```
Ingeniería en Computación
Asignatura: Minería de Datos
Examen 1 – Temas incluidos: 1 Descubrimiento al Conocimiento de los Datos y
2 Arquitectura de Minería de Datos
Descripción del Caso de Negocio
Barrera Peña Victor Miguel
```

# Caso de estudio

Una compañía de Seguros está preocupada por la incidencia de altos siniestros en el último año principalmente en las zonas urbanas, los incidentes que se han incrementado principalmente son la coalición entre autos. La venta está enfocada a pólizas de autos, el monto de la póliza depende del tipo de auto, marca, modelo, año de compra, si es flotilla, familiar o individual, su emisión es anual. Las pólizas emitidas y siniestradas se encuentran centralizadas. La venta de pólizas la realizan sus agentes, compañías asociadas y agencias regionales. Para identificar las causas de porque el incremento de siniestros, deciden contratar a una compañía para que los asesore y proporcione algún servicio que ayude a identificar la problemática y poder minimizar los costos por los siniestros constantes.
Basados en el caso de negocio responder a las siguientes preguntas:

# 1) (10 puntos) Que información adicional se tendría que preguntar para conocer con mayor detalle el caso de negocio y poder utilizarla en el proceso de Minería de Datos, para dar respuesta a la necesidad del cliente.

- 1. ¿Qué información ya se posee cuando se quiere realizar la plataforma?
  2. ¿Qué es lo que se quiere realizar con dicha información?
  3. ¿Qué registros se tienen?
  4. ¿Cuál es el presupuesto que se tiene?
  5. ¿Cómo se quiere que se gestione la información?
  6. ¿Hay limitaciones en el uso de la información?
  7. ¿Por qué solo en zonas urbanas? ¿Qué determinó esto?
  8. ¿Cómo se valoran las pólizas?
  9. ¿A partir de cuándo se dio el incremento de incidentes y qué los lleva a creer esto?
  10. ¿Al descubrir a que se deben los incidentes que quiere hacer con dicha información, segmentar las causas asociadas para poder reparar su sistema de valuación?

# 2) (30 puntos) Elaborar un diagrama de Arquitectura de Minería de datos para la propuesta de minería de datos y además describir los siguientes puntos:

## a. Repositorio de datos que participa y sus características

La información se almacena desde las sucursales que son las que levantan los incidentes de cuando se realizan los accidentes. Además de las sucursales para contratar los planes en caso de siniestro, ello está conectado a una API que logra poder almacenarlo en una arquitectura de AWS. Ya que es la primera implementación, en cuanto esta crezca es necesario cambiar el modelado de los datos y cómo están guardados.


## b. Identificar los datos participantes

- Información de las pólizas emitidas: tipo de auto, marca, modelo, año de compra, si es flotilla, familiar o individual, agente vendedor, compañía asociada o agencia regional.
- Información de los siniestros: fecha, hora, ubicación, tipo de incidente, autos involucrados, daños materiales y personales.
- Información de los clientes: edad, género, historial de conducción, antigüedad como asegurado, ubicación geográfica.
- Información de las zonas urbanas: densidad de tráfico, niveles de contaminación, infraestructura vial, señalización, clima.

## c. Limpieza y Tratamiento de datos recomendados

- Identificar y corregir errores de escritura o codificación.
- Normalizar datos, como convertir fechas a un formato estándar o agrupar valores en categorías.
- Completar datos faltantes mediante técnicas de imputación o inferencia a partir de otros datos disponibles.
- Transformar datos para facilitar el análisis, como agregar variables calculadas o convertir variables categóricas en numéricas.

## d. Proponer que algoritmos participaran para resolver la problemática y ¿porque se proponen?

- Segmentación: porque lo que queremos encontrar es las principales causas y con ello tomar accion

## e. Ambiente de visualización para los usuarios Ejecutivos, Operativos, Supervisores, etc.

- Tableu para mostrar graficas, yo la utilizo y es muy potente.

# 3) (10 Puntos) Describir las fases del plan de trabajo que se presentaría al cliente para realizar el Proceso de Minería de Datos

En este caso, primero se tendría que saber cómo está el sistema en donde se registran los incidentes para saber cómo están almacenados y, de esta manera, cómo reestructurarlos para que sean fácilmente tratados.

Ya, si tienen que estar con una alta disponibilidad y con una rapidez alta, o enfocarse para que el sistema sea un poco más lento pero seguro y fácil de mantener a la larga.

Fase 1:

- Reestructuración de la entrada de información (introducción mediante la API).
- Manejo de la información que ya se posee para que se ajuste con el propósito del sistema.
- Considerar que la valuación incidental no tendría que ser fija, ya que las características de los valores en los que se valúa afectan a todos los clientes.
- Evaluar si se crea una nueva zona de peligro por diseño de tráfico.
- La propuesta es una idea de valor que permite saber con exactitud dónde se incrementa el riesgo o las zonas de peligro, para con ello poder tomar decisiones informadas.

Fase 2:

- Creación de la migración a la nube para poder tener cómputo distribuido en las terminales que desee.

Fase 3:

- Creación de la API para poder realizar la extracción de datos hacia las terminales.

Fase 4:

- Crear el pipeline para poder entrenar los modelos predictivos (parte en donde se utilizará la minería de datos y con ello poder realizar lo que desee). Modelo general de segmentación de principiales

Fase 5:

- Optimización para poder tener cómputo distribuido usando la plataforma de Amazon Prime y de esta manera poder optimizar recursos para que el costo de mantenimiento no se extienda y se pueda crear servicios alrededor del servicio principal.

Fase 6:

- Capacitación de los empleados para poder usar el sistema.

Fase 7:

- Creación de feedback tanto de clientes como usuarios.

Fase 8:

- Al tener un diferenciador de no solo medir por qué el incremento en zonas urbanas, sino que puede ser trasladable a otras zonas ya que el modelo segmenta las principales razones de incremento en una zona dada, puede portabilizarlo por zonas y tener un modelo de negocio más grande.

# 4) (10 Puntos) ¿Qué elementos se deben consideran para dar un nivel de confianza al Cliente?

- certeza de que el modelo predictivo sea preciso
- La rapidez con que el modelo pueda segmentar las principales causas para poder categorizarlas.
- La coherencia de los datos.

# Tema: Base de datos y Repositorio de datos Preguntas:

## 5) (10 Puntos) ¿Cuáles son las particularidades de un DW, un DM o un DL y qué características deben tener estas opciones para ser consideradas como una opción para una organización?

- Data Lake: el orden es no repudio
- Data warehouse: Tratamiento de datos

| Característica | Data Warehouse (DW) | Data Mart (DM) | Data Lake (DL) |
|---|---|---|---|
| Propósito | Almacenamiento y análisis de grandes cantidades de datos de diferentes fuentes | Almacenamiento y análisis de datos enfocados en una sola área de negocio | Almacenamiento de grandes cantidades de datos estructurados y no estructurados en su forma original |
| Datos | Integrados, orientados a la gestión y enfocados en el tiempo | Subconjunto de datos del DW enfocados en una sola área de negocio | Datos sin procesar, estructurados y no estructurados |
| Complejidad | Alta | Baja | Media |
| Estructura | Altamente estructurada | Semi-estructurada | No estructurada |
| Costo | Alto | - | Bajo |

# 6) (10 Puntos) Cuáles son las características de un Repositorio y de un servidor de Base de Datos

| Característica | Repositorio | Servidor de Base de Datos |
|---|---|---|
| Propósito | Almacenamiento centralizado y control de versiones de archivos y documentos | Almacenamiento y gestión de datos estructurados |
| Alcance | Desarrollo de software, control de versiones y colaboración | Aplicaciones empresariales, sitios web y análisis de datos |
| Datos | Archivos y documentos | Tablas, registros y consultas |
| Estructura | Jerárquica o plana, dependiendo del sistema de control de versiones utilizado | Tablas relacionales con claves primarias y foráneas |
| Seguridad | Autenticación y autorización de usuarios, cifrado de datos (especialmente en github) | Autenticación y autorización de usuarios, cifrado de datos, respaldos y recuperación de desastres |
| Enfoque | Guardar información muy especifica que sufre modificaciones muy amenudo y poder seleccionar la mejor version que cumpla el proposito | Almacenar grandes cantidades de información de manera ordenada. |

| Característica | Repositorio | Servidor de Base de Datos |
| --- | --- | --- |
| Ejemplos | Git, Mercurial,gitlab | MySQL, PostgreSQL, Oracle, Microsoft SQL Server |

# 7) (10 Puntos) Cuáles son los principales beneficios que tiene una organización el migrar a un ambiente Cloud

Básicamente la idea es que tanta responsabilidad en infrestructura quiere esta involucrado la organizacion,  por ejemplo si una organizacion unicamente quiere implementar su programacion lo puede hacer y toda la infrestructura se encarga la compañia de cloud, viceversa si la organizacion quiere encargarse de la infrestructura lo puede hacer, pero de problemas tecnicos del servidor fisicamente  se encarga la compañia de cloud.

Los beneficios:

- Es mas facil de hacer lo que quieras, ya que puedes ser participe de la infrestructura tanto como quieras
- Puedes escalar verticalemente cuanto quieras mientras pagues mas
- Puedes solicitar servicio a demanda, es decir solo pagas lo que utilizas
- La ventaja es que fisicamente todo el costo esta cubierto y no necesitas especialistas para cada area  fisica, eso se encarga el cloud.

# Tema: Modelos Predictivos

## 8) (10 Puntos) Cuáles con los principales modelos predictivos de Minería de Datos y describa sus características y en qué caso de aplicaría

1. Clustering: La cantidad de grupos, la distancia entre grupos y el criterio de similaridad utilizado para agrupar los datos son las principales características del clustering. En resumidas cuentas, agrupa y dime en que difieren, si quiero crear grupos que comparten carateristicas lo usaria
2. Decision Trees: Los nodos de la árbol de decisión representan los valores de las variables y los hojas representan las subconjuntos de datos, así como el criterio utilizado para segmentar los datos. Si quiero mediante numeros guiarme en un camino hacia posibles opciones este lo usaria

3. Neural Networks: La arquitectura de la red neuronal, la cantidad de capas y los números de neuronas en cada capa son las principales características de los neural networks utilizados para segmentación. Es el caballo de batalla para todo, ya tengo datos  y quiero perfeccionarlo para realizar una accion en concreto pero que yo le brindo una cantidad  grande iformacion , mediante las capas lo puedo adecuar, por ello es una gran herramienta para tomar decisiones muy precisas, con las que los numeros ya estan enterados, aunque es mas dificil de aplicar y se extiene mucho.