Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Estudios de Postgrado
Introducción a la Minería de Datos
Cuarto Trimestre 2024
Catedrático: M.A. Ing. Kevin Lajpop

Proyecto - parte 1

Objetivo

El objetivo de este proyecto es que los estudiantes apliquen técnicas de minería de datos para analizar conjuntos de datos relacionados con distintas problemáticas en Guatemala. A partir de estos análisis, se espera que generen propuestas concretas para abordar y mejorar estas problemáticas.

Conjunto de datos

A cada estudiante se le asignará un conjunto de datos de manera aleatoria para poder desarrollar su proyecto, entre las opciones de los conjuntos de datos hay temáticas como salud, seguridad, educación, medio ambiente, entre otros, los datos puede incluir diversos tipos y es necesario que el estudiante deba limpiar los datos. Es importante mencionar que si necesitan más datos los deben de buscar y complementar su análisis e indicar que fuente agregó a su proyecto.

Enlace de dataset asignados:
https://docs.google.com/spreadsheets/d/1l7ghOX444tNeFTBpWTaL4ykmvMrALSwclYhOUHBoGjE/edit?usp=sharing

Descripción general

El estudiante luego de haber seleccionado los datos con los que desea trabajar el estudiante debe de realizar lo siguiente:

1. Reglas de asociación Apriori

- Empleando el algoritmo **Apriori**, los estudiantes deben descubrir patrones y relaciones interesantes entre diferentes variables en los conjuntos de datos (mínimo 4 patrones interesantes).
- Deben identificar asociaciones significativas que puedan ayudar a comprender mejor la problemática en cuestión y sugerir posibles intervenciones.
- Se espera que proporcionen interpretaciones sobre el significado y la relevancia de estas reglas de asociación para abordar la problemática identificada.

2. Reglas de asociación FP-Growth

- Empleando el algoritmo **FP-Growth**, los estudiantes deben descubrir patrones y relaciones interesantes entre diferentes variables en los conjuntos de datos (mínimo 4 patrones interesantes).
- Deben identificar asociaciones significativas que puedan ayudar a comprender mejor la problemática en cuestión y sugerir posibles intervenciones.
- Se espera que proporcionen interpretaciones sobre el significado y la relevancia de estas reglas de asociación para abordar la problemática identificada.

3. Análisis de clúster

- Utilizando técnicas de clustering (K-Means), el estudiante debe identificar grupos o segmentos de datos relacionados con una problemática específica seleccionada en el set de datos.
- Deben explorar posibles correlaciones entre los grupos identificados y variables relevantes en los conjuntos de datos.
- Se espera que proporcionen interpretaciones significativas de los resultados del clustering y discutan cómo estos grupos pueden ser útiles para comprender y abordar la problemática identificada.
- **Se debe de graficar los clúster.**

4. Propuestas

Basándose en los resultados de los análisis anteriores, el estudiante debe generar propuestas concretas para abordar y mejorar la problemática identificada. Estas propuestas deben ser viables y estar respaldadas por los hallazgos del análisis de datos. Para realizar sus propuestas debe de basarse con documentación científica,

así mismo debe realizar una investigación del contexto guatemalteco para validar la viabilidad de sus propuestas.

Presentación y calificación

- Es importante tener en cuenta que la documentación técnica debe de ser lo más explícita para que el análisis el catedrático pueda implementarlo en su computadora.
- Compartir el repositorio de GitHub al usuario: **Adiel13**
- Si bien es cierto no hay un límite ni mínimo de páginas para presentar sus propuestas, recuerde que debe de ser lo suficientemente nutrido y creativo para poder exponer los puntos necesarios, ni ser tan denso que rompa la lectura en la cual el único camino sea dejar de leer el documento.
- Citar con formato APA 7.
- La documentación técnica debe de contener bibliotecas utilizadas, algún instrucción en R que no hayamos visto en clase y sea necesario para su análisis, forma de ejecutar el código y todo lo necesario para que el catedrático pueda replicar el proyecto.

Restricciones

1. El proyecto debe de realizarse de manera individual
2. El lenguaje a utilizar es R
3. El repositorio debe de contar con 5 commits como mínimo
4. Si no utiliza un año del dataset que se le ha compartido, debe indicar el porqué se está descartando.
5. No se tolerarán copias en los trabajos.

Fecha de entrega

- Último día: **9 de noviembre a las 23:59**

Entregables

1. Repositorio en GitHub con código en R el cual debe incluir todo lo que están realizando para su análisis, como podría ser limpieza y transformación de datos.
2. Documentación técnica en su repositorio en formato Markdown.

3. Documento con resultados y propuestas que realiza basándose en los resultados obtenidos de los distintos algoritmos (dicho documento con bibliografía).
4. Plataforma de entrega: Aula Virtual.

Evaluación:

Aspecto a evaluar	**Punteo**
Repositorio GitHub con 5 commits mínimo	5
Documentación técnica Markdown: Explica detalladamente el código Instrucciones de como implementarlo en otro ambiente	15
Implementación de algoritmo de reglas de asociación: Código Se replica en otro ambiente	25
Implementación de algoritmo de clúster: Código Se replica en otro ambiente Gráfica	25
Documento con propuestas: Basado en su código Respeta las citas APA Documento entendible Documento no es denso Bibliografía científica Entre otros aspectos	30
Total	100