METODOLOGÍA DE EXPERIMENTACIÓN EN BIG DATA
Por: Msc. Carlos Andres Lopez

Modelado de experimento.

1) Modelo o descripcion de los datos.
- Año
- trimestres
- mes_trimestre
- empresa
- empaquetado
- tipologia
- medio_atencion
- numero_quejas

Se debe estar seguro sobre el contenido o descripcion de los datos.
Objetivo es contrastar la observación de los datos del cientifico de datos con la vision de la mepresa.

2) Definir datos con valor, seleccionar datos importantes para el análisis.
Ejemplo Datos:
- año
- trimestres
- empresa
- tipologia
- medio de atencion
- numero_quejas

Conteo de datos y porcentaje (seleccionados 11 descarados 18):
6/ 8 en porcentaje 75% procesar / 25% descartar

3) Descripción de datos. Definir tipo de dato y posible explotación o exploración.
Si son numéricos:
- mínimo
- máximo
- promedio
- desv. Estándar

Si son texto:
- longitud de la cadena (x)
- cadena iterable (x)
- formato de cadena (x)
- fecha (x)
- codificación
- agrupacion

Los procesos de transformacional debe apuntar a generar este tipo de procesos o resultados.

4) Establecer peso o valor basado en importancia, separados por tipo:
Numéricos:

- año 2
- trimestre 3
- numero_quejas 5

Texto:

- empresa 5
- tipologia 4
- medio_atencion 3

Generamos escala de calificacion, 0 poco importante - 5 muy importante y calificamos cada dato.

En ocasiones las cadenas son numeros y se debe definir si se usan como números o como cadenas.
Conteo de datos y porcentaje (actualizar porcentaje):
Nos quedamos con:

- año
- trimestre
- numero_quejas
- empresa
- tipología

5/8 en porcentaje 62.5% procesar / 37.5% descartar

5) Modelar la pregunta.

Objetivo análisis: establece limites alcanzables del análisis, definición de tesis.
Vamos a plantear algunas preguntas que podríamos responder analizando los datos.

- Cual es la empresa con mayor numero de quejas?
- Cual es la empresa con menor numero de quejas?
- Cual es el tipo de queja mas común?
- Cual es el tipo de queja menos común?
- Cual es el trimestre donde se presentan mas quejas?
- Cual es el trimestre donde se presentan menos quejas?
- Cual es la empresa que trimestralmente con mayor numero de quejas?
- Cual es la empresa que trimestralmente con menor numero de quejas?
- Cual es le trimestre con mayor numero de quejas?
- Cual es le trimestre con menor numero de quejas?
- Cual es el tipo de queja mas común por trimestre?
- Cual es el tipo de queja menos común por trimestre?

Si N preguntas se puede proyectar, el promedio de explotación de es de tres productos por pregunta. Si se tienen 10 objetivos entonces se generan 30 productos de análisis.
Al combinar 2 objetivos se pueden generar mas productos.

Los entregales básicos son:

- Estadísticos de la observación
- La grafica de la observación
- Análisis de los estadísticos

El análisis de los estadísticos son las propuestas de valor que pueden surgir de las observaciones.

6) Modelar productos de datos.
Se deben enumerar los posibles productos a los que apuntará el análisis.
Se establecerá el objetivo general.

Generar información de tipo de queja por empresa y por año/trimestre.

7) Establecer modelos de consulta, agregación y desagregacion de datos.
Este paso puede requerir modelos de imputación de datos (transformación de datos crudos a datos pre procesados)

- Crear una conjunto de datos con las columnas que se emplearan para contestar las preguntas
- Tomar una pregunta y definir los pasos de pre proceso de datos
- Plantear alternativas de análisis:
  - cantidades
  - promedios
  - mejores, peores
  - graficas
  - listas

Preproceso: tipo_queja. Se requiere generar codificación, dará como resultado el agrupamiento.

Agregación: empresa y tipo_queja, a este resultado agregar año/trimestre

8) Imputación de datos (ETL)

Se enumeran las consultas o filtros que serán necesarias.

========
9) Generación descriptiva de datos (primer paso analítico)
- Análisis y procesamiento de datos (estadístico)
- Flujo de trabajo (workflow) vs sopa de scripts
- Visualización
- Generación de micro sitio
- Generación de micro servicio

10) Minería de datos (extraer conclusiones generales del modelo de datos)
- Sistemas de reportes
- Modelos de almacenamiento
- Perspectiva de implementaciones futuras

11) Desarrollo del modelo, implementar el modelo investigado a gran escala

Trabajos futuros:
12) Modelo prescriptivo
13) Modelo predictivo (optimización)