Universidad San Carlos de Guatemala
Facultad de ingeniería.
Ingeniería en ciencias y sistemas
Segundo semestre
Escuela de vacaciones de diciembre

Título del Proyecto:

InsightCluster

PONDERACIÓN: 45

Horas Aproximadas: 30

Resumen Ejecutivo

Las empresas que ofrecen productos o servicios suelen disponer de grandes volúmenes de datos sobre sus clientes, tales como información de consumo y opiniones expresadas en reseñas. Sin embargo, en muchos casos estos datos no cuentan con etiquetas que permitan un análisis directo, lo que dificulta la identificación de patrones de comportamiento y preferencias.

InsightCluster propone el desarrollo de una aplicación que integra técnicas de **aprendizaje no supervisado** para realizar la **segmentación de clientes** y el **agrupamiento de reseñas de productos**, con el fin de descubrir patrones ocultos en los datos.

Objetivos del Proyecto

Objetivo General

Desarrollar una aplicación basada en aprendizaje no supervisado que permita al estudiante aplicar los conocimientos adquiridos en el curso para entrenar, evaluar y ajustar modelos de predicción.

Objetivos Específicos

- Aplicar algoritmos de aprendizaje no supervisado para identificar patrones y agrupaciones en datos numéricos y textuales.
- Implementar procesos de limpieza, normalización y vectorización de datos como parte del flujo de análisis.
- Analizar e interpretar los resultados obtenidos, justificando las decisiones técnicas tomadas durante el desarrollo del proyecto.

Enunciado del Proyecto

En entornos comerciales y digitales, los datos de clientes y las opiniones expresadas en reseñas suelen crecer rápidamente y carecer de una clasificación previa. Esto dificulta comprender el comportamiento de los usuarios, identificar segmentos con características similares o detectar temas recurrentes en las opiniones.

Detalles del proyecto

Carga masiva y preprocesamiento de datos

El proyecto **InsightCluster** cuenta con un módulo encargado de la carga, validación y preprocesamiento de los datos de entrada. Este módulo permite incorporar conjuntos de datos que contengan información de clientes y reseñas textuales, asegurando su correcta limpieza y transformación para el análisis posterior.

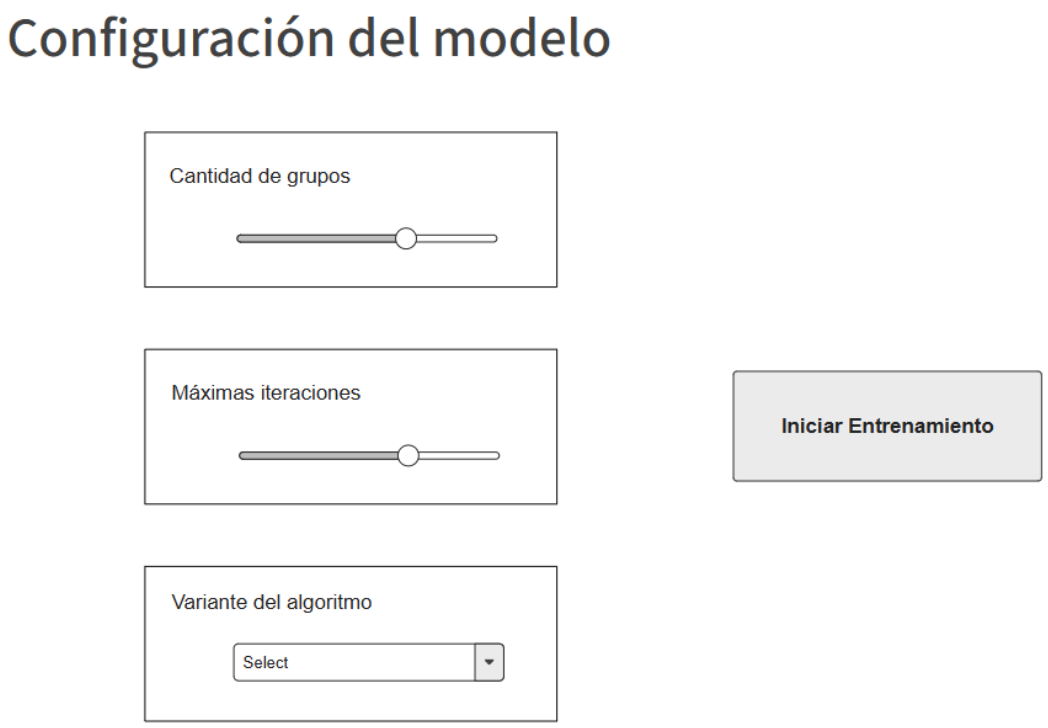
Nota: Esta imagen es solamente de referencia.

Las variables que contendrá el archivo son:

1. **cliente_id:** Identificador único del cliente.
2. **frecuencia_compra:** Número de compras realizadas en un periodo definido.
3. **monto_total_gastado:** Gasto acumulado del cliente.
4. **monto_promedio_compra:** Promedio gastado por compra.
5. **dias_desde_ultima_compra:** Tiempo transcurrido desde la última compra.
6. **antiguedad_cliente_meses:** Tiempo que el cliente lleva utilizando el servicio.
7. **canal_principal:** Canal más utilizado (web, móvil, tienda física, etc.).
8. **numero_productos_distintos:** Cantidad de categorías o productos diferentes comprados.
9. **reseña_id:** Identificador único de la reseña.
11. **texto_reseña:** Opinión escrita por el cliente.
12. **fecha_reseña:** Fecha en que se realizó la reseña.
13. **producto_categoria:** Categoría del producto reseñado.
14. **longitud_reseña:** Número de palabras o caracteres (deberá calcularse si se considera necesaria).

