$A_5 \longrightarrow A_2 \longrightarrow A_3 \longrightarrow A_6$
6	$A_4 \longrightarrow A_1 \longrightarrow A_5 \longrightarrow A_3$

Buscando Patrones de Navegación mediante HPGs (Borges and Levene 2000):

Las sesiones o log files de navegación toman la forma de secuencias de enlaces recorridos por un usuario, conocidos como *navigation trails* o sesiones.
Consideraremos al término enlace como sinónimo de página, documento, URL o visita.

Los 'navigation trails' se utilizan para construir una Hypertext Probabilistic Grammar (HPG), podríamos traducirlo como Gramática Probabilística de Hipertexto.

Una HPG es una tupla <V, ∑, S, P>. No es más que un tipo especial de gramáticas probabilísticas regulares, con la característica especial que tienen el mismo número de terminales S que no terminales V (con lo que se hace una correspondencia 1 a 1 entre ellos).

Se construye el grafo de transiciones de la gramática de la siguiente manera:

- Se añade un único nodo inicial S y un nodo final F, que no corresponden con ningún URL.
- Se añaden tantos nodos como URLs distintos haya en los distintos trails.

¿Qué valores probabilísticos ponemos en las flechas? (para saber las reglas de producción probabilística P)

Existen dos parámetros para construir el HPG:

- α: importancia de inicio.

• Si α = 0 sólo habrá flechas de S a los nodos que han sido alguna vez inicio de sesión, y el valor de la flecha dependerá de cuántas veces lo han sido.
• Si α = 1 el peso de las flechas dependerá de la probabilidad de visitas a cada nodo, independientemente de que fueran iniciales.
• Si α > 0 habrá flechas con peso > 0 de S a todos los nodos.
• N (donde N ≥ 1): valor de N-grama. Determina la memoria cuando se navega la red, es decir el número de URLs anteriores que pueden influir en la elección del próximo URL. Si N=1 el resultado será una cadena de Markov.

De aquí extraemos los no terminales y los terminales correspondientes:

$V = \{S, A_1, A_2, A_3, A_4, A_5, A_6, F\}$

$S = \{a_1, a_2, a_3, a_4, a_5, a_6\}$

Tenemos 6 trails y 24 visitas, donde A_1, por ejemplo, fue visitada 4 veces, 2 de las cuales como página de inicio.
Por tanto, tomando α =0,5 y N=1, podemos calcular la probabilidad de la producción $p(S \longrightarrow a_1A_1)$ que corresponde con la flecha de S a A_1 en el grafo de transiciones de la siguiente manera:

$p(S \longrightarrow a_1 A_1) = (0,5 * 4)/24 + (0,5 * 2)/6 = 0,25$

En primer lugar, tenemos en cuenta las ocurrencias en total del Nodo A_1 respecto a los 24 enlaces posibles, más las veces que el Nodo A_1 se encuentra como página de inicio respecto a los 6 posibles caminos dados. Las flechas interiores se calculan de manera similar.

Por ejemplo, si A_4 se ha visitado 4 veces, 1 justo antes del final, otra antes de A_6 y dos antes de A_1 tenemos:

$\underline{p}(A_4 \longrightarrow a_1A_1) = 2/4$ $\underline{p}(A_4 \longrightarrow a_6A_6) = 1/4$ $\underline{p}(A_4 \longrightarrow F) = 1/4$

En las siguientes tablas podemos expresar el conjunto de producciones probabilísticas *p* derivadas de la tabla dada (para α=0,5 y N=1), representamos por separado cada una y luego al final unificamos en una sola tabla:

Inicio producción	(visitas de cada Nodo) + (veces Nodo al inicio)	
$S \longrightarrow a_1 A_1$	$p(S \longrightarrow a_1 A_1) = (0,5 * 4)/24 + (0,5 *2)/6$	0,25
$S \longrightarrow a_2 A_2$	$p(S \longrightarrow a_2 A_2) = (0,5 * 4)/24$	0,08
$S \longrightarrow a_3 A_3$	$p(S \longrightarrow a_3 A_3) = (0,5 * 5)/24$	0,11
$S \longrightarrow a_4 A_4$	$p(S \longrightarrow a_4 A_4) = (0,5 * 4)/24 + (0,5 *1)/6$	0,17
$S \longrightarrow a_5 A_5$	$p(S \longrightarrow a_5 A_5) = (0,5 * 5)/24 + (0,5 *3)/6$	0,35
$S \longrightarrow a_6 A_6$	$p(S \longrightarrow a_6 A_6) = (0,5 * 2)/24$	0,35

Transición producción (visitas de cada Nodo origen al siguiente Nodo destino)		
$A_1 \longrightarrow a_2 A_2$	$p(A_1 \longrightarrow a_2 A_2) = ¼$	0,25
$A_1 \longrightarrow a_5 A_5$	$p(A_1 \longrightarrow a_5 A_5) = 2/4$	0,50
$A_2 \longrightarrow a_3 A_3$	$p(A_2 \longrightarrow a_3 A_3) = ¾$	0,75
$A_2 \longrightarrow a_4 A_4$	$p(A_2 \longrightarrow a_4 A_4) = ¼$	0,25
$A_3 \longrightarrow a_4 A_4$	$p(A_3 \longrightarrow a_4 A_4) = 2/5$	0,40
$A_3 \longrightarrow a_6 A_6$	$p(A_3 \longrightarrow a_6 A_6) = 1/5$	0,20
$A_4 \longrightarrow a_1 A_1$	$p(A_4 \longrightarrow a_1 A_1) = 2/4$	0,50
$A_4 \longrightarrow a_6 A_6$	$p(A_4 \longrightarrow a_6 A_6) = ¼$	0,25
$A_5 \longrightarrow a_2 A_2$	$p(A_5 \longrightarrow a_2 A_2) = 3/5$	0,60
$A_5 \longrightarrow a_3 A_3$	$p(A_5 \longrightarrow a_3 A_3) = 2/5$	0,40

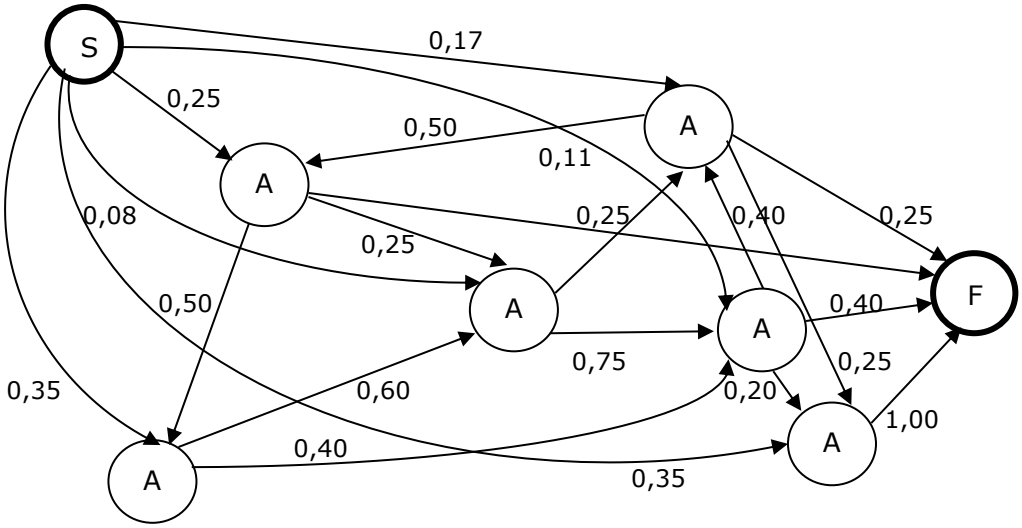
Final producción	(veces Nodo al final)	
$A_1 \longrightarrow F$	$p(A_1 \longrightarrow F) = ¼$	0,25
$A_3 \longrightarrow F$	$p(A_3 \longrightarrow F) = 2/5$	0,40
$A_4 \longrightarrow F$	$p(A_4 \longrightarrow F) = 1/4$	0,25
$A_6 \longrightarrow F$	$p(A_6 \longrightarrow F) = 2/2$	1,00

Representamos en una única tabla las probabilidades de producción de cada una de las etapas (inicio, transición y final):

Inicio producción	**Transición producción**	**Final producción**

$S \longrightarrow a_1\ A_1$	0,25	$A_1 \longrightarrow a_2\ A_2$	0,25	$A_1 \longrightarrow F$	0,25
$S \longrightarrow a_2\ A_2$	0,08	$A_1 \longrightarrow a_5\ A_5$	0,50	$A_3 \longrightarrow F$	0,40
$S \longrightarrow a_3\ A_3$	0,11	$A_2 \longrightarrow a_3\ A_3$	0,75	$A_4 \longrightarrow F$	0,25
$S \longrightarrow a_4\ A_4$	0,17	$A_2 \longrightarrow a_4\ A_4$	0,25	$A_6 \longrightarrow F$	1,00
$S \longrightarrow a_5\ A_5$	0,35	$A_3 \longrightarrow a_4\ A_4$	0,40		
$S \longrightarrow a_6\ A_6$	0,35	$A_3 \longrightarrow a_6\ A_6$	0,20		
		$A_4 \longrightarrow a_1\ A_1$	0,50		
		$A_4 \longrightarrow a_6\ A_6$	0,25		
		$A_5 \longrightarrow a_2\ A_2$	0,60		
		$A_5 \longrightarrow a_3\ A_3$	0,40		

De acuerdo a la tabla, podemos representar el grafo:



Bueno, ¿y ahora esto para qué sirve?

En primer lugar, permite estimar la probabilidad de cualquier 'navigation trail' todavía no producido.

Esto es útil para:

- Calcular la probabilidad de llegar a una cierta página si el usuario está en una página dada.
- La prueba de aplicaciones con los trails más comunes.
- El diseño ajustado a estos trails más comunes.
- La detección de usuarios anómalos (aquéllos que realizan trails con muy baja probabilidad).

En segundo lugar, y más importante, nos interesa ver aquellos "patrones de navegación".