**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 2

## 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. Predicción por medio de árboles de decisión

- Empleando el algoritmo de **árboles de decisión**, los estudiantes deben descubrir 4 situaciones las cuales deseen realizar la predicción. Deben de ser

predicciones diferentes, no es válido utilizar la misma predicción pero con diferentes variables de entrada.
- Deben de ser situaciones que fueron identificadas en su proyecto parte 1, ya sea con FP-Growth, Apriori o K-Means.
- Se debe de presentar diversas situaciones o escenarios en los cuales su modelo se ponga a prueba, por ej. para predecir una nota aprobatoria o no, en un primer escenario se probará “mujer de 21 años, llevando el curso por primera vez y en el segundo semestre”, para el segundo se probará “mujer de 27 años, llevando el curso por cuarta vez y en el primer semestre”, debe de evidenciar todos los casos que le sean posibles.
- Se debe de graficar los árboles.
- Se espera que proporcionen interpretaciones sobre el significado y la relevancia de las predicciones realizadas por su modelo.

## 2. Predicción por medio de bosques aleatorios

- Empleando el algoritmo **Random Forest**, los estudiantes deben profundizar en mínimo 2 situaciones, basadas en los árboles de decisión, sobre las cuales deseen realizar la predicción. Deben de ser predicciones diferentes, no es válido utilizar la misma predicción pero con diferentes variables de entrada.
- Se debe de presentar diversas situaciones o escenarios en los cuales su modelo se ponga a prueba, por ej. para predecir una nota aprobatoria o no, en un primer escenario se probará “mujer de 21 años, llevando el curso por primera vez y en el segundo semestre”, para el segundo se probará “mujer de 27 años, llevando el curso por cuarta vez y en el primer semestre”, debe de evidenciar todos los casos que le sean posibles.
- Se debe de graficar los bosques aleatorios.
- Se espera que proporcionen interpretaciones sobre el significado y la relevancia de estas reglas de asociación para abordar la problemática identificada.

## 3. Predicción por medio de redes neuronales

- Utilizando técnicas de **redes neuronales**, el estudiante debe de generar redes neuronales de dos modelos realizados con árboles de decisión.
- Se debe de presentar diversas situaciones o escenarios en los cuales su modelo se ponga a prueba, por ej. para predecir una nota aprobatoria o no, en un primer escenario se probará “mujer de 21 años, llevando el curso por primera vez y en el segundo semestre”, para el segundo se probará “mujer de 27 años, llevando el curso por cuarta vez y en el primer semestre”, debe de evidenciar todos los casos que le sean posibles.

- Se espera que proporcionen interpretaciones sobre el significado y la relevancia de las predicciones realizadas por su modelo y la comparación contra el modelo expuesto en árboles de decisión.

### 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 o Python 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 para árboles de decisión y random forest
3. El lenguaje a utilizar es Python para redes neuronales
4. El repositorio debe de contar con 5 commits como mínimo
5. Si no utiliza un año del dataset que se le ha compartido, debe indicar el porqué se está descartando.
6. No se tolerarán copias en los trabajos.

# Fecha de entrega

- Último día: **7 de diciembre a las 23:59**

# Entregables

1. Repositorio en GitHub con código en R y Python 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:<br>Explica detalladamente el código<br>Instrucciones de como implementarlo en otro ambiente | 10 |
| Implementación de árboles de decisión:<br><br>Código<br>Se replica en otro ambiente<br>Gráfica | 15 |
| Implementación de algoritmo de random forest:<br><br>Código<br>Se replica en otro ambiente<br>Gráfica | 25 |
| Implementación de redes neuronales:<br><br>Código<br>Se replica en otro ambiente | 25 |

| | |
|---|---|
| Documento con propuestas:<br><br>Basado en su código<br>Respeta las citas APA<br>Documento entendible<br>Documento no es denso<br>Bibliografía científica<br>Entre otros aspectos | 20 |
| Total | 100 |