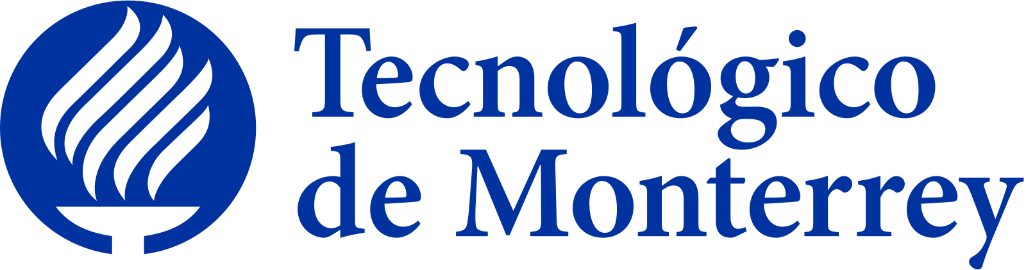
**Maestría en Inteligencia Artificial Aplicada**

**Curso: Inteligencia Artificial y Aprendizaje Automático**

**Tecnológico de Monterrey**



**Prof Luis Eduardo Falcón Morales**

**Adtividad de la Semana 9**

**Actividad 10: Taxonomía de Métricas de Clasificación**

* Aurelio Antonio Lozano Rabago A01081266
* Josias Ruiz Peña A00968460
* David González A01794025
* Jerson David Pérez Contreras A01793810
* Jose Alberto Mtanous Treviño A00169781

### Resumen: Comparación experimental del desempeño de métricas para clasificación

Las métricas de desempeño en clasificación son fundamentales para evaluar la calidad de los métodos de aprendizaje y los modelos generados, se analizaron 18 métricas de desempeño con la finalidad de ayudar a elegir la mas adecuada (o mas adecuadas) para cada aplicación.

#### Introducción

La correcta evaluación de modelos de aprendizaje automático es muy importante en temas de reconocimiento de patrones, esta evaluación puede apoyarse en 3 grupos de métricas:

* Métricas basadas en un umbral y entendimiento *cualitativo* del error: Se usan cuando se pretende minimizar el número de errores en un modelo
* Métricas basadas en entendimiento *probabilístico* del error: son útiles especialmente cuando se quiere evaluar la confiabilidad de los clasificadores
* Métricas basadas en que tan bien un modelo clasifica los ejemplos: se usan para seleccionar las mejores *n* instancias de un set de datos para una buena separación de clases.

#### Trabajos relacionados.

Hay varios estudios que muestran que, dependiendo las métricas usadas, se obtienen un mejor modelo. El propósito de este estudio es analizar que familias de algoritmos se comportan mejor según la familia de métricas a usar.

#### Métricas

Algunas ocasiones lo más complicado es entender la diferencia entre una buena clasificación y probabilidad.

* Umbral del error (cualitativas): Acc, MAvA, MFM, MAvG, KapS
* Clasificación (probabilísticas): AUNU, AUNP, AU1U, AU1P,
* Probabilidad: SAUC, PAUC, MAPR, MPR, MAE, MSE, LogLoss, CalL, CalB.

#### Metodología

Se usaron 6 algoritmos de aprendizaje automático, con 30 sets de datos evaluados usando validación cruzada de 20x5, produciendo 18,000 resultados en total. se construyó matriz de correlación de Pearson para medir la relación entre métricas y dendrogramas para visualizar los datos.

#### Análisis de resultados

Hay una relación cercana entre: 1- métricas cualitativas; 2- AUC; 3- MPR, MAE, PAUC, MAPR; quedando aislados MSE; LogL; CalL y CalB. La correlación baja en sets de datos desbalanceados, generado incluso diferencias entre métricas de umbral del error donde en sets de datos balanceados no las había. Los resultados entre sets de datos pequeños y grandes también afectaron la correlación.

#### Análisis de sensibilidad

Se realizaron mas experimentos para entender mejor la relación entre las métricas y la habilidad de encontrar errores, añadiendo ruido a las clases, a las probabilidades, clasificaciones y a la frecuencia de las clases a 2 modelos controlados. Para ruido en clases es mejor usar métricas cualitativas; para ruido probabilístico es mejor usar AUC y MSE. Con ruido de clasificación mostraron mejor desempeño SAUC, PAUC, MSE, MPR, MAPR, MAE, CalL y AUC. Finalmente con ruido de frecuencia de clases MAvA y MAvG, MPR/MAE, LogL, CalL y MSE son robustas.

#### Discusión

Métricas cualitativas son mejores cuando el ruido esta presente en el dataset, pero son muy malas si se tiene un mal algoritmo o un set de datos pequeño (es mejor usar AUC en estos casos). Si hay pocos ejemplos en la clase, evitar usar los maro promedios o 1vs1, 1 vs-n.