**Introducción a la ciencia de datos**

**Objetivo General:** Identificar la calidad y estructura de los datos históricos con los que cuento, basado en ese análisis poder tomar una decisión sobre qué modelo Machine learning se adapta mejor a mi problema y resuelve mi necesidad. Así como de que herramientas se necesita y el proceso a llevar a cabo.

A screenshot of a computer

Description automatically generated

**Fuente**: Video :Curso de análisis exploratorio (EDA) y visualización de datos por Bootcamp de ciencia de datos, Código Facilito.

-Como primer paso hay que conocer los datos con los que contamos…A screenshot of a computer

Description automatically generated**Fuente**: Video :Curso de análisis exploratorio (EDA) y visualización de datos por Bootcamp de ciencia de datos, Código Facilito.

A close-up of a text

Description automatically generated

* George Pólya : How to solve It ; Libros de matematicas

\*\*Existen estándar en el mundo de la ciencia de datos, Y también esta presenta las ISOS. Una de ellas es la ISO 8601 que dice que se debe seguir un formato en las fechas : [ISO 8601 - Wikipedia, la enciclopedia libre](https://es.wikipedia.org/wiki/ISO_8601)

A white paper with black text

Description automatically generated

Se establece que cuando se trabaja con un conjunto de datos estos siempre deben de estar acompañados de un Diccionario de datos; Cada vez que hay un conjunto de datos, lo acompaña un conjunto de datos.

¿Qué nos dice un diccionario de datos?

….Esta es una tabla que tiene n columnas, y esta columna y esta columna es formato Texto, esta es formato numérico. –Ahí no se respeta estándares o esta estandarizado el formato de un diccionario de datos, pero entre más detallado mucho mejor.

DataFrame:

Están desarrollados en base a Numpy. Cada una de las columnas del dataframe, es en esencia, un arreglo de de numpy y que además tiene un índice. Cuando se forma un dataframe, todas las series que están en un dataframe, la cual es un arreglo de numpy con índice, comparten el mismo índice: 0,1,2,3,4.

A screenshot of a computer

Description automatically generated

\*\*Recomendación: Guardar y almacenar tabla de datos en formato .parquet. Las cuales se manejan en el big data y se manejan en hadoop.

¿Por que usar archivos .parquet? Porque a comparación de archivos .CSV los archivos .parquet no se altera su formato cuando es transformado a un dataframe, y no tienes que modificar algunas detalles que se hayan movido. Y a comparación con Excel, parquet es mucho mas liviano cuando se trabaja con muchos datos. Referido a datos tabulares.

* Existe una funcion para poder convertir CSV a Parquet.

A screenshot of a computer code

Description automatically generated

Conversión de archivo CSV a Parquet.