Guía de Análisis Exploratorio.
Proyecto

INTRODUCCIÓN

Para hacer una investigación formal es necesario que ésta se base en una situación problemática y por consiguiente un problema que justifique la investigación. Es importante revisar la teoría que rodea la problemática y los antecedentes de investigaciones similares. La investigación puede ser aplicada a cualquier campo como salud, educación, agricultura, seguridad, finanzas, economía y negocios, etc. Para hacer la investigación, es necesario contar con un buen conjunto de datos. Afortunadamente hay muchos sitios donde se pueden encontrar datos confiables y de uso gratuito. Basta con explorar sitios como:

- Instituto Nacional de Estadística (Guatemala): https://www.ine.gob.gt/portal-estadistico-1-0/
- Naciones Unidas (global): https://data.un.org/
- FAO (Naciones Unidas): https://data.apps.fao.org/catalog/dataset
- Banco Mundial: https://data.worldbank.org/

Además del INE, a nivel nacional, la mayoría de los ministerios grandes: Educación, Agricultura, etc., también han puesto a disposición pública múltiples conjuntos de datos.

El objetivo del proyecto es elaborar modelos de aprendizaje de máquinas para predecir y/o clasificar basados en la selección de variables respuesta. En esta entrega particular, el objetivo es explorar el conjunto de datos que el equipo de proyecto seleccione.

CONJUNTO DE DATOS

Para su proyecto, piense en un área que sea de interés para su grupo y luego investigue qué sitios pueden tener datos que puedan serles útiles. Algunos requisitos que deben cumplir sus datos son:

- Los datos deben cubrir un período igual o mayor de 10 años.
- Es imperativo trabajar con datos originales. Algunos sitios ofrecen datos “agregados”, es decir, que ya han sido procesados y los datos ofrecidos son un resumen de los datos originales. Estos conjuntos de datos agregados no serán aceptables para el proyecto.
- Es posible que no encuentre todas las características (variables) que necesita en un solo archivo. Para remediar esto, se pueden unir diferentes conjuntos de datos
- El número de observaciones (registros) que tenga el archivo consolidado no debe ser menor de 1,000
- El número de características que tiene cada observación debe ser mayor a 7.

En el caso del INE se sugiere considerar los siguientes conjuntos de datos:

En la de estadísticas vitales, podremos encontrar 5 conjuntos de datos por año desde 2009 hasta 2021 (https://www.ine.gob.gt/ine/vitales/):

- Nacimientos.

- Matrimonios.
- Divorcios.
- Defunciones.
- Defunciones Fetales

En la de violencia se pueden encontrar 5 conjuntos de datos que resultan interesantes también, pero particularmente 3 resultan interesantes:

- Hechos delictivos
- Violencia intrafamiliar
- Violencia en contra de la mujer y delitos sexuales.

ACTIVIDADES

1. Explore los datos para encontrar preguntas interesantes y líneas de investigación. Para esto:
    a. Comience describiendo cuántas variables y observaciones tiene disponibles, y el tipo de cada una de las variables.
    b. Haga un resumen de las variables numéricas e investigue si siguen una distribución normal y tablas de frecuencia para las variables categóricas, escriba lo que vaya encontrando.
    c. Cruce las variables que considere que son las más importantes para hallar los elementos clave que lo pueden llevar a comprender lo que está causando el problema encontrado.
    d. Haga gráficos exploratorios que le dé ideas del estado de los datos.
    e. Haga un agrupamiento (clustering) e interprete los resultados.
2. Una vez que haya explorado los datos
    a. Describa la situación problemática que lo lleva a acotar un problema a resolver.
    b. Enuncie un problema científico y unos objetivos preliminares.
    c. Describa los datos que tiene para responder el problema planteado. Esto incluye el estado en que encontró el, o los, conjunto(s) de datos y las operaciones de limpieza que se realizaron, en caso de que hayan sido necesarias.
    d. Escriba unas conclusiones con los hallazgos encontrados durante el análisis exploratorio

EVALUACIÓN

Notas: Para tener derecho a calificación debe mostrar evidencias de contribuciones significativas tanto en el repositorio como en el documento.

- **(10 puntos) Situación Problemática:** Describe la situación problemática que da lugar al problema.
- **(10 puntos). Problema científico:** Se enuncia el problema científico que se desprende de la situación planteada. Se comprende bien cuál es el problema.

- **(10 puntos). Objetivos:** Se plantean los objetivos a cumplir para darle solución al problema planteado. Se enuncian, al menos, un objetivo general y 2 específicos. Los objetivos deben ser medibles y alcanzables durante la investigación.
- **(20 puntos). Descripción de los datos:** Se describen los datos, tanto las variables y observaciones como las operaciones de limpieza que se le hicieron si fueron necesarias.
- **(30 puntos). Análisis Exploratorio:**
    - Estudia las variables cuantitativas mediante técnicas de estadística descriptiva
    - Hace gráficos exploratorios como histogramas, diagramas de cajas y bigotes, gráficos de dispersión, que ayudan a explicar los datos.
    - Analiza las correlaciones entre las variables, trata de explicar los datos atípicos (outliers) y toma decisiones acertadas ante la presencia de valores faltantes.
    - Estudia las variables categóricas.
    - Elabora gráficos de barra, tablas de frecuencia y de proporciones
    - Explica muy bien todos los procedimientos y los hallazgos que va haciendo.
    - Determina la tendencia al agrupamiento y el mejor número de “clusters” a utilizar.
    - Hace el agrupamiento con cualquiera de los algoritmos estudiados.
    - Verifica la calidad del agrupamiento, incluya el método de la silueta.
    - Interpreta los grupos, usando para eso las variables numéricas y categóricas dentro de cada grupo.
- **(20 puntos). Hallazgos y conclusiones:**
    - Hace un resumen de los hallazgos en el análisis exploratorio
    - Le pone un nombre a los grupos que reflejen sus características principales
    - Llega a conclusiones sobre los siguientes pasos a seguir.

MATERIAL A ENTREGAR

- Vínculo de Google docs con el informe de análisis exploratorio. Se debe poder verificar el historial de cambios
- Script de R (.r o .rmd) o de Python que utilizó para responder las preguntas con el código utilizado.
- Vínculo de repositorio de github