**Segundo Examen Parcial – Inteligencia Artificial**

**Nombre**: _____

Catedrático: M.I.C Normado Ali Zubia Hernández

## Introducción:

La industria taquillera es una de las áreas más redituables hoy en día, lamentablemente no todas las películas son exitosas, es por eso que es muy importante conocer las variables que hacen a una película exitosa. Por eso se han requerido tus servicios para analizar dichos patrones y dar una respuesta a dicha problemática.

https://www.kaggle.com/c/tmdb-box-office-prediction

## Información:

1. Se tiene un conjunto de 7398 registros correspondientes a información de películas.
2. El dataset contiene un total de 21 atributos.
3. La etiqueta de clase la contiene el último atributo llamado “revenue”, que indican las ganancias obtenidas de cada película.
4. Para calcular la precisión del modelo se utilizará el método “Root Mean Squared Logarithmic Error (RMSLE)“.
5. 20% de la calificación total será por competencia, los lugares serán dados dependiendo del lugar en la tabla de competición:
    a. Primer lugar: 20%
    b. Segundo lugar: 15%
    c. Tercer lugar: 10%

## Actividad:

1. Se deberá analizar el conjunto de datos y plantear una estrategia de ataque con respecto a este análisis.

   En esta parte expondrán de que trata el problema, los atributos de los que se compone el conjunto de datos, que tipo de datos contienen cada uno de esos atributos y para que sirven cada uno de dichos atributos

2. Aplicar las siguientes técnicas de pre-procesamiento según considere adecuado:
   a. Lidiar con datos vacíos
   b. Lidiar con datos outliers
3. Aplicar las siguientes técnicas de reducción de dimensionalidad y comparar cual es la que tienen mejores resultados en los algoritmos:
   a. PCA
   b. Attribute subset selection (Puede ser la técnica de selectKBest o selección de atributos por medio de árbol de decisión)
4. Aplicar las siguientes técnicas de normalización y comparar cual es la que tienen mejores resultados en los algoritmos:
   a. MinMaxScaler
   b. StandardScaler
5. Aplicar los algoritmos vistos en clase para la solución del problema, además de un algoritmo nuevo no visto en clase.

   En este apartado se implementaran los algoritmos al conjunto de datos para obtener los mejores resultados. Al menos se deben de implementar los siguientes modelos:
   a. Redes Neuronales
   b. Árboles de decisión
   c. Bagging
   d. Boosting
   e. Random Forest

6. Utilizar para cada algoritmo utilizado los siguientes métodos de evaluación:
    a. Matriz de confusión
    b. Kappa statistic
    c. F-measure

   Realizar un análisis de los resultados y explicarlos.
7. Escribir las conclusiones de la iteración.

   Apartado donde se expondrán las observaciones de la iteración, incluyendo que estrategias piensa implementar para la siguiente iteración para mejorar el resultado.

## Entregables:

- Documento en formato PDF donde se expliquen las siguientes secciones:
    - Análisis del problema y estructura del conjunto
    - Pre-procesamiento
    - Implementación de modelo de machine learning
    - Conclusiones
- Repositorio donde se tiene almacenado el proyecto

- Nota: Es necesario al menos realizar 3 iteraciones, en las cuales cada iteración tendrá una conclusión acerca de que es necesario mejorar para obtener mejores resultados. Así mismo es necesario especificar que resultados se obtuvieron en dicha iteración.

Nota: La calificación será otorgada en base a los siguientes criterios:

- Técnicas de análisis y planteamiento de estrategia ……. 20%
- Pre-procesamiento ……42%
- Implementación de modelo de machine learning……..……30%
- Conclusiones ……8%