Organización de Datos (75.06/95.58)
Primer Cuatrimestre 2021
Trabajo Práctico N°2

Curso 1: Argerich

Introducción:

En el año 2015 Nepal fue afectado por el terremoto Gorkha, un sismo que registró una magnitud de 7.8 en la escala Richter y tuvo su epicentro en la ciudad de Kathmandu. Aproximadamente 600,000 estructuras en el centro y pueblos aledaños fueron dañadas o destruidas. Un análisis posterior al sismo llevado por la Comisión Nacional de Planeamiento de Nepal comunicó que la pérdida total económica ocasionada por el terremoto fue de aproximadamente $7 mil millones (USD; NPC, 2015).

El dataset para el presente TP está compuesto de encuestas realizadas por Kathmandu Living Labs y el Central Bureau of Statistics y contiene información sobre el impacto del terremoto, estado de viviendas y estadísticas sociodemográficas.

Particularmente el dataset se enfoca en cómo eran las condiciones de una determinada vivienda y cuál fue su grado de daño luego del accidente.

Objetivo:

El TP consiste en realizar un análisis exploratorio de los datos provistos con el objetivo de determinar características y variables importantes, descubrir insights interesantes, y analizar la estructura de los mismos.
Tener en cuenta que el trabajo realizado podrá ser utilizado en el TP2 y se puede considerar un paso previo al mismo.

Datos

Se recomienda previamente ver la explicación de la plataforma donde encontraran los datos. Pueden acceder a dicho video haciendo clic **aquí**.

Los datos a analizar se pueden encontrar haciendo clic **aquí**.

La variable de interés es *"**damage_grade**"*, que representa el nivel del daño que recibió la edificación y puede ser:

- 1 - Low damage
- 2 - Medium damage
- 3 - Serious damage

Referencia de los datos:

Los datos se encuentran en dos archivos diferentes:

- "train_values.csv": aquí encontraremos 38 features que describen a la edificación identificada por un id.
- "train_labels.csv": aquí encontraremos la variable `damage_grade` y el correspondiente id de la edificación.

Es importante destacar que algunos datos del tipo categórico se encuentran ofuscados con caracteres ASCII. Esto implica que si bien la consistencia se mantiene, no conocemos los valores originales de dichos datos. Los caracteres ASCII repetidos en distintas columnas **no** representan el mismo valor.

- `building_id` (tipo: ID): identificador único de la edificación.
- `geo_level_1_id`, `geo_level_2_id`, `geo_level_3_id` (tipo: enteros): región geográfica en la cual la edificación existe, desde la más general (level 1) a la más específica (level 3). Valores posibles:
  - level 1: 0-30,
  - level 2: 0-1427,
  - level 3: 0-12567.
- `count_floors_pre_eq` (tipo: entero): número de pisos en la edificación antes del terremoto.
- `age` (tipo: entero): antigüedad de la edificación en años.

- `area_percentage` (tipo: entero): superficie normalizada ocupada por la edificación.
- `height_percentage` (tipo: entero): altura normalizada ocupada por la edificación.
- `land_surface_condition` (tipo: categórico): condición de la superficie terrestre donde el edificio fue construido. Valores posibles: n, o, t.
- `foundation_type` (tipo: categórico): tipo de cimientos usados cuando se construyó la edificación. Valores posibles: h, i, r, u, w.
- `roof_type` (tipo: categórico): tipo de techo usado cuando se construyó la edificación. Valores posibles: n, q, x.
- `ground_floor_type` (tipo: categórico): tipo de construcción usado en la planta baja cuando se construyó la edificación. Valores posibles: f, m, v, x, z.
- `other_floor_type` (tipo: categorical): tipo de construcción usado en otros pisos cuando se construyó la edificación (exceptuando el techo). Posibles valores: j, q, s, x.
- `position` (tipo: categórico): orientación de la edificación. Posibles valores: j, o, s, t.
- `plan_configuration` (tipo: categórico): formato de construcción de la edificación (para diseño sísmico). Valores posibles: a, c, d, f, m, n, o, q, s, u.
- `has_superstructure_adobe_mud` (tipo: binario): variable que indica si la edificación fue construida con adobe/barro.
- `has_superstructure_mud_mortar_stone` (tipo: binario): variable que indica si la edificación fue construida con barro - piedra.
- `has_superstructure_stone_flag` (tipo: binario): variable que indica si la edificación fue construida con piedra.
- `has_superstructure_cement_mortar_stone` (tipo: binario): variable que indica si la edificación fue construida con cemento - piedra.

- `has_superstructure_mud_mortar_brick` (tipo: binario): variable que indica si la edificación fue construida con barro - ladrillos.
- `has_superstructure_cement_mortar_brick` (tipo: binario): variable que indica si la edificación fue construida con cemento - ladrillos.
- `has_superstructure_timber` (tipo: binario): variable que indica si la edificación fue construida con Timber (madera específica para la construcción).
- `has_superstructure_bamboo` (tipo: binario): variable que indica si la edificación fue construida con Bambú (caña).
- `has_superstructure_rc_non_engineered` (tipo: binario): variable que indica si la edificación fue construida con concreto reforzado no-diseñado.
- `has_superstructure_rc_engineered` (tipo: binario): variable que indica si la edificación fue construida con concreto reforzado diseñado.
- `has_superstructure_other` (tipo: binario): variable que indica si la edificación fue construida con otro material.
- `legal_ownership_status` (tipo: categórico): estado legal de la tierra donde la edificación fue construida. Valores posibles: a, r, v, w.
- `count_families` (tipo: entero): número de familias que vivían en la edificación.
- `has_secondary_use` (tipo: binario): variable que indica si la edificación era usada con un uso secundario.
- `has_secondary_use_agriculture` (tipo: binario): variable que indica si la edificación era usada con propósitos de agricultura.
- `has_secondary_use_hotel` (tipo: binario): variable que indica si la edificación era usada como oficina de gobierno.
- `has_secondary_use_rental` (tipo: binario): variable que indica si la edificación se alquilaba.
- `has_secondary_use_institution` (tipo: binario): variable que indica si la edificación era usada como sede de una institución.

- `has_secondary_use_school` (tipo: binario): variable que indica si la edificación era usada como escuela.
- `has_secondary_use_industry` (tipo: binario): variable que indica si la edificación era usada con propósitos industriales.
- `has_secondary_use_health_post` (tipo: binario): variable que indica si la edificación era usada como puesto de salud.
- `has_secondary_use_gov_office` (tipo: binario): variable que indica si la edificación era usada como oficina de gobierno.
- `has_secondary_use_use_police` (tipo: binario): variable que indica si la edificación era usada como estación de policía.
- `has_secondary_use_other` (tipo: binario): variable que indica si la edificación era usada con otro uso secundario.

Requisitos:

Los requisitos de la entrega son los siguientes:

- El análisis debe estar hecho en Python Pandas o R.
- El análisis debe entregarse en formato PDF vía gradescope. En el informe no debe contener código.
- Informar el link a un repositorio Github en donde pueda bajarse el código completo para generar el análisis. El repositorio debe ser público.

Evaluación:

La evaluación del TP1 se realizará en base al siguiente criterio:

- Originalidad del análisis exploratorio.
- Calidad del reporte. ¿Está bien escrito? ¿Es claro y preciso?
- Calidad del análisis exploratorio
    - Qué tipo de preguntas se hacen y de qué forma se responden, ¿es la respuesta clara y concisa con respecto a la pregunta formulada?
    - ¿Se plantean hipótesis sobre lo observado?
    - ¿Se realiza un mínimo preprocesamiento o limpieza de los datos?
    - ¿Se profundiza en los datos más allá de un simple análisis estadístico?

- Calidad de las visualizaciones presentadas.
    - ¿Tienen todos los ejes su rótulo?
    - ¿Tiene cada visualización un título?
    - ¿Están numeradas las visualizaciones?
    - ¿Es entendible la visualización sin tener que leer la explicación?
    - ¿El tipo de plot elegido es adecuado para lo que se quiere visualizar?
    - ¿Es una visualización interesante?
    - ¿El uso del color es adecuado?
    - ¿Hay un exceso o falta de elementos visuales en la visualización elegida?
    - ¿La visualización es consistente con los datos?
    - ¿Presenta el grupo un listado de "insights" aprendidos sobre los datos en base al análisis realizado? ¿Es interesante?
- Conclusiones presentadas.

El grupo que realice el mejor análisis exploratorio obtendrá 10 puntos para cada uno de sus integrantes que podrán ser usados en el parcial además de ser publicado en el repositorio de la materia como ejemplo para los siguientes cuatrimestres.

Recursos:

Algunos recursos interesantes para visualizaciones en Python y R:

- https://python-graph-gallery.com/
- https://datavizproject.com/
- https://www.r-graph-gallery.com/