• **Nombre de la institución:**

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

• **Nombre del alumno:**

JACKELINE MONGE NUNEZ

• **Nombre del profesor:**

JOSUE PARRA

• **Materia:**

PROGRAMACION AVANZADA

• **Título del trabajo:**

ANÁLISIS DE CONCEPTOS DE EXTRACCIÓN DE DATOS

• **Fecha de entrega:**

16 DE MARZO DEL 2025

• **Grupo:**

951

Análisis de conceptos de preprocesamiento de datos

Introducción

El preprocesamiento de datos es una etapa fundamental en todo proyecto de análisis o minería de datos. Consiste en preparar, limpiar y transformar los datos crudos para que sean útiles y confiables antes de aplicar modelos estadísticos o de aprendizaje automático. En este documento se analizan sus principales conceptos y tareas básicas.

¿Qué es el Pre procesamiento de datos?

El preprocesamiento de datos es el conjunto de técnicas utilizadas para convertir los datos originales (que pueden tener errores, valores faltantes, o estar desorganizados) en un formato limpio y listo para el análisis. Este paso es esencial para evitar sesgos, errores o resultados inexactos en cualquier estudio basado en datos.

Importancia del preprocesamiento

• Mejora la calidad de los datos.

• Elimina errores, inconsistencias y duplicados.

• Aumenta la eficiencia de los algoritmos de análisis.

• Es fundamental para obtener conclusiones válidas y confiables.

Tareas básicas de preprocesamiento de datos

1. Valores faltantes

Los datos faltantes ocurren cuando no se dispone de un valor en una o más columnas. Se pueden tratar de distintas formas:

• Eliminar las filas o columnas afectadas.

• Rellenar con la media, moda, mediana o algún valor específico.

• Usar métodos estadísticos o algoritmos de imputación.

2. Columnas irrelevantes

Son aquellas columnas que no aportan información útil al análisis, como identificadores únicos o códigos internos. Eliminarlas reduce el ruido y mejora el rendimiento de los modelos.

3. Valores atípicos (outliers)

Son datos que se alejan significativamente del resto. Se detectan con técnicas como boxplots o desviación estándar. Dependiendo del caso, se pueden corregir, eliminar o conservar.

4. Normalización de datos

Escalar los datos numéricos para que estén en un mismo rango (por ejemplo, entre 0 y 1) es importante para algoritmos sensibles a magnitudes como redes neuronales o KNN.

5. Discretización de datos

Consiste en convertir variables numéricas en categorías. Por ejemplo, la edad puede agruparse en rangos: joven, adulto, adulto mayor. Esto simplifica ciertos análisis y modelos.

6. Datos duplicados

Registros exactamente iguales o muy parecidos pueden distorsionar los resultados. Se identifican y eliminan para mantener la integridad de los datos.

7. Procesamiento de datos categóricos

Los modelos de aprendizaje automático requieren datos numéricos. Por eso, los datos categóricos (como “masculino” o “femenino”) se convierten en números mediante técnicas como:

• One-hot encoding (crear columnas binarias)

• Label encoding (asignar números a cada categoría)

Conclusión

El preprocesamiento de datos es una etapa crucial para aprender porque garantiza que el análisis posterior sea confiable, eficiente y asi poder validarlo. Sin un preprocesamiento adecuado, cualquier modelo o estudio de datos puede arrojar resultados con errores. Por lo tanto, comprender y aplicar correctamente las técnicas de limpieza y transformación es esencial para todo programador, analista o científico de datos.

Referencias (formato APA 7.0)

• García, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining*. Springer.

• Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

• IBM. (2020). *What is Data Preprocessing?* Recuperado de https://www.ibm.com/cloud/learn/data-preprocessing

• Towards Data Science. (2021). *Data Preprocessing for Machine Learning in Python*. https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825