Actividad previa

- Revisar conjuntos de datos seleccionados por los estudiantes y brindar retroalimentación.
- Explicación de la distancia de Mahalanobis.

Resultado de Aprendizaje

RdA de la asignatura:

- **RdA 1:** Plantear los conceptos fundamentales del aprendizaje automático, incluyendo los principios básicos, técnicas de preprocesado de datos, métodos de evaluación y ajuste de modelos, destacando su importancia en el análisis y resolución de problemas de datos.

RdA de la clase:

- Comprender las técnicas principales de preprocesado de datos: escalado, discretización y reducción de dimensionalidad.
- Aplicar división de conjuntos de datos en entrenamiento y prueba para evaluar modelos de aprendizaje automático.

Introducción

Pregunta inicial:

- Si en un modelo una de las varaibles, por ejemplo, la edad, tiene un rango de valores entre 0 y 100, y otra variable, por ejemplo, el ingreso anuel, tiene un rango entre $0 y $100000, ¿qué problemas podrían surgir al calcular distancias entre puntos en este espacio de características?
- ¿Si en un examen de matemática te evalúan con exactamente los mismos ejercicios que practicaste, qué tan bien crees que mediría tu conocimiento real de la materia?

Desarrollo

Actividad 1: Preprocesado de Datos

En esta actividad los estudiantes aprenderán las técnicas básicas de preprocesado de datos mediante clase magistral y práctica en cuaderno de Jupyter, incluyendo escalado, discretización y reducción de dimensionalidad para preparar datos de aprendizaje automático.

¿Cómo lo haremos?

- **Conceptos fundamentales:** Se presentarán las técnicas de preprocesado: escalado, discretización, one-hot-encoding y reducción de dimensionalidad.
- **Implementación en Python:** Los estudiantes accederán a un cuaderno de Jupyter previamente preparado.

  Enlace al cuaderno: 03-1-Preprocesado-Datos.ipynb.
- **Experimentación:** Realizar los ejercicios propuestos en el cuaderno.

Actividad 2: Conjuntos de Entrenamiento y Test

En esta actividad se abordará la importancia de dividir un conjunto de datos en entrenamiento y prueba mediante exploración de cuaderno de Jupyter.

¿Cómo lo haremos?

- **Métodos de partición:** Se presentarán las necesidad de particionar conjuntos de datos.
- **Implementación en Python:** Los estudiantes accederán a un cuaderno de Jupyter previamente preparado.

  Enlace al cuaderno: 03-2-Conjuntos-Entrenamiento-Prueba.ipynb.
- **Experimentación:** Realizar los ejercicios propuestos en el cuaderno.

Cierre

Verificación de aprendizaje:

1. ¿En qué rangos queda una variable luego de aplicar cada tipo de normalización?
2. ¿Cuál es la diferencia entre discretizar una variable usando igual amplitud o igual frecuencia?
3. ¿Dónde ingresan el concepto de valores propios en el ACP?

Preguntas tipo entrevista:

1. Describe la diferencia entre discretización y one-hot encoding.
2. ¿Qué método de normalización utilizarías para variables categóricas antes de entrenar un modelo?

Tarea: Realizar el procesamiento de datos necesarios al conjunto de datos seleccionado en la clase anterior y subirlo al repositorio de GitHub.

Pregunta de investigación:

1. ¿Qué estrategias existen para manejar conjuntos de datos desbalanceados al realizar particiones de entrenamiento y prueba?
2. ¿Realizar one-hot encoding agrega colinealidad a las caracterísitcas? Ver: The Problem With One-Hot Encoding

Para la próxima clase: Se debe tener el conjunto de datos preprocesados y listos en el repositorio.