Aplicación de Métodos Multivariados en Ciencia de Datos

Reto MA2003 sesión 1

Dr. Hugo García Tecocoatzi
Dr. Andrés Ramírez Morales

Instituto Tecnológico y de Estudios Superiores de Monterrey

Contexto de la empresa

Interlub es una empresa fundada en 1984 dedicada a la producción de lubricantes especializados. A diferencia de grandes competidores como *Shell* o *Mobil*, Interlub se enfoca en desarrollar soluciones **personalizadas** para necesidades específicas de sus clientes.

Características clave de Interlub

- Desarrollo de **tecnología propia** y soluciones a la medida.

Actividad 0

- Analizar 5 productos externos y entender las "características típicas"

El reto propuesto

El desafío planteado para la materia es el desarrollo de un:

Recomendador de productos de lubricación especializada.

Descripción del problema

- Catálogo de más de **2500 productos** con distintas características técnicas. (Pero no tenemos acceso a éste)
- La selección del producto óptimo es compleja y depende de expertos internos.

Objetivo del reto

Diseñar un modelo que:

- Use las **fichas técnicas** de los productos de Interlub.
- Permita recomendar lubricantes según un producto competidor o una necesidad técnica.
- Agilice el proceso de selección.

Recursos disponibles

- Base de datos en CSV con alrededor de 50 productos.
- Fichas técnicas de competidores en PDF (5 productos).
- Posible acceso a más información conforme avance el proyecto. (Luis trabaja en ello)

Detalles adicionales del reto

Durante la sesión de preguntas se aclararon los siguientes puntos clave:

- El modelo debe funcionar **por similitud**.
- Salida: un **ranking** de productos (ej. Top 5 más similares).
- Identificar diferencias y características modificables.
- Enfoque híbrido: datos técnicos + necesidades del cliente.

Sobre los datos

- Fichas técnicas incluyen pruebas comunes (viscosidad, punto de fusión, pH, etc.).
- Existen **valores faltantes** en varias pruebas.
- No hay datos temporales (históricos de uso o ventas).

Consejos de Luis Ibarra (Interlub)

- Familiarizarse con los datos y las pruebas técnicas.
- Aplicar análisis multivariado (correlaciones, reducción de dimensionalidad, etc.).
- Modelo flexible y útil para consultores e ingenieros de ventas.
- Mantener creatividad: se valoran soluciones innovadoras.
- Soporte de Luis Ibarra (9 am – 8 pm) y posible canal grupal (Vamos a crear un canal donde todos estemos) .

Entregable

- Entregable: prototipo funcional + reporte académico + Notebooks.
- Presentación final en diciembre (semana 16-17 del semestre).

Contexto del Dataset

Base de Datos de Grasas Lubricantes

- 50 productos diferentes de grasas
- 40+ variables técnicas por producto
- Datos de composición, propiedades físicas y aplicaciones
- Lo malo: Presencia significativa de valores faltantes

Ejemplos de Variables

- **Físicas:** Viscosidad, Penetración, Punto de Gota
- **Desempeño:** Estabilidad Mecánica, Resistencia a Cargas
- **Aplicación:** Temperaturas, Industrias, Condiciones

Valores faltantes

Un problema típico

- Datos incompletos en múltiples variables
- Imposibilidad de análisis con dataset completo
- Riesgo de sesgos en las conclusiones

Solución Propuesta

Creación de **sub-bases de datos** específicas para análisis particulares

Actividad 1: Limpieza y análisis de la base de datos

Objetivo

Analizar la base de datos

- Eliminar cololumans no tengan más de 5 datos.
- Separa variables númericas
- Separar variables categóricas
- Explorar como crear subbases de datos

Datos de las grasas de otros provedores

Extraer la información de las grasas de la competencia

Actividad 2: Selección por Condiciones Específicas

Enfoque Práctico

Crear sub-bases para condiciones operativas específicas del usuario

Criterios de Filtrado:

- Rango de temperatura
- Resistencia al agua
- Carga extrema específica

Ejemplo: Alta Temperatura

- Temp. máxima ¿ 180 C
- 15 productos cumplen
- Análisis de propiedades comunes

Actividad 3: Análisis de Correlaciones Técnicas

Objetivo

Identificar relaciones entre propiedades

Sub-base Utilizada:

- Productos con datos completos
- 20+ variables técnicas

Correlaciones Identificadas

Hallazgos Principales:

- **Alta correlación**
- **Relación inversa**
- **Patrón claro**
- Etc...

Limitación

Correlación no implica causalidad - validar

Lecciones aprendidas

Ventajas del Enfoque por Sub-bases

- **Máximo aprovechamiento** de datos disponibles
- **Análisis específicos** para preguntas concretas
- **Reducción de complejidad** en cada análisis
- **Resultados accionables** para diferentes usuarios

Precauciones Necesarias

- Documentar criterios de filtrado
- Evaluar posibles sesgos
- Validar con conocimiento experto
- Comunicar limitaciones claramente

Recomendaciones: Datos con Valores Faltantes

Recomendaciones:

1. Diagnóstico completo de patrones de datos faltantes
2. Definir objetivos específicos
3. Crear sub-bases justificadas
4. Validar representatividad
5. Documentar todo el proceso

con conocimiento de dominio