Configuración y entrenamiento del modelo

InsightCluster cuenta con un módulo que permite configurar y ejecutar el proceso de aprendizaje no supervisado de forma controlada. En este componente, se debe integrar al menos un algoritmo de clustering, permitiendo definir parámetros esenciales como el número de grupos o la métrica de similitud. El módulo será responsable de entrenar el modelo, asignar cada cliente o reseña a un segmento y almacenar los resultados generados.

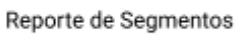


Nota: Esta imagen es solamente de referencia, las opciones de configuración deben coincidir con el modelo seleccionado.

Interpretación y perfilado de los segmentos descubiertos

InsightCluster busca ser una herramienta para las empresas y más allá de realizar el entrenamiento de un modelo de machine learning busca otorgar valor y generar conocimiento sobre sus clientes y productos. Se debe incluir la representación de los resultados obtenidos del clustering, de forma clara y sencilla, a través de gráficas y tablas que apoyen su comprensión.

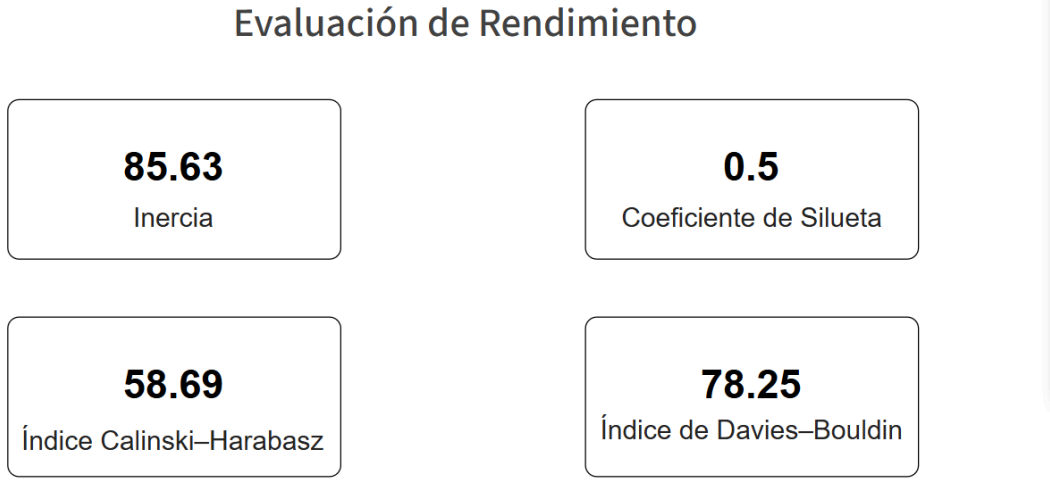
Se deberá generar descripciones comprensibles de cada segmento identificado, destacando las características más representativas de los clientes y los temas predominantes en las reseñas asociadas. El objetivo es traducir los resultados técnicos del modelo en información útil y entendible, reforzando la conexión entre el análisis de datos y la toma de decisiones.



Nota: Esta imagen es solamente de referencia.

Evaluación y validación del análisis

Los usuarios de InsightCluster deben poder confiar en el análisis que el modelo ha realizado de sus datos, por ende se debe proporcionar un apartado en el que se evaluará la calidad del mismo a través del cálculo de métricas internas de validación, como medidas de cohesión y separación entre clusters.



Nota: Esta imagen es solamente de referencia.

Exportación de reportes

Finalmente, InsightCluster facilita la exportación de los resultados del análisis permitiendo la generación de archivos con los datos segmentados y resúmenes de los clusters, de forma que puedan ser utilizados en informes y presentaciones.

Nota: Esta imagen es solamente de referencia.

Restricciones

- No se permite el uso de APIs externas (como la de chatgpt) o modelos preentrenados, el modelo a utilizarse debe ser de su completa creación.
- El lenguaje de programación para el entrenamiento del modelo y el desarrollo del backend debe ser Python.
- El lenguaje de programación para el frontend queda a elección del estudiante. Se recomienda React, Streamlit o Vue.
- La entrega se realizará por medio de Github. Pueden usar el repositorio creado para el primer proyecto agregando una carpeta con el nombre Proyecto2 y a su nuevo compañero o crear un repositorio nuevo con el nombre OLC2_2SEVD25_Proyecto2_#grupo.
- Por medio de UEDI se hará entrega del link del repositorio.
- Los repositorios deben mantenerse en privado para evitar copias de código entre estudiantes. **Usuario de github**: KatherinneGomez
- Se debe realizar el manual técnico en el archivo README dentro de la carpeta de la respectiva, explicando el proceso de limpieza de datos, la selección del modelo, las decisiones de diseño tomadas, conclusiones y lecciones aprendidas.
- Todas las dudas durante el proceso de desarrollo deberán realizarse por medio de los foros de UEDI.
- Se recomienda realizar commits frecuentes en Github para demostrar el avance progresivo del proyecto.
- El código debe estar debidamente comentado y seguir buenas prácticas de programación.
- **Fecha de entrega 30/12/2025 a las 23:59 horas.**

Penalizaciones

- Copias detectadas obtendrán una nota de 0 puntos y se reportará a la Escuela de Ciencias y Sistemas.
- No se permiten entregas tarde.