Universidad de San Carlos de Guatemala
Escuela de Estudios de Postgrado
Maestría en Ingeniería para la Industria
Con Especialidad en Ciencias de la Computación
Introducción a la Minería de Datos

# Proyecto Final

| | |
|---|---|
| **Fecha de entrega** | 29 de abril |
| **Proyecto del curso:** | 50 puntos de zona |
| **Integrantes:** | 4-5 integrantes por grupo. |
| **Objetivo general:** | Comprender y aplicar los algoritmos básicos utilizados en la Minería de Datos. |

El presente trabajo evaluará los siguientes temas cubiertos en el curso de Introducción a la Minería de Datos:

- Entradas
  - Conceptos
  - Instancias
  - Atributos
- Representación del conocimiento
  - Tablas
  - Modelos lineales
  - Árboles
  - Reglas
  - *Clusters*
- Algoritmos
  - Modelación estadística
  - Árboles de decisión
  - Modelos lineales
  - *Clustering*
- Credibilidad
  - Pruebas y entrenamientos
  - Predicción del rendimiento
  - Validación cruzada.

**Análisis de Minería de Datos:**

Analizarán en grupo uno de los *datasets* que se encuentran publicados en el sitio Web del curso, este, no puede ser repetido entre grupos (un único *dataset* debe de ser analizado por un solo grupo). Su análisis objetivo hará uso de los conocimientos transmitidos en clase y en los textos de clase. La selección del *dataset* será muy importante para el análisis que efectuarán de acuerdo con los algoritmos aplicados a la Minería de Datos.
Las conclusiones de su análisis, deberá de contener la interpretación de los resultados obtenidos utilizando el lenguaje de programación *Python*.

Un representante deberá enviar un correo con los integrantes del grupo y la preferencia de los *datasets* en orden jerárquico de prioridad (en el anexo 1 se explica como deberá de enviarse la información para elegir el *dataset* de su preferencia).

**Su análisis deberá contener:**

1. Listado de cada uno de los atributos del *dataset*, conteniendo el tipo de dato, descripción y el análisis que obtendría de cada uno de ellos.
2. Explicación del problema a resolver.
3. Definición de los puntos relevantes del *dataset* a analizar, así como de los algoritmos seleccionados (se debe seleccionar un mínimo de 2 algoritmos).
4. De los algoritmos seleccionados para resolver el problema deberá de describir justificando: ventaja entre cada uno de ellos, recomendación de elección.
5. Descripción de los limites de rendimiento optimo.
6. Justificación de los intervalos de confianza utilizados.
7. Análisis de la estimación de la tasa de éxito y de la estimación de error.
8. Análisis, fundamento y descripción de la muestra de entrenamiento.
9. Planteamiento de la hipótesis del problema a resolver.
10. Interpretación de los resultados obtenidos.
11. Sus conclusiones deberán ser soportadas por gráficos y anotaciones en los gráficos que se utilicen.

**Recomendaciones:**
Recuerde que un análisis riguroso será prueba fiel de su calidad como analista de Minería de Datos. Todo contenido deberá contener análisis puntual de dicha información. Analice la información relevante de forma que le permita presentarla en su trabajo (tablas, gráficos, etc.) para respaldar sus conclusiones. Su análisis, tanto fundamental como técnico deberá estar respaldado con datos y gráficos, indicando en ellos el porqué de su uso. Es fundamental la presentación, ortografía y redacción de su trabajo.

# Anexo 1

**Elección del *dataset*:**

En el Aula Virtual, en la sección: “Proyecto Final -> *datasets”*, se encuentran publicados 8 *datasets,* el cual, el grupo deberá de elegir el de su preferencia.
Ya que un solo grupo puede trabajar un único *dataset* (relación de uno a uno), deberán de enviar un correo electrónico con el nombre de los integrantes del grupo, así como en orden de preferencia el listado de los 7 *dataset* que le gustaría trabajar.

**Ejemplo:**

1. Integrante 1
2. Integrante 2
3. Integrante 3
4. Integrante 4
5. Integrante 5

Preferencia de *datasets*:

| | |
|---|---|
| 1 | 9000+ Movies |
| 2 | Banks – Loan |
| 3 | Adult Income |
| 4 | NBA Players(1950-2017) |
| 5 | Netflix subscription fee Dec-2021 |
| 6 | Amazon Top 50 Bestselling Books 2009 - 2022 |
| 7 | Reprasentative Sample of Bitcoin Blockchain Data |
| 8 | Taxi trip data NYC.csv |

En el orden que se muestra, indica que la principal preferencia es el *dataset* “9000+ Movies”, su segunda preferencia es el *dataset* Banks – Loan, y así sucesivamente.

En el ejemplo anterior, el primer grupo en enviar el correo será asignado automáticamente el *dataset*: 9000+ Movies.

En el caso, un segundo grupo desea trabajar con el mismo *dataset* que se mencionó anteriormente, y es el mismo listado que se muestra en el ejemplo, será asignado el segundo *dataset* de su preferencia: “Banks – Loan”. Y así sucesivamente.

En caso un tercer grupo envié en prioridad No. 1, un *dataset* que no ha sido elegido por otro grupo, se le asignará automáticamente el *dataset* de su preferencia.

**Anexo 2**

**Rubrica de calificación:**

| | |
|---|---|
| **Introducción**<br>Resumen de los puntos más relevantes: *dataset* y selección del algoritmo | 10 |
| **Script de solución**<br>Solución del algoritmo elegido, utilizando código *Python*. | 40 |
| ***Output***<br>Selección apropiada para mostrar los resultados obtenidos | 10 |
| **Conclusiones**<br>Interpretación de los resultados obtenidos. | 30 |
| **Presentación, ortografía y redacción** | 10 |