**INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY**

**A blue and white logo

Description automatically generated**

# 

# Clasificación automática de descripciones de carga contenerizada en códigos del Sistema Armonizado (HS) usando modelos de Procesamiento de Lenguaje Natural

**Prof. Titular: Dra. Grettel Barceló Alonso**

**Juan Pablo Acosta López (A01794035)**

**Oscar Villa Cárdenas (A01794052)**

**Genaro Rodriguez Vázquez (A01150931)**

**29 septiembre 2024**

**Índice**

[Clasificación automática de descripciones de carga contenerizada en códigos del Sistema Armonizado (HS) usando modelos de Procesamiento de Lenguaje Natural](#_heading=h.crzpksqdu7ux)

[Construcción de características](#_heading=h.r0z25814vwo2)

[Normalización](#_heading=h.g9gbwkq18las)

[Selección / extracción de características](#_heading=h.ao0fgfb04ccy)

[Conclusiones](#_heading=h.xhep8orgpmfj)

## Construcción de características

**Criterio**: Se espera la creación de características nuevas a partir de variables categóricas y justificar las decisiones.

**Por qué no aplica**: En un proyecto de LLM, las características que se usan son directamente el texto (tokens o secuencias de palabras), y no variables categóricas tradicionales. La representación de las palabras o tokens ya está embebida en vectores mediante modelos preentrenados (como Word2Vec, BERT, etc.). Aquí no se crean características manualmente porque los LLMs aprenden las relaciones de manera automática. La generación de embeddings o representaciones vectoriales del texto es parte del modelo y no requiere una construcción explícita de características a partir de variables predefinidas.

## Normalización

**Criterio**: Escalar variables para asegurar que todas tengan un impacto equitativo.

**Por qué no aplica**: En los proyectos de LLM no es necesario escalar las variables como en los modelos tradicionales. Los textos ya están representados en vectores de alta dimensionalidad mediante técnicas como embeddings, y los LLM manejan directamente las relaciones entre estos vectores. Los modelos basados en texto no requieren la normalización típica que aplicas en datos numéricos. Además, los embeddings ya suelen tener una distribución bien controlada, y cualquier ajuste adicional se realiza internamente dentro del proceso de aprendizaje del modelo.

## Selección / extracción de características

**Criterio**: Reducir la complejidad mediante selección de características.

**Por qué no aplica**: Los LLMs no funcionan con características específicas seleccionadas manualmente. En lugar de eso, manejan secuencias de texto completo, y es el propio modelo el que determina qué partes del texto son relevantes para una tarea dada. Los modelos de lenguaje como BERT, GPT, etc., ya están diseñados para interpretar y aprender de todo el texto de manera automática, por lo que la selección de características manuales no es necesaria.

## Conclusiones

Aunque es posible ofrecer una conclusión sobre el preprocesamiento de los datos textuales (como tokenización, limpieza de texto, etc.), la metodología CRISP-ML está más orientada a proyectos de ciencia de datos tradicionales con modelos numéricos o categóricos. En los proyectos de LLM, la preparación de datos es diferente, ya que se centra más en la limpieza textual y menos en la transformación de variables numéricas.