

Aspectos relevantes para el desarrollo
del proyecto T1 de Análisis de Datos

Septiembre, 2024

Prof. Boris Panes

Objetivo general del proyecto T1

El objetivo de este proyecto es producir un video donde se presenta la descripción de un proceso de preparación de datos. El objeto del analisis es una conjunto de datos seleccionados por cada grupo que permita ejecutar las tareas basicas de preparacion de datos. Para ser especificos se recomienda seguir el algoritmo básico de preparación de datos planteado en la clase 2 (pagina 3 de este documento).

Estructura de los equipos de trabajo

- 1- Responsable de búsqueda de datos y descripción del objetivo usando Kaggle
- 2- Responsable de realizar el análisis de datos en Notebook de python
- 3- Responsable de armar la presentación en un PPT
- 4- Responsable de exponer el trabajo en formato video de ZOOM

Entrega del trabajo T1: 15-09-24

Requerimientos mas específicos

Materiales y formato de entrega:

Set de datos en formato csv (comma separated values) plano, multicolumna
Notebook escrito en python donde se carga el set de datos y se realiza el análisis
Presentación en PDF, 10-15 láminas
Video describiendo el trabajo **(envio por correo)**

Contenido del Video:

Introducción al trabajo por parte del presentador (1 min)
Presentación de los integrantes y resumen de su contribución (5 minutos)
Discusión del notebook (14 minutos)

Algoritmo de limpieza y transformación de datos

Tiempo total aproximado: 15-20 minutos

Algoritmo de limpieza y transformación de datos (Contenido de un notebook)

Nota interesante: en general se puede apreciar que los set de datos de **Kaggle** no contienen muchos datos erróneos, nulos o con formatos inconsistentes, al menos ninguno contiene todas estas posibilidades juntas al mismo tiempo

Las acciones relacionadas con **limpieza y transformación de datos se pueden chequear todo el tiempo**, pero no necesariamente se deben aplicar.

El contenido del notebook y la presentación debería contener los siguientes pasos

Lectura del set de datos

Exploración de columnas y filas

Búsqueda de valores nulo

Protocolo de acción al respecto es contextual

Distinción entre contenidos numéricos y texto (Object)

Búsqueda y tratamientos de registros duplicados

Aplicación de proceso de normalización

Aplicación de proceso de estandarización

Evaluación y aplicación de ingeniería de columnas

*En este primer notebook
no se requieren discutir
relaciones de 2 o mas variables
como scatter plots y boxplots